

BAYESIAN VIRAL SUBSTITUTION ANALYSIS AND COVARIANCE ESTIMATION
VIA GENERALIZED FIDUCIAL INFERENCE

Wen Jenny Shi

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the
Department of Statistics and Operations Research.

Chapel Hill
2015

Approved by:

Jan Hannig

Corbin Jones

Shankar Bhamidi

Lu Shu

Kai Zhang

©2015
Wen Jenny Shi
ALL RIGHTS RESERVED

ABSTRACT

Wen Jenny Shi: Bayesian Viral Substitution Analysis and Covariance Estimation via
Generalized Fiducial Inference
(Under the direction of Jan Hannig and Corbin Jones)

With the advances in biology and computing technologies, there have been increasing amount of big bio data awaiting to be analyzed. Aiming to develop statistical tools for omics data, we focus on the problem of viral sequencing data modeling as well a fundamental statistics question with applications in both biology and many other fields. This dissertation is comprised of three major parts.

Motivated by a multi-time sampled, case-control influenza viral population study, in the first part we model the sequencing data of a viral population under a Bayesian Dirichlet mixture distribution. We have developed an efficient clustering scheme that enables us to distinguish treatment causal changes from variation within viral populations. As a proof of concept, we applied our method to a well-studied HIV dataset, and successfully identified known drug resistant regions and additional potential sites. For the influenza data, our algorithm revealed two genome sites with strong evidence of treatment effect.

The second part of the thesis concerns the covariance matrix estimation in a high-dimensional multivariate linear models and sparse covariate settings using fiducial inference. The sparsity imposed on the covariate matrix allows to estimate relationships between a list of gene expressions and several metabolic levels under a high dimension low sample size setting. Aiming to quantify the uncertainty of the estimators without having to choose a prior, we have developed a fiducial approach to the estimation of covariance matrix. Built upon the Fiducial Bernstein-von Mises Theorem, we show that the fiducial distribution of the covariance matrix is consistent under our framework. Furthermore, we propose an adaptive efficient reversible jump Markov chain Monte Carlo algorithm for sampling from the fiducial

distribution, which enables us to define a meaningful confidence region for the covariance matrix.

In the last part of the thesis, we examine the stochastic models for capturing the evolutionary processes of gene expression levels. Generalizing a microarray Brownian motion (BM) model, we have developed a BM model for high-throughput sequencing data that takes sampling variance into account. To allow conservation in the evolution process, we also investigate Ornstein-Uhlenbeck (OU) models. Applying to a multiple-tissue mammalian dataset, we showed that the OU model is more appropriate for the top 10 highly expressed genes in the dataset, and we performed hypothesis testing for significant changes in gene expression levels along specific lineages.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES.....	ix
1 Introduction	1
2 Viral substitution analysis	3
2.1 Introduction	3
2.2 Parametric Bayesian Mixture Framework	6
2.2.1 Sequencing Data	6
2.2.2 Dirichlet Mixture Model.....	7
2.2.3 Hellinger Distance.....	9
2.3 Methodology	10
2.3.1 Preprocess	12
2.3.2 Processing	13
2.3.3 Postprocess	16
2.4 Simulation Study	19
2.5 Real Data Analysis	23
2.5.1 Human immunodeficiency virus 1 (HIV-1)	24
2.5.2 H1N1 Influenza A (IVA).....	27
2.6 Discussion.....	38
2.7 Appendix	39
2.7.1 Proof of Theorem 2.1.....	39
2.7.2 Additional IVA Ht plots.....	42
2.7.3 Raw data plot.....	43

3	Covariance estimation via fiducial inference	58
3.1	Introduction.....	58
3.2	Generalized fiducial inference	60
3.2.1	Brief background	60
3.2.2	Generalized fiducial distribution	61
3.3	A fiducial approach to covariance estimation	63
3.3.1	Data generating equation.....	63
3.3.2	Jacobian	64
3.4	Theoretic results	68
3.5	Reversible jump Markov chain Monte Carlo.....	70
3.5.1	Algorithm flow	71
3.5.2	Jump map	72
3.5.3	Zeroth-order method	73
3.6	Implementation	74
3.6.1	Special Case I: No fixed zero entries in A	74
3.6.2	Special Case II: Clique model.....	74
3.6.3	General case with sparsity known.....	76
3.6.4	General case with sparse locations unknown	78
3.7	Discussion.....	79
3.8	Proofs.....	80
3.8.1	Proof of Proposition 3.1	80
3.8.2	Proof of Proposition 3.2	81
3.8.3	Proof of Proposition 3.3	82
3.8.4	Proof of Theorem 3.1.....	85
3.9	Appendix	89
3.9.1	Förstner-Moonen distance (FM-distance)	89
4	Phylogenetically dependent gene expressions study	97

4.1	Introduction.....	97
4.2	Stochastic models for phylogenetically dependent gene expressions.....	98
4.2.1	Brownian motion (BM).....	99
4.2.2	Ornstein-Uhlenbeck (OU)	100
4.2.3	Lévy processes	102
4.2.4	Parametric bootstrap.....	103
4.3	Implementation	103
4.3.1	Multi-species multi-tissues Mammalian data	105
4.3.2	BM vs OU	107
4.3.3	Mean shift test	110
4.4	Discussion.....	110
	BIBLIOGRAPHY.....	113

LIST OF TABLES

2.1	High through-put sequencing count data	6
2.2	Data preprocessing	12
2.3	Simulation result summary	23
2.4	Simulation efficiency summary	24
2.5	H1N1 result using Passages 1, 3, 9, 12, and the end time point data	34
2.6	H1N1 result with Passages 1, 3, 9, and 12	36
3.1	Jump map (part1)	73
3.2	Jump map (part 2)	73

LIST OF FIGURES

2.1	H1N1 experiment setup.....	5
2.2	Comparing distributions with Hellinger distance	11
2.3	Three steps clustering procedure	14
2.4	Experiment setup for the toy example and HIV-1 dataset	18
2.5	Simulation experimental design	20
2.6	Test 1 Ht plot	22
2.7	HIV-1 Ht plot	26
2.8	H1N1 Seg6E1 Ht plot	28
2.9	H1N1 Seg6E2 Ht plot	29
2.10	S6-822 raw read	30
2.11	H1N1 Seg7E1 Ht plot	31
2.12	H1N1 Seg7E2 Ht plot	32
2.13	S7-91 raw read	33
2.14	H1N1 experiment setup with 12 passages per line	37
2.15	Seg4E1 Ht plot	43
2.16	Seg4E2 Ht plot	44
2.17	Seg5E1 Ht plot	45
2.18	Seg5E2 Ht plot	46
2.19	Seg8E1 Ht plot	47
2.20	Seg8E2 Ht plot	48
2.21	S8-80 raw read	49
2.22	S1-2299 raw read	50
2.23	S1-2303 raw read	51
2.24	S3-2193 raw read	52
2.25	S4-1210 raw read	53
2.26	S5-24 raw read	54
2.27	S5-389 raw read	55

2.28	S5-1103 raw read	56
2.29	S8-819 raw read	57
3.1	Clique example 1	91
3.2	Clique example 2	92
3.3	Clique example 3	93
3.4	General case with sparsity known example 1	94
3.5	General case with sparsity known example 2	94
3.6	General case with sparsity known example 3	95
3.7	General case with sparsity unknown example 1	95
3.8	General case with sparsity unknown example 2	96
3.9	General case with sparsity unknown example 3	96
4.1	Simple phylogeny	98
4.2	BM trace plot	99
4.3	OU trace plot	101
4.4	Mammalian phylogeny	104
4.5	Heat map of mammalian gene expression data	106
4.6	Heat map of gene expression in mammalian brain tissue	107
4.7	Multiple dimensional scaling plot	108
4.8	BM vs OU test	109
4.9	OU model mean shift test	110

CHAPTER 1

Introduction

In recent years, with the advances in data collection technologies and computing, large amounts of data have been and are continuously harvested. There is a critical need for developing powerful analytical tools and extracting the important information embedded. In this dissertation we present several statistical methods developed for analyzing omics data in the field of biology, which may be extended to other fields as well.

In Chapter 2, we focus on modeling the RNA sequencing (RNA-seq) data and detecting drug resistant regions on the genome for viral populations. To describe the RNA-seq read count distributions, we suggest a Bayesian Dirichlet mixture framework. We develop an efficient three-step clustering procedure to generate the mixture clusters without requiring to specify the number of mixture components *a priori*. Our method analyzes data collected from multiple time points and/or under control and treatment environment simultaneously, and compares posterior distributions for the same genomic location across time and treatment environments. Through simulations we showed that our clustering algorithm is much more efficient comparing to direct Gibbs sampler. We further applied our method to a well-studied HIV-1 dataset and an H1N1 data with two biological duplicates. Our method revealed the most common known drug resistant sites along with a few other interesting genomic locations.

Next, in Chapter 3, we look into covariance estimation, a rather classical statistics problem. Instead of the common sparse covariance constraint, we impose a sparsity structure on the covariate matrix, a scenario that arises in proteomics and metabolomics. Aiming for a distribution of estimators without requiring priors, we considered a fiducial approach. Under the assumption that there is a one-to-one correspondence between the covariate and covariance matrices, (which is often true under the sparse setting), we prove that the derived fiducial distribution satisfies the Fiducial Bernstein von-Mises Theorem (Sonderegger

and Hannig, 2012). To sample from the fiducial distribution, we suggest to use Markov chain Monte Carlo (MCMC) methods. In the general case where the sparse structure of the covariate is unknown, we propose an adaptive Reversible Jump MCMC (RJMCMC) that incorporates the zeroth-order method to improve efficiency.

Finally, we present a review of stochastic modeling for phylogenetically dependent continuous traits in Chapter 4. How much a gene is expressed can determine important phenotypes and other characteristics of an organism. The study of gene expression modeling has been a popular area in the past decades. The stochastic nature of the evolutionary process of gene expression needs to be taken into account when the expression levels are compared across related species. Here, we review the stochastic modeling of gene expression levels for phylogenetically dependent species using Brownian motion (BM), Ornstein-Uhlenbeck (OU), and general Lévy processes. For illustration, the BM and OU methods were applied to the RNA-seq data of nine mammalian species. Based on the top 10 highly expressed genes, we showed that the OU model is more appropriate than the BM model, and that there is no significant mean shift on the mouse branch, even though its expression levels appear to be much more different from the others.

CHAPTER 2

Viral substitution analysis

2.1 Introduction

RNA viruses and retroviruses, such as SARS, influenza, hepatitis C, polio, and HIV, use RNA as their genetic material. The RNA polymerases of these viruses lack the proof-reading ability of DNA polymerases, which results in a high mutation rate in these RNA genomes and a high rate of genome evolution. This rapid rate of evolution can be advantageous for the virus as it can confound the immune system and lead to the emergence of resistance to antiviral drugs (Boutwell et al., 2010; Rambaut et al., 2004).

Phylogenetic and molecular evolutionary analyses of viral genes and genomes are standard tools for investigating RNA virus evolution at a molecular level (Norström et al., 2012). However, the high mutation rate and the complex secondary structures of RNA viruses genomes often compromise sequence based methods of analysis (Simmonds and Smith, 1999; Damgaard et al., 2004; Watts et al., 2009; Cuevas et al., 2012). These aspects of viral biology complicate teasing apart the evolutionary signal of adaptation, such as evolution of drug resistance, from the signal of neutral evolutionary processes, such as genetic drift. Further complicating sequence analysis are compensatory mutations that offset structural defects and other pleiotropic costs of adaptive alleles, which often arise and sweep to fixation in viral populations (Knies et al., 2008). Thus there is a clear need for analytical methods that are robust to these complications, make minimal assumptions as to how the virus should evolve, and can identify regions of the viral genome that have changed over time in response to treatment.

The wealth of new viral sequence data made possible by recent advances in sequencing technology has amplified the need for new analytical tools (Jabara et al., 2011). Increasingly, populations of thousands of viruses are sampled and sequenced from an infected individual.

This approach captures a snapshot of the viral genetic variation within an individual. A few studies have combined this approach with traditional passage experiments or sampling during the course of an infection (Eriksson et al., 2008; Kuroda et al., 2010; Leitner et al., 1993; Wright et al., 2010). This powerful experimental design reveals how a population of viruses genomically responds to evolutionary pressure. With the ever-decreasing cost of sequencing, these studies are expected to become commonplace.

Our motivating dataset came from a study of influenza A H1N1 viruses (IVA) in response to an inhibitor of neuraminidase, oseltamivir (a.k.a. Tamiflu). Oseltamivir has been used both for prevention and treatment of influenza viruses. It prevents the virus from budding from the host cell, thereby slowing viral reproduction. How the IVA respond to oseltamivir on the genomic level has not been fully understood. Our goal is to find the genomic regions of the virus that evolved in response to oseltamivir. The dataset contains replicate populations of IVA sampled over many generations (“passages”) in the presence and absence oseltamivir (Renzette et al., 2014). The IVA were first adapted from chicken eggs to Madin-Darby canine kidney (MDCK) cells for three passages. Then the samples were serially passaged in MDCK cells in either the absence or presence of oseltamivir in replicated experiments (Figure 2.1). At the end of each passage, whole-genome high throughput sequencing data were collected (Renzette et al., 2014).

RNA viruses evolve rapidly even within the untreated group. It is important to distinguish genetic changes selected for by the inhibitor from those that arise due to other population genetic forces. The time series data and control-treatment setup provides multiple samples for the virus populations with and without the administration of oseltamivir. Two biological replicates allow to crosscheck sites for drug resistance. We take advantage of the replicated longitudinal data and develop a novel statistical approach for identifying evolved nucleotides in a viral genome without relying on sequence annotation or the nature of the change (non-synonymous or synonymous; transition or transversion).

Our approach analyzes multiple time-sampled observations simultaneously, models viral sequence position indices under a Bayesian Dirichlet mixture distribution, performs a series of clustering algorithms, and identifies treatment causal substitution sites via comparing the before and after treatment posterior distributions for the corresponding regions on the

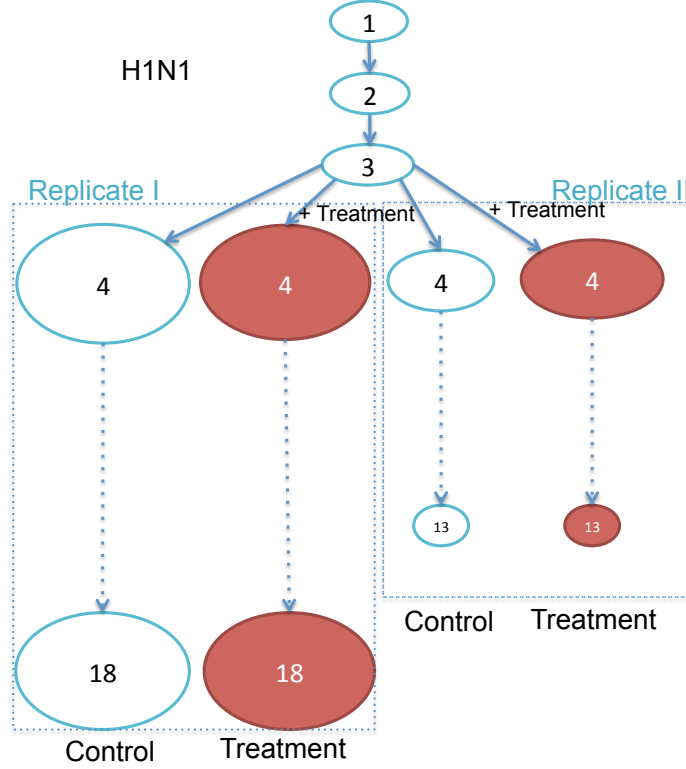


Figure 2.1: IVA adapted from chicken egg to MDCK cells for passages 1-3, then serially passaged in either absence (white) or presence (red) of oseltamivir environments. There are two biological replicates. The size of the oval corresponds to the average total read count per site. The number in the oval corresponds to the generation.

viral genome. Our algorithm also allows us to identify genomic locations that have similar patterns of change.

We first validated our approach with synthetic test data. Then we used a well-studied HIV-1 data set (Jabara et al., 2011) as a positive control. We showed that this approach identifies key changes that have been experimentally shown to be important to the evolution of drug resistance. Finally, we applied our method to the longitudinal time-sampled IVA data in the absence and presence of oseltamivir. We identified two genome sites (S6-822 & S8-80) that presented the greatest evidence of drug resistance along with a set of locations might have been affected by adaptation to the host or genetic drift.

The rest of the chapter is arranged as follows. Section 2.2 describes the viral genome data type, a Bayesian framework used to model the viral populations, and an f-divergence measure used in our study. Section 2.3 introduces a three-step sequential approach we have

developed to identify treatment causal substitutions. Simulation results are presented in Section 2.4, and in Section 2.5 implementation of our method to a well described HIV-1 dataset as a proof of concept, followed by the analysis of the IVA dataset. Section 5 concludes the chapter with a few remarks and a discussion.

2.2 Parametric Bayesian Mixture Framework

In this section, we first briefly introduce the whole genome high-throughput sequencing data. We then define the Bayesian Dirichlet mixture framework used to model a viral population and state a distance measure for comparing the distributions for the same genome position across time.

2.2.1 Sequencing Data

Advances in high-throughput whole genome shotgun sequencing allow deep genome sequencing of viral populations within a host (Muers, 2011). This technology produces millions of short DNA or RNA sequences. These sequences are aligned to a reference genome and differences between the reference and sequenced population are noted. With this advanced shotgun sequencing method, we are able to combine the reads from each individual and work with data with the following form:

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	\cdots			Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	\cdots	
					C	A	T	\cdots			A	0	0	0	0	2	0	2	0	\cdots
C	T	C	T	A	C	A		\cdots			C	1	1	3	0	1	5	1	0	\cdots
		C	T	A	C	C	M	\cdots	\Rightarrow		G	0	1	0	1	0	0	0	0	\cdots
				G	C	T	T	\cdots			T	0	1	0	3	1	0	2	2	\cdots
	C	M	G	T	C	T		\cdots			M	0	0	1	0	0	0	0	1	\cdots
	G	C	T	C				\cdots												

Table 2.1: High through-put sequencing data from all samples are pooled and aligned (left panel) and then compressed into a five-row count matrix for the genome of interest (right panel).

Letters A, C, G, T, M stand for five possible read types in this toy example: Adenine, Cytosine, Guanine, Thymine, Missing/deleted data, respectively. Left-hand side of Table 2.1 illustrates a high through-put sequencing alignment result. Its read-specific compressed

view is shown in the right panel of Table 2.1. Counts of each read type (A, C, G, T, M) at the i^{th} position are recorded as $Y_i = (y_i^1, y_i^2, y_i^3, y_i^4, y_i^5)$.

2.2.2 Dirichlet Mixture Model

To describe the genomic site specific variation residing within a viral population we constructed a parametric Bayesian mixture model based on observed nucleotide read counts. Assume that total number of read types is J . Given the probability parameters, the collection of different read counts at each genomic site is assumed to follow a J -dimensional multinomial distribution. For an arbitrary i^{th} position on the sequence, the probabilities of having each of the J read types are denoted as $P_{c_i} = (p_{c_i}^1, p_{c_i}^2, \dots, p_{c_i}^J)$. Every p_{c_i} lies between 0 and 1; their summation $\sum_{j=1}^J p_{c_i}^j = 1$. We assume a finite collection of K possible probability parameters, $\mathbb{P} = \{P_1, \dots, P_K\}$, each genomic site could take on, i.e. every P_{c_i} is a member of \mathbb{P} . The subscript c_i is an assignment indicator denoting which probability parameter in the set \mathbb{P} the i^{th} genomic site is associated with, $c_i \in \{1, \dots, K\}$. The number of elements in \mathbb{P} , K , is the number of mixture components in the Bayesian mixture framework. Because many sites in the genome sequence share the same tendencies of having certain kinds of genetic variation (as captured by the reads), it is intuitive that K is much smaller than the length of the viral sequence of interest, N . Furthermore, a weakly informative symmetric Dirichlet prior is applied to all the elements of \mathbb{P} to ensure probability properties of P'_k s, $k = 1, \dots, K$. With total J possible read types, a corrected Perks prior, Dirichlet $(\frac{1}{J^2}, \frac{1}{J^2}, \dots, \frac{1}{J^2})$ is chosen for the multinomial parameters. The corrected Perks prior reduces the prior strength (concentration) by a factor proportional to the number of categories of the multinomial to ensure that the Bayesian estimator is preferred to maximum likelihood estimators for the parameters (Walley, 1996; de Campos and Benavoli, 2011). With an additional assumption that there is an equal chance of getting any P_k in \mathbb{P} , we constructed the following hierarchical Dirichlet mixture model:

$$\begin{aligned} Y_i | c_i, \mathbb{P} &\overset{indep}{\sim} Multinomial(m_i; P_{c_i}) \\ c_i | \mathbb{P} &\overset{iid}{\sim} Uniform\ Discrete\left(\frac{1}{K}\right) \end{aligned}$$

$$P_k \stackrel{iid}{\sim} \text{Dirichlet}\left(\frac{1}{J^2}, \frac{1}{J^2}, \dots, \frac{1}{J^2}\right)$$

where m_i indicates the total number of reads observed at the i^{th} position, i.e. $\sum_{j=1}^J y_i^j = m_i$. Component number K is some fixed unknown integer. Integrating the posterior density $\pi(c_1, \dots, c_N, \mathbb{P} | Y_1, \dots, Y_N)$ over \mathbb{P} , the marginal posterior for the assignments given reads on the sequences is

$$\pi(c_1, \dots, c_N | Y_1, \dots, Y_N) = \frac{1}{h(Y_1, \dots, Y_N)} \prod_{k=1}^K \frac{\prod_{j=1}^J \Gamma\left(\sum_{i=1}^N y_i^j \mathbf{1}_{\{c_i=k\}} + \frac{1}{J^2}\right)}{\Gamma\left(\sum_{i=1}^N m_i \mathbf{1}_{\{c_i=k\}} + \frac{1}{J}\right)}, \quad (2.1)$$

where $h(Y_1, \dots, Y_N)$ is the normalizing constant.

Furthermore, if both read counts and assignments are given for the entire sequence sample, we have

$$P_k | c_1, \dots, c_N, Y_1, \dots, Y_N \stackrel{indep}{\sim} \text{Dirichlet}(\alpha_k^1, \alpha_k^2, \dots, \alpha_k^J), \quad (2.2)$$

where $\alpha_k^j = \sum_{i=1}^N y_i^j \mathbf{1}_{\{c_i=k\}} + \frac{1}{J^2}$, for $j = 1, 2, \dots, J$; and $k = 1, 2, \dots, K$.

In the methodology section we will introduce a sequence of efficient Markov chain Monte Carlo (MCMC) procedures used to cluster the genome sequence positions and generate assignment labels c'_i s for each viral genome site. Notice that the posterior distribution (2.1) is defined for a fixed mixture component number K . One may choose K *ad hoc*, however, if the chosen K is smaller than the real number of mixture components, at least one resulting cluster contains members from multiple true clusters ; if the chosen K is too large, the clustering procedure can be infeasible due to the high dimensionality of most genome sequence data. At every iteration of the MCMC updating step, one coordinate or a class of coordinates will be altered into one of the K possible assignments. As K increases, the probability of assigning the correct label to each position decreases. Equation (2.1) naturally places an AIC-like penalty on non-empty clusters. It encourages empty groups by scaling the marginal posterior $\pi(c_1, \dots, c_N | Y_1, \dots, Y_N)$ by $[\Gamma(\frac{1}{J^2})]^J$. This shrinkage property allows our algorithm to start with a liberal upper bound of component number instead of the truth and naturally reduces it to a close upper bound of K . In Section 2.3 we

will introduce a tree-like MCMC step that provides the liberal upper bound and a block-MCMC procedure that produces a reasonably close upper bound of K . In Section 2.4 we will show through a simulation study that with the close upper bound of K , our algorithm correctly identifies the genomic regions with evolutionary changes.

2.2.3 Hellinger Distance

In order to capture significant evolutionary changes within the genomes of the viral populations, we need a measure for quantifying the changes. We chose an f-divergence, the Hellinger distance, H , to measure the similarity between two probability distributions (Hellinger, 1909). Under Lebesgue measure, for two probability density functions f and g , the squared Hellinger distance can be expressed as following:

$$H^2(f, g) = \frac{1}{2} \int \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx = 1 - \int \sqrt{f(x)g(x)} dx. \quad (2.3)$$

The Hellinger distance is a metric. The larger H is, the more different f and g are.

We prefer the Hellinger distance over relative entropy, the Kullback-Leibler divergence (KL), because symmetry is a desired property for the comparison of distributions. One can also use a symmetrised KL, such as the Jensen-Shannon divergence. We used the Hellinger distance to compare two marginal posterior distributions of the probability parameters given all cluster assignments and every read count, $P_{k_1}|c_1, \dots, c_N, Y_1, \dots, Y_N$, and $P_{k_2}|c_1, \dots, c_N, Y_1, \dots, Y_N$. The distance measures how similar the two allelic positions or same allelic position at two different time points are. Applying the squared measure (2.3) to two marginal posteriors with form (2.2), we have

$$H^2(P_{k_1}, P_{k_2}|c_1, \dots, c_N, Y_1, \dots, Y_N) = 1 - \frac{B(\vec{\beta}_{k_1, k_2})}{\sqrt{B(\vec{\alpha}_{k_1})B(\vec{\alpha}_{k_2})}}, \quad (2.4)$$

where

$$\begin{aligned} \vec{\alpha}_i &= (\alpha_i^1, \alpha_i^2, \dots, \alpha_i^J), \text{ for } i = k_1, k_2; \\ \vec{\beta}_{k_1, k_2} &= \left(\frac{\alpha_{k_1}^1 + \alpha_{k_2}^1}{2}, \frac{\alpha_{k_1}^2 + \alpha_{k_2}^2}{2}, \dots, \frac{\alpha_{k_1}^J + \alpha_{k_2}^J}{2} \right); \end{aligned}$$

$$B(a^1, a^2, \dots, a^J) = \frac{\prod_{j=1}^J \Gamma(a^j)}{\Gamma\left(\sum_{j=1}^J a^j\right)}.$$

To better visualize Hellinger distances for the viral data we further applied a monotonic transformation on H : $f(H) = \ln(1 - \ln(1 - H^2))$. With the definition (2.4) the Hellinger distance can then be transformed into

$$Ht(P_{k_1}, P_{k_2} | c_1, \dots, c_N, Y_1, \dots, Y_N) = \ln \left(1 - \ln \left(\frac{B(\vec{\beta}_{k_1, k_2})}{\sqrt{B(\vec{\alpha}_{k_1})B(\vec{\alpha}_{k_2})}} \right) \right). \quad (2.5)$$

Consider the toy example where three data collections, baseline (t_1), pre-treatment (t_2), and post-treatment (t_{3D}), were obtained (Figure 2.2). To see if the i th genomic site has been affected by the treatment, we compute the marginal posterior distributions for site i at all three time points: $\pi_i^{t_1}, \pi_i^{t_2}, \pi_i^{t_{3D}}$, perform pairwise comparison with the transformed Hellinger distance Ht , and check if the comparisons between the treated and non-treated populations, $Ht(\pi_i^{t_1}, \pi_i^{t_{3D}})$ & $Ht(\pi_i^{t_2}, \pi_i^{t_{3D}})$, are much greater than the variation within the untreated group, $Ht(\pi_i^{t_1}, \pi_i^{t_2})$.

2.3 Methodology

In order to perform the comparisons illustrated in Figure 2.2, we first need the group assignments c_1, c_2, \dots to compute the marginal posterior distributions. In general, we assume that the viral population was sampled and sequenced before and after the treatment. To see whether a genome site has been affected by the treatment, we cluster the genome sites, generate the assignment labels, derive its marginal posterior distribution for each site from each sample and compare the posteriors across time points and treatments. If a site shows significant change over time under treatment but not under control environment, it is identified as a substitution site due to treatment. The details of this procedure are described below.

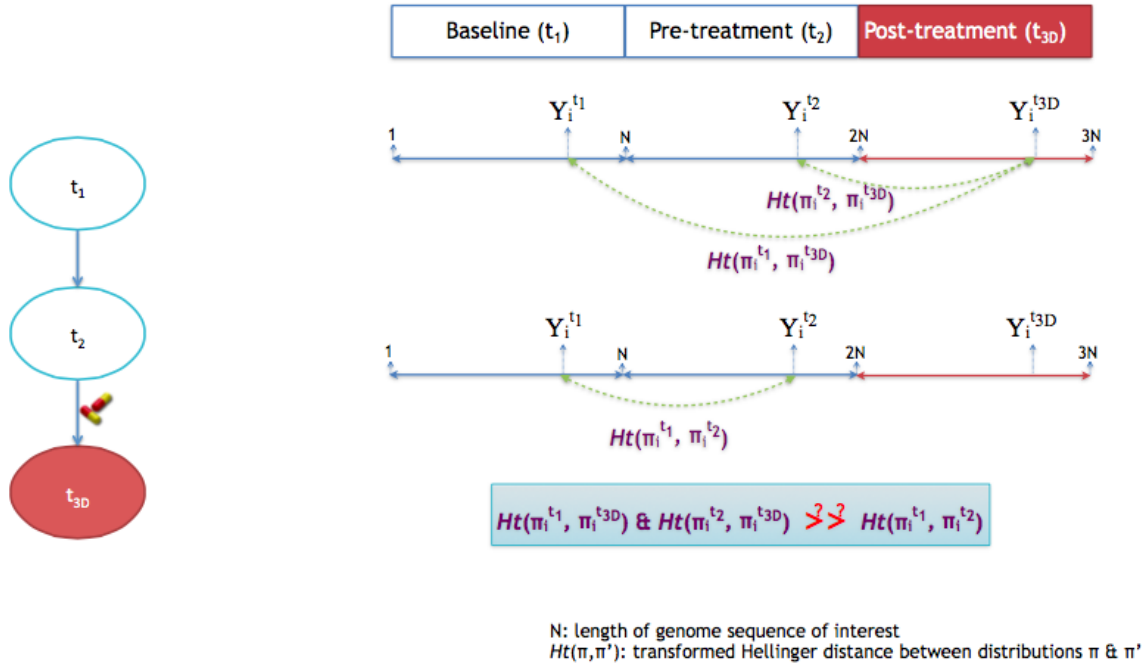


Figure 2.2: The transformed Hellinger distance is used to compare the marginal posterior distributions for the same genomic site across time. If the comparison between the treated and non-treated groups are much larger than the variation within drug-free environment, then the site is identified to be affected by the treatment.

	$Y_1^{t_1}$	$Y_2^{t_1}$	$Y_3^{t_1}$	$Y_4^{t_1}$	$Y_5^{t_1}$		$Y_1^{t_2}$	$Y_2^{t_2}$	$Y_3^{t_2}$	$Y_4^{t_2}$	$Y_5^{t_2}$		$Y_1^{t_3}$	$Y_2^{t_3}$	$Y_3^{t_3}$	$Y_4^{t_3}$	$Y_5^{t_3}$
<i>A</i>	8	0	0	0	0	<i>A</i>	10	0	0	0	0	<i>A</i>	11	0	0	0	0
<i>C</i>	0	0	6	1	0	<i>C</i>	0	0	9	2	0	<i>C</i>	0	0	10	0	0
<i>G</i>	0	0	0	0	0	<i>G</i>	0	0	0	0	0	<i>G</i>	0	0	0	0	0
<i>T</i>	0	7	0	6	8	<i>T</i>	0	10	0	9	5	<i>T</i>	3	9	0	7	8
<i>M</i>	0	0	0	0	0	<i>M</i>	0	0	0	0	0	<i>M</i>	0	0	1	0	1

$$\Downarrow$$

	$Y_1^{all\ t}$	$Y_2^{all\ t}$	$Y_3^{all\ t}$	$Y_4^{all\ t}$	$Y_5^{all\ t}$	$Y_6^{all\ t}$	$Y_7^{all\ t}$	$Y_8^{all\ t}$
<i>A</i>	18	0	0	0	0	11	0	0
<i>C</i>	0	15	0	1	2	0	10	0
<i>G</i>	0	0	0	0	0	0	0	0
<i>T</i>	0	0	46	6	9	3	0	8
<i>M</i>	0	0	0	0	0	0	1	1

$$\Downarrow$$

$$\begin{aligned}
Y_1^{all\ t} &= Y_1^{t_1} + Y_1^{t_2}, \\
Y_2^{all\ t} &= Y_3^{t_1} + Y_3^{t_2}, \\
Y_3^{all\ t} &= Y_2^{t_1} + Y_5^{t_1} + Y_2^{t_2} + Y_5^{t_2} + Y_2^{t_3} + Y_4^{t_3} + Y_5^{t_3}, \\
Y_4^{all\ t} &= Y_4^{t_1}, \\
Y_5^{all\ t} &= Y_4^{t_2}, \\
Y_6^{all\ t} &= Y_1^{t_3}, \\
Y_7^{all\ t} &= Y_3^{t_3}, \\
Y_8^{all\ t} &= Y_5^{t_3}.
\end{aligned}$$

Table 2.2: Toy example of joining and preprocessing three 5×5 data matrices. The first few columns in the joint data matrix (second row) are the consolidation of columns with single nucleotide read in the sampled data panels (first row). The remaining columns of the joint data matrix are the copies of non-homogeneous reads of the sample (first row). The detail of the consolidation process is described in the panel in the third row.

2.3.1 Preprocess

Continuing with the toy example in Figure 2.2, the first step is to combine the datasets collected at different time points and consolidate the invariant read sites (Table 2.2). Assume that in the toy example $J = 5$. The five possible reads are A, C, G, T, M, as in Table 2.1.

The three small tables in the first row of Table 2.2 show the read counts obtained at time points t_1 , t_2 , and t_3 ; the second row table shows the combined data of the first row produced by merging all the sites with a particular homogeneous read type. The first few columns of the joint data (second row table in Table 2.2) are the consolidation of columns with single read type A, C, G, T, M, respectively. The sites with multiple read types (non-invariant sites) are copied to joint data matrix after all the combined invariant sites

($Y_1^{all\ t}, Y_2^{all\ t}, Y_3^{all\ t}$ in the toy example). In particular, $Y_1^{all\ t}$ in the joint data matrix (second row in Table 2.2) is formed by merging columns $Y_1^{t_1}$ and $Y_1^{t_2}$. Similarly, $Y_2^{all\ t}, Y_3^{all\ t}$ are formed from the sites that have a homogenous read of C and T , respectively:

$$Y_2^{all\ t} = Y_3^{t_1} + Y_3^{t_2},$$

$$Y_3^{all\ t} = Y_2^{t_1} + Y_5^{t_1} + Y_2^{t_2} + Y_5^{t_2} + Y_2^{t_3} + Y_4^{t_3} + Y_5^{t_3}.$$

The following columns in the second row are

$$Y_4^{all\ t} = Y_4^{t_1}, Y_5^{all\ t} = Y_4^{t_2}, Y_6^{all\ t} = Y_1^{t_3}, \dots$$

The exact mapping is shown in the third row of the table. Note that this preprocessing step consolidates invariant sites and reduces the dimensionality of sequencing data without losing any significant information.

2.3.2 Processing

After preprocessing the read counts, a series of MCMC methods are implemented to cluster the geomic locations and obtain the assignment labels c'_i s (Figure 2.3).

The first step is a “top down” hierarchical clustering with 2-means initial states (*hierarchical SCMH*) based on a two-component Single Coordinate updating Metropolis Hastings algorithm (Fishman, 2005). Under the divisive hierarchical model, sibling nodes are mutually exclusive and complementary respect to their parent node. In the case that one child node is empty, that branch stops growing and its parent node is recorded as a leaf node. Eventually this branching process stops. A block Metropolis Hastings (*block MH*) step (Fishman, 2005; Robert and Casella, 2004) is then applied to the leaf nodes, each treated as a block. After assessing convergence (e.g. Geweke diagnostic (Geman, 1992)), T thinned-out iterations of the assignment labels are reserved. Finally, one run of a fixed scan Gibbs sampler (Geman and Geman, 1984) is implemented on the joint data with the reserved assignment labels as initial states (*Gibbs*). At every stage of a MCMC, a label proposal for each site is given according to the posterior likelihood for the joint sequence if the label is assigned.

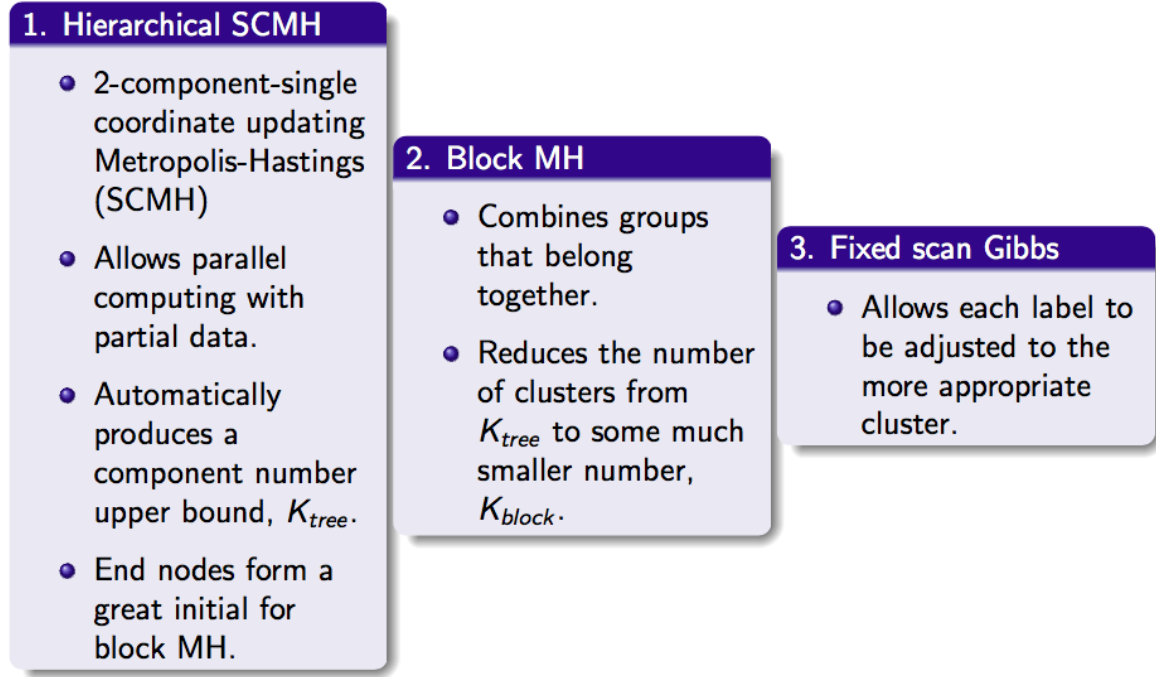


Figure 2.3: Three steps clustering procedure. It automatically produces an upper bound from K , assigns cluster labels to genomic sites at each time point, and allows parallel computing.

The divisive hierarchical clustering model allows us to avoid choosing a K , number of the mixture components. The total number of leaf nodes in the tree forms a reasonable upper bound for the number of mixture components regarding the entire joint data matrix. With sufficient number of observations at each site, m_i (*n.b.* most sequencing data have hundreds to thousands of sequencing reads), the Metropolis-Hasting splitting algorithm clusters correctly with probability one. This result is a direct consequence of the following theorem:

Theorem 2.1. *Suppose that $Y = [Y_1, Y_2]$, Y_1 and Y_2 are $J \times 1$ random read count vectors. $J \in \{2, 3, 4, \dots\}$. Further assume that $Y_i|c_i, \mathbb{P} \stackrel{indep}{\sim} \text{Multinomial}(m_i; P_{c_i})$, for $i = 1, 2$.*

If $c_1 = c_2 = 1$ and m_i 's are sufficiently large, then the marginal posterior likelihood ratio of assigning different labels over current state goes to zero almost surely,

$$i.e. LR = \frac{\pi(c_1 = 1, c_2 = 2|Y)}{\pi(c_1 = 1, c_2 = 1|Y)} \rightarrow 0, \text{ a.s.}$$

The proof of Theorem 2.1 can be found in the appendix.

Most genomes are lengthy, which results in high dimensional data. Direct application of multiple mixture component MCMC methods to this high dimensional data quickly becomes infeasible. Our hierarchical tree model enables efficient computing on this high dimensionality data. After the first split (at the root node) only a portion—typically only a small portion—of the original data set is analyzed at a time. With this reduced input size, each Markov chain converges much faster. At the root node, although the entire joint sequence is used, there are only two possible values for each assignment label. Thus the Markov chain typically reaches convergence inexpensively. The *hierarchical SCMH* step is also easily parallelized for high performance computing systems. As a result, we have an efficient clustering procedure that automatically produces a component number upper bound and initial assignments for the block MH step.

In practice, the MCMC tree usually splits the data into too many groups. The *block MH* step, however, allows clusters to combine. The binary hierarchical clustering process therefore does not require the true number of components to be representable by binary clusters. The shrinkage property of the marginal distribution (2.1) favors combining leaf nodes that belong to the same group. As a result of this natural penalty on non-empty groups, the number of distinct groups at the end of the *block MH* step is almost always much smaller than the total leaf number in the hierarchical tree. The *block MH* step in essence tunes the assignments for each genome site and reduces the total number of mixture components. As shown later in Section 2.4, our clustering algorithm with only the first two steps (*hierarchical SCMH* & *block MH*) produces reasonable results with slightly higher error rate, compared to the full algorithm.

Occasionally a few indices in some end nodes can be misplaced in the tree splitting step. Because all the indices in each leaf node are kept in the same cluster throughout the *block MH* process, those position indices do not get a chance to be moved to a different cluster. We solve this by adding a fixed scan Gibbs sampler step, *Gibbs*, that can modify the assignments for individual indices and move them to more appropriate clusters.

It is worth noting that when the preprocessed dataset is very length (i.e. many columns), even one scan of Gibbs can take a large amount of time to compute with standard methods. For this very reason, direct Gibbs sampler can become infeasible even if a the number of

mixture components K is provided. Furthermore, one must choose such a K ad hoc and risk either having a small mixture number that always misgroups elements from multiple true clusters together or risk having a large mixture number that might make computation infeasible. Further, most sequencing data sets are large. Direct Gibbs sampler might be infeasible even with a small K . Because the *hierarchical SCMH* step enables parallel computing and the *block MH* step works with a dataset of smaller dimension than the original, the computational cost is much lower in comparison to direct MCMC approaches. In the simulation study section (Section 2.4), we will compare the result and computation time using our three-step clustering approach to a direct Gibbs sampler with several choices of K , including the truth. We will show that our clustering is much more efficient, it outperforms the direct Gibbs sampler given the true K .

Alternatively, one can apply a Dirichlet process model to the joint dataset. A Dirichlet process model can be viewed as a Dirichlet mixture model with infinite number of components. With this framework, the point when cluster number stops growing depends heavily on the shrinkage power of the prior. Hence for a Dirichlet process model to work a more careful choice of prior is required. Intuitively, the computational time for the Dirichlet process is at best as good as a direct Gibbs sampler with the true K and a good initial state.

2.3.3 Postprocess

After implementing the three steps: *hierarchical SCMH*, *block MH*, and *Gibbs*, we obtain T running sets of assignment labels for the joint data. By reversing the preprocessing step (illustrated in Table 2.2) each genome position gets an assignment label for each time point from each Gibbs result. The posterior distribution per genome position per time point can now be computed. For each position i , we use the transformed Hellinger distance, Ht (Equation 2.5), to compare posterior distributions before and after treatment. Given two time points t_{k_1}, t_{k_2} , a collection of Hellinger distance values are obtained from the clustering result for each location i . We take the median of those distance values and denote it as $Ht(\pi_i^{t_{k_1}}, \pi_i^{t_{k_2}})$. In principle, one can use another measure of center instead of the median. We chose the median for its simplicity and straightforward interpretation. The summary

statistic of the Hellinger distance between treated and non-treated times for location i is denoted as $Ht(D_i)$. Large values in $Ht(D_i)$ indicate evolutionary changes in the viral genome. Those changes can be caused by the treatment or non-treatment related reasons, such as genetic drift and adaptation to the host. To distinguish between these potential causes of changes, we denote a summary statistic $Ht(N_i)$ for the comparison between time points without treatment.

Exact form of the statistics $Ht(D_i)$ and $Ht(N_i)$ depends on the experimental design. The basic idea is that, at genomic location i , $Ht(D_i)$ is the minimum change between the last treated time point and all pre-treatment times, while $Ht(N_i)$ is the maximum change among pairwise comparisons between untreated samples. At position i , if $Ht(D_i)$ is large, the last sampled population after treatment is significantly different from *all* samples before treatment; if $Ht(N_i)$ is large, *some* untreated sample is significantly different from *some* other untreated sample.

Intuitively, if the nucleotide read count distribution at site i has been affected by the treatment, $Ht(D_i)$ shall be *large*, relative to $Ht(N_i)$ and the comparisons for all the other sites that are not affected by treatment. How large is *large* will be determined by thresholding. For any given cutoff d , we define the following three sets:

$$\begin{aligned} S_1^d &= \{i : Ht(D_i) > d\}, \\ S_2^d &= \{i : Ht(N_i) > d \ \& \ Ht(N_i) > Ht(D_i)\}, \\ S_3^d &= S_1^d \setminus S_2^d. \end{aligned}$$

The first set, S_1^d , includes all the genomic locations that have large changes when comparing the treated and untreated groups. It is a *potential* set for substitutions. The second set concerns the large differences within the untreated group. It consists of all the genomic locations that have large variation which is not due to the treatment. The set S_2^d can be viewed as a *noise* set. Taking the set difference between the potential list and the noise list, the resulting set S_3^d is the list containing the *signals*. As d decreases, the sets S_1^d and S_2^d grow. We expect the growth of the sets accelerates as the cutoff moves from the signal to noise portion of the data. Therefore, the threshold d_0 used for inference is

determined by the tipping point where the size of S_2^d starts to increase dramatically, as d decreases.

To illustrate the derivation of $Ht(D_i)$ and $Ht(N_i)$, we continue with the toy example. It can be easily generalized to more than three time point collections, with duplicates, or with a treatment-control setup (see Sections 2.4 & 2.5). Return to the toy example (Figure 2.2), which shares the same experimental setup as the HIV-1 study in Section 2.5.1; the treatment is administrated after t_2 ; by time t_3 the viral population have completely responded (see Figure 2.4).

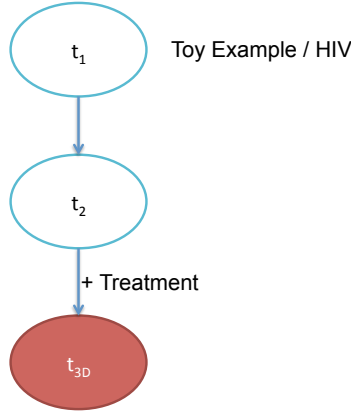


Figure 2.4: Illustration of the experimental setup the toy example and the HIV data. This setup includes two untreated populations (t_1, t_2) and one post-treatment population (t_{3D}). Observations are collected from each time point.

For each genome site i , we compare its marginal posterior distributions (Eq 2.2) at time t_1 , t_2 , and t_{3D} , denoted as $\pi_i^{t_1}$, $\pi_i^{t_2}$, $\pi_i^{t_{3D}}$, respectively, using the transformed Hellinger distance (Eq 2.5). Taking the clustering results from parallel chains, for a site i , we may define the summary statistics $Ht(D_i)$ and $Ht(N_i)$ as

$$Ht(D_i) = \min\{Ht(\pi_i^{t_1}, \pi_i^{t_{3D}}), Ht(\pi_i^{t_2}, \pi_i^{t_{3D}})\}, \quad Ht(N_i) = Ht(\pi_i^{t_1}, \pi_i^{t_2}), \quad \forall i. \quad (2.6)$$

If the change of read count distribution is caused by the treatment, the posterior distribution $\pi_i^{t_{3D}}$ ought to be much different from $\pi_i^{t_1}$ and $\pi_i^{t_2}$. Consequently, both $Ht(\pi_i^{t_1}, \pi_i^{t_{3D}})$ and $Ht(\pi_i^{t_2}, \pi_i^{t_{3D}})$ result in large values. Large $Ht(D_i)$ value guarantees that both $Ht(\pi_i^{t_1}, \pi_i^{t_{3D}})$ and $Ht(\pi_i^{t_2}, \pi_i^{t_{3D}})$ are large, it is therefore sufficient to look at $Ht(D_i)$.

Depending on the noise level of the data, the boundary of noise and signal portions of the data can be approximated by the curvature of noise set size function as the cutoff d decreases. The size of S_2^d is a step function of d . We suggest to plot the size of S_2^d against a decreasing series of cutoffs. We approximate the curvature of the plot by looking at the total segment length of every consecutive Δ number of steps. We then pick the point whose left Δ steps minus its right Δ steps is the largest as the optimal point. The default Δ value is 3 in our program. Larger Δ values lead to more coarse yet more robust approximation of the curvature. We also require a minimum length for the step on the left of the optimal point to guarantee that the noise set did not enlarge shortly after the value that is slightly greater than the cutoff. If the step on the left of the optimal point is shorter than the required minimum length, we move the optimal point to the left by one step and check the length of the next step. The final threshold is chosen to be the optimal point shifted to the left by the minimum length. This default sets the minimum length to be half ($\alpha = 0.5$) of the average length of the left Δ steps. A larger α leads to a more conservative result while a smaller α corresponds to a more liberal result. Both Δ and $\alpha = 0.5$ are introduced to mathematically capture the boundary of noise and signal sections of the data. In practice, we suggest users to verify the output by examining the site count plot directly.

2.4 Simulation Study

In this section, we use simulations to test our algorithm, with and without the *Gibbs* modification step, and compare its efficiency with direct Gibbs samplers.

Consider the following experiment setup for a viral population with genome length 300 nucleotides and five possible nucleotides at each genome site: A, C, G, T, M (see Figure 2.5).

The simulation mimics the experiment which first samples the RNA data twice before the administration of the treatment (t_1, t_2), then obtains a control group (t_3) and a treatment group (t_{3D}). For each genome site at time t_1, t_2, t_3 , the sequencing read count data are generated from multinomial distribution invariant in time and dependent on the genomic location. For the treated group t_{3D} , the evolved drug resistance sites 1, 21, 41, 61,

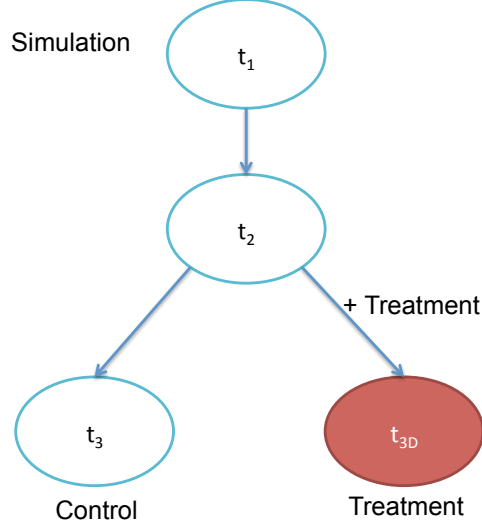


Figure 2.5: The experimental design used for the simulated test data. After two generations without treatment (t_1, t_2) the population is split into a control group t_3 and a treatment group t_{3D} . The treatment group is given a drug and allowed to evolve resistance to that drug for a few generations. The before treatment, control group, and treatment populations are sampled and sequenced.

81 are generated from alternative multinomial distributions, while the rest follow the same multinomial distributions as other time points.

As discussed in Section ??, we assume that the probability parameter for the nucleotide at each genomic location is sampled from a Dirichlet mixture model. For the sample without treatment, total 15 probability parameters, P_1, P_2, \dots, P_{15} , are used to generate the five possible reads: A, C, G, T, M. Five additional probability parameters, $P_{16}, P_{17}, \dots, P_{20}$, are introduced to generate the treated population. The total number of mixture component for the joint dataset is therefore 20.

At each genomic location i , the corresponding summary statistics are

$$\begin{aligned}
 Ht(D_i) &= \min\{Ht(\pi_i^{t_1}, \pi_i^{t_{3D}}), Ht(\pi_i^{t_2}, \pi_i^{t_{3D}})\} \\
 Ht(N_i) &= \max\{Ht(\pi_i^{t_1}, \pi_i^{t_2}), Ht(\pi_i^{t_1}, \pi_i^{t_3}), Ht(\pi_i^{t_2}, \pi_i^{t_3})\}
 \end{aligned} \tag{2.7}$$

For a simulated dataset (Figure 2.6), we analyzed the clustering results from both without (top two panels) and with (bottom two panels) the Gibbs modification step. The left two panels in Figure 2.6 show the sizes of sets S_1^d, S_2^d, S_3^d (potential, noise, signal) as threshold d decreases (zoom-in view). The dashed and the dotted lines are the thresholds

obtained using our method at two different choices of α ($\alpha = 0.5, 0.25$). At either threshold, without or with *Gibbs*, both S_1^d and S_3^d have five elements; S_2^d is empty. It is clear that a wide range of α parameter would produce different threshold, yet the same results here. The right two panels present the summary statistics $Ht(D_i)$ (small green circle) and $Ht(N_i)$ (blue cross) at each genome position. The horizontal dashed and dotted lines correspond to the thresholds chosen in the left panels. Above the dashed line, five large red circles highlights the $Ht(D_i)$ corresponding to the signal sites in right panels. They show much larger values than the rest and reveal clear separation between signals and noise. The potential set, noise set, and signal set are:

$$S_1^{d0} = \{1, 21, 41, 61, 81\}, \quad S_2^{d0} = \emptyset, \quad S_3^{d0} = S_1^{d0}. \quad (2.8)$$

Compared between without and with the *Gibbs* step, the inference results are equally good for this simulated dataset.

We repeat above data generating procedure and analysis 100 times. All of the 100 test sets precisely identified the five substitution sites with our full algorithm. Without the *Gibbs* step, 97 out of 100 simulated data sets were able to correctly identify the five signal sets robustly regarding to the choices of α parameter. The few tests that did not produce perfect result each included one false positive identification. The overall result with default parameter setting ($\Delta = 3$, $\alpha = 0.5$) is summarized in Table 2.3, along with results from direct Gibbs samplers with $K = 20, 40, 60, 80$. PR, FN, FP, FNP correspond to perfect results, only false negatives, only false positives, both false negatives and false positives, respectively. The numbers under each category are test counts. The *Gibbs* step improves the result obtained for the *block MH* step. The median of K derived from our algorithm is $K = 42$. Although the derived K is larger than the true K , the inference result is still 100% correct. This suggests that an overestimated K can still result in correctly identified signal sites. Similarly, direct Gibbs with larger K 's show perfect result for all 100 tests. As the goal is to compare posterior distributions not to obtain precise cluster number, an upper bound of mixture component number suffices. Note that when using our algorithm, we are

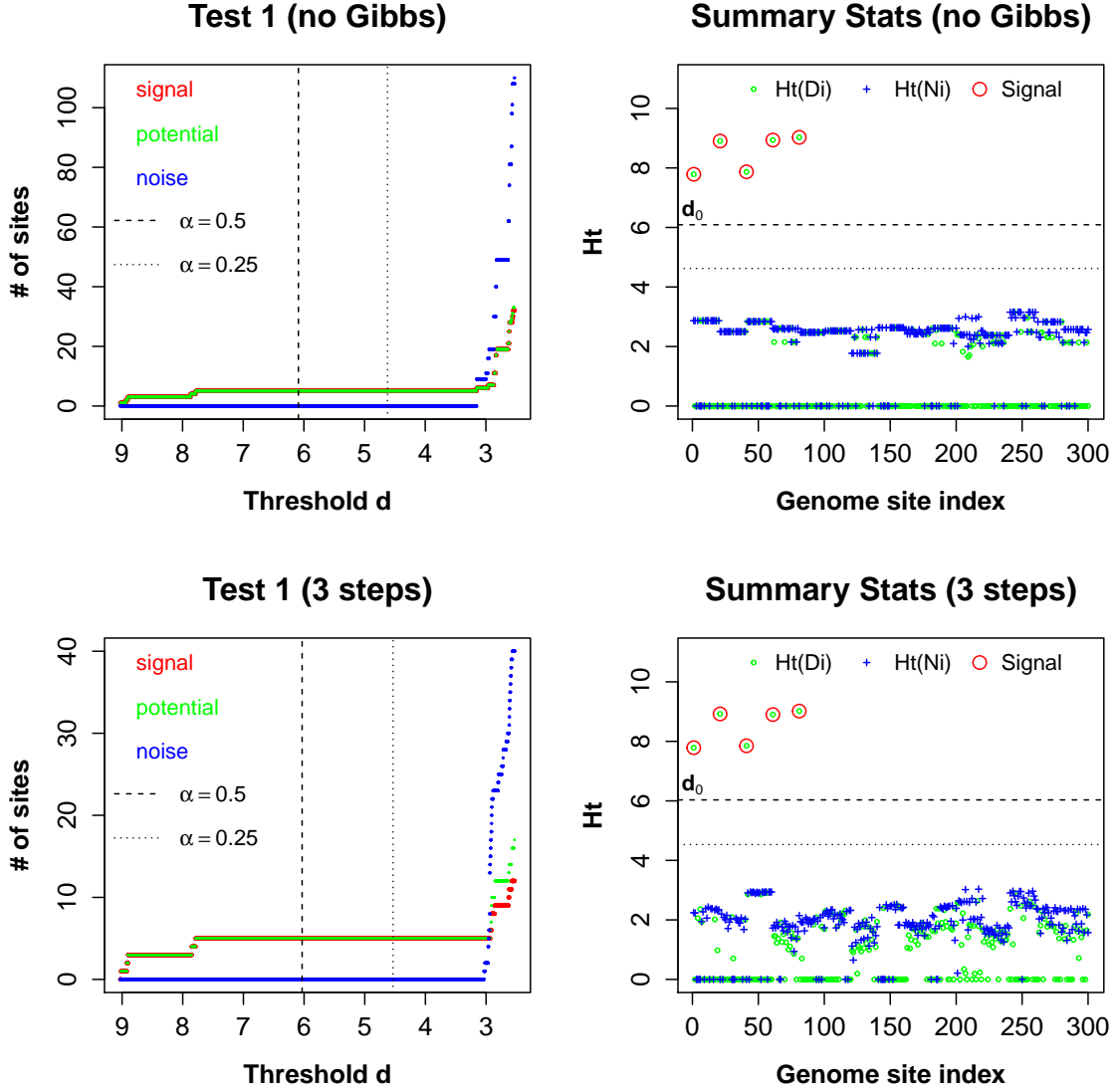


Figure 2.6: The result plots for Test 1 without (top) and with (bottom) the *Gibbs* step. The left two panels show the number of elements of S_1^d, S_2^d, S_3^d as the threshold d decreases with thresholds indicated in dashed ($\alpha = 0.5$) and dotted lines ($\alpha = 0.25$); the right two panels are the summary statistics plots with correspond thresholds to the left. The small green circles and blue pluses are the $Ht(D_i)$ values and $Ht(N_i)$ values, respectively. For genome site i that belongs to the signal set, its $Ht(D_i)$ value is highlighted in a large red circle. The five red circles on the top left correspond to the true substitution sites: 1, 21, 41, 61, 81. There is a clear separation between signals and the rest of the sites.

guarantee to be working with an appropriate upper bound of K . This insurance does not exist if K is chosen ad hoc, as would be required for a direct Gibbs.

Result	Our method		Direct Gibbs			
	w/ <i>Gibbs</i>	w/o <i>Gibbs</i>	K = 20	K = 40	K = 60	K = 80
PR	100	97	99	100	100	100
FN	0	0	1	0	0	0
FP	0	3	0	0	0	0
FNP	0	0	0	0	0	0

Table 2.3: Result comparison of our method, Gibbs with $K = 20$, Gibbs with $K = 40$, and Gibbs with $K = 60$ using a variety of thresholds. PR, FN, FP, FNP are the number of tests with perfect results, only false negatives, only false positives, both false negatives and false positives, respectively. All methods show good results. The *Gibbs* step improves the result over the *block MH* step alone.

To assess the efficiency of our sequential MCMC algorithm, we also compared the clustering time (measured in CPU time) of the methods discussed in Table 2.3 for the 100 synthetic data sets (see Table 2.4). For each test, the reported time under our method was the waiting time for the processing step (with or without *Gibbs*); the reported time for the direct Gibbs samplers was the time needed for one chain completing with corresponding K-means initial states. The CPU time is based on compute nodes including 122 blade servers, each with 8-cores 2.80 GHz Intel processors, 2×4M L2 cache (Model X5560), and 48GB memory for a total of 976 processing cores, two similar 8 core blades with 96 GB memory, and three more blades with 192 GB memory and 24 total cores. Summaries including means and standard deviations of computing time are recored for each method. As shown below, our method takes only a fraction of the time needed for the direct Gibbs, even when the true K is given. The variation of clustering time among the 100 tests is also much smaller using our algorithm. As K increases, the processing time for the direct Gibbs grows rapidly. It is worthy noted that the general computational issue with Gibbs sampler also affects the modification step in our algorithm. Our algorithm without the *Gibbs* step does not suffer the same issue. At the price of slight higher error rate, it produces reasonable results promptly.

2.5 Real Data Analysis

We applied the our algorithm to an HIV-1 dataset collected from three longitudinal plasma samples from tan individual participating in an anti-retroviral drug trial and an H1N1 viral

Clustering Time	Our method		Direct Gibbs			
	w/ <i>Gibbs</i>	w/o <i>Gibbs</i>	K = 20	K = 40	K = 60	K = 80
Min	55.78	41.49	95.94	289.3	576.6	959.1
1st Quartile	60.96	48.83	104.9	320.7	647.2	1083
Median	67.61	56.04	124.7	371.0	721.4	1178
Mean	67.72	55.42	148.8	394.8	766.8	1241
3rd Quartile	73.46	61.01	171.4	423.0	805.7	1368
Max	86.20	74.86	422.8	706.6	1467	1961
Standard Deviation	7.996	7.747	63.22	94.92	166.8	208.6

Table 2.4: Clustering time comparison in CPU time. For each test set, the corresponding process time of the direct Gibbs was that of a single Markov chain with K-means initial state for Gibbs sampler given a pre-chosen number of clusters. For our method, the corresponding process time records the total waiting time (in CPU time) needed for the processing step to finish 100 parallel Markov chains for each test set. The medians, means, and standard deviations here are from all 100 test sets. Our algorithm shows clear advantage in computational efficiency.

dataset produced by serially passaging the virus in kidney cells both in the presence and in the absence of an anti-viral drug.

2.5.1 Human immunodeficiency virus 1 (HIV-1)

As a "positive control", we applied our approach to an experimentally well characterized HIV-1 dataset (Jabara et al., 2011). Viral RNA was extracted from three longitudinal blood plasma samples taken from one individual infected with subtype B HIV-1, participating in a protease inhibitor (ritonavir) efficiency trial (Cameron et al., 1998). 454 sequencing was used to survey the genetic variation at the protease (*pro*) gene within the viral population. This population variation was surveyed twice, separated by six months, prior to ritonavir drug selection (t_1, t_2) and then once after the initiation of therapy (t_{3D}). HIV-1 is known to rapidly evolve resistance to ritonavir and several resistance mutations in the *pro* gene have already been identified and confirmed with *in vitro* experiments. Thus, if our method is efficacious we should recover these same sites through our analysis.

The length of the protease gene is 297. As in the toy example, there are five possible reads and the corresponding Hellinger summary statistics are

$$Ht(D_i) = \min\{Ht(\pi_i^{t_1}, \pi_i^{t_{3D}}), Ht(\pi_i^{t_2}, \pi_i^{t_{3D}})\}, \quad Ht(N_i) = Ht(\pi_i^{t_1}, \pi_i^{t_2}), \quad i = 1, \dots, 297. \quad (2.9)$$

The inference results based on clustering without and with the *Gibbs* step are shown in the top and bottom panels of Figure 2.7 respectively. The left two panels set size plot of S_1^d, S_2^d, S_3^d as the threshold d decreases (zoom-in); the right are the summary statistics plots. The thresholds according to two α parameter levels, $\alpha = 0.5$ (dashed line) and $\alpha = 0.5$ (dotted line), are plotted as well. In the summary statistics panels, the large red circles highlight the signal with default $\alpha (= 0, 5)$. Looking at the trajectory of the S_2^d size function in the top left panel, the default α appears to be too conservative, only site 245 was identified. The smaller α seems to be more appropriate, with which, three additional sites, 48, 55, 268, were added to the signal set. The two α levels considered in the full algorithm produce similar thresholds and the same inference result:

$$S_1^{d0} = \{48, 55, 243, 245, 250, 264, 268\}, \quad S_2^{d0} = \{70, 72, 168, 219, 289\}, \quad S_3^{d0} = S_1^{d0}. \quad (2.10)$$

Due to the noise level and limited time points of this dataset, the clustering without the *Gibbs* produced more conservative results. In the detected signal set, sites 48, 55, 245, 250, 268 correspond to known drug resistance mutations (Jabara et al., 2011). The other two sites identified, positions 243 and 264, both correspond to synonymous amino acid variation prior to treatment that disappeared post treatment. Meanwhile, the corresponding amino acids to sites 70, 72, 168, 219, 289, in the noise set S_2^{d0} , were identified as high variability in the study of genetic variation in the untreated environment (Jabara et al., 2011).

As shown above, our full method reveals not only the well-known drug resistant sites but also additional genomic locations that show clear evolutionary changes. The sites identified correspond to major resistance sites manually identified and curated from the literature in Jabara et al. (Jabara et al., 2011). Site 245, which corresponds to the major ritonavir resistant variant, V82A, shows a strong signal (Baldwin et al., 1995). Similar patterns are seen at other known resistance sites. These data suggest that our approach can identify biologically important genetic changes. We also note that in contrast to the earlier work, we were able to identify these sites with minimal *a priori* knowledge of genome. That is, we assumed nothing about where the genes were in the genome, if the change altered an amino

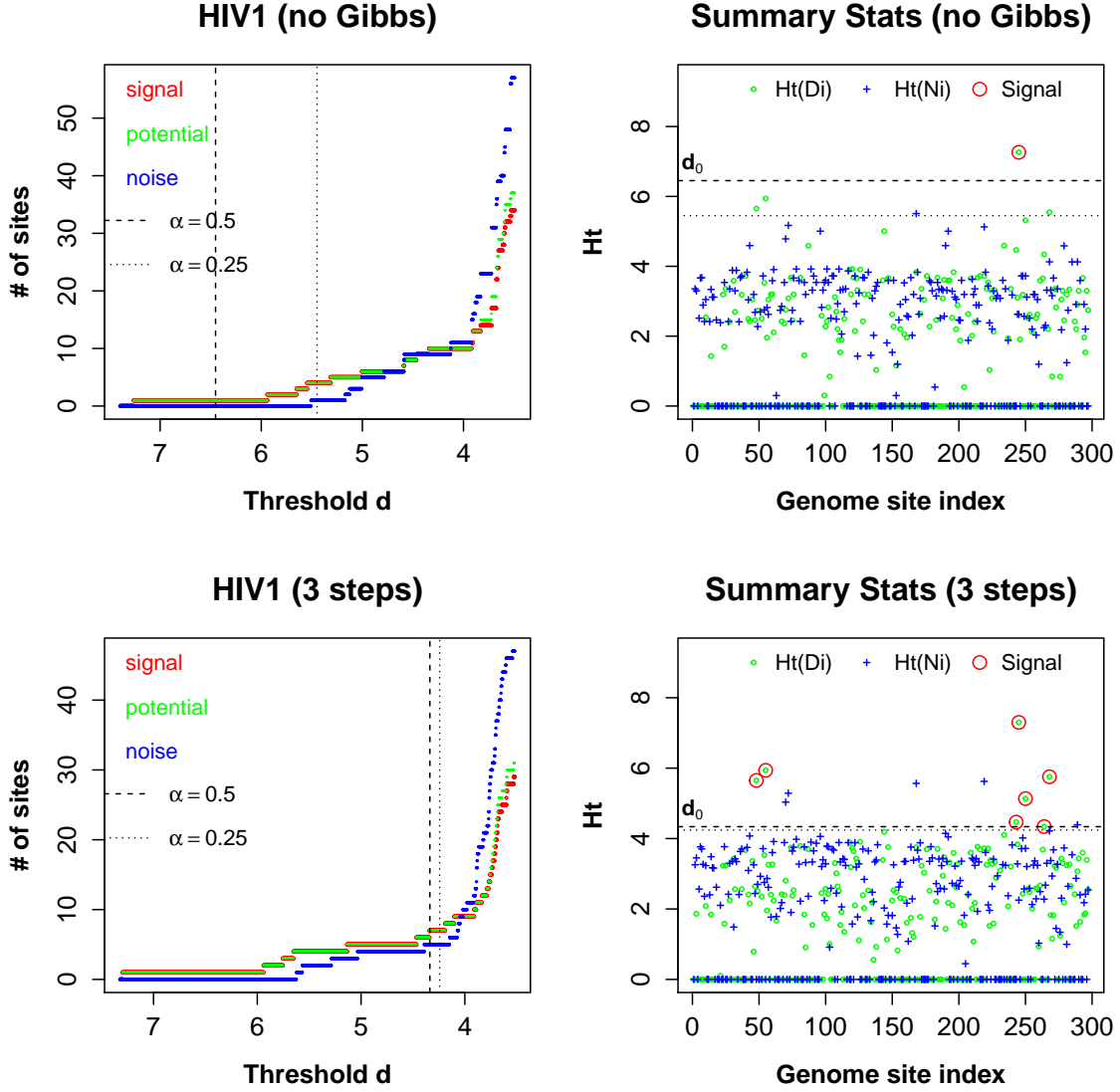


Figure 2.7: Results from the HIV-1 protease genome data set, without (top) and with (bottom) the *Gibbs* step. The left panels show the sizes of sets S_1^d, S_2^d, S_3^d as the threshold d decreases; the right panels are the summary statistics plots with signal identified (with default parameters) highlighted in red circles. Without the modification step, the default α appears to be too conservative. For the full algorithm result, either choice of α produced the same inference result with seven signal sites: 48, 55, 243, 245, 250, 264, 26, and five noise sites: 70, 72, 168, 219, 289.

acid or other structural element, etc. Thus, we can apply our method with confidence to viral genomes that are not nearly as well studied as HIV-1.

2.5.2 H1N1 Influenza A (IVA)

We applied our method to the whole-genome sequencing time series data of influenza A virus A/Brisbane/59/2007 strain (NIH Biodefence and Emerging Infectious Research Resources Repository NIAID, NIH; NR-12282; lot 58550257). The data were collected from multiple passages in the presence and absence of an inhibitor of neuraminidase, oseltamivir, for a total of two biological replicates (E1 & E2) (see Figure 2.1). At the end of each passage, whole-genome high throughput sequencing data were collected. The read counts are unbalanced between the two experiments, as the first replicate, E1, consistently had more reads than the second one. There are four possible nucleotides: A, C, G, T, i.e. $J = 4$.

This IVA strain consists of 8 segments: PB2 (2313 nucleotides (nts)), PB1 (2301 nts), PA (2303 nts), HA (1775 nts), NP (1396 nts), NA (1426 nts), M1/2 (1005 nts), and NS1/2 (869 nts). To reduce computational intensity, we examine each segment per replicate separately. Within each duplicate, we analyze the control and treatment groups over selected time points simultaneously. In particular, we choose five time points: 1, 3, 9, 12, and the end (13 and 18 for E1 and E2, respectively). As the first three passages were shared across groups, we analyze total of 8 time-samples, three of which were treated, for each biological replicate. Denote the 8 collection times as $t_1, t_2, t_3, t_4, t_5, t_{3D}, t_{4D}, t_{5D}$. The summary statistics are then formulated as

$$\begin{aligned} Ht(D_i) &= \min\{Ht(\pi_i^{t_1}, \pi_i^{t_{5D}}), Ht(\pi_i^{t_2}, \pi_i^{t_{5D}})\} \\ Ht(N_i) &= \max\{Ht(\pi_i^{t_1}, \pi_i^{t_2}), Ht(\pi_i^{t_1}, \pi_i^{t_j}), Ht(\pi_i^{t_2}, \pi_i^{t_j}), j = 3, 4, 5\} \end{aligned} \quad (2.11)$$

To allow additional response time for the drug, the comparisons to t_{3D} and t_{4D} are not directly included in $Ht(D_i)$.

Taking segment 6 as an example, we analyzed both replicates simultaneously, without and with the *Gibbs* step. The result plots for E1 and E2 are presented in Figures 2.8 & 2.9, respectively. Both replicates revealed site 833 (S6-822). The clear separation between $Ht(D_{S6-822})$ and the rest indicates that there is strong signal attributable to the treatment for S6-822.

To further investigate this finding, we plot the proportions of nucleotide type at each time point using the raw data (see Figure 2.10). The two panels on top are based on E1,

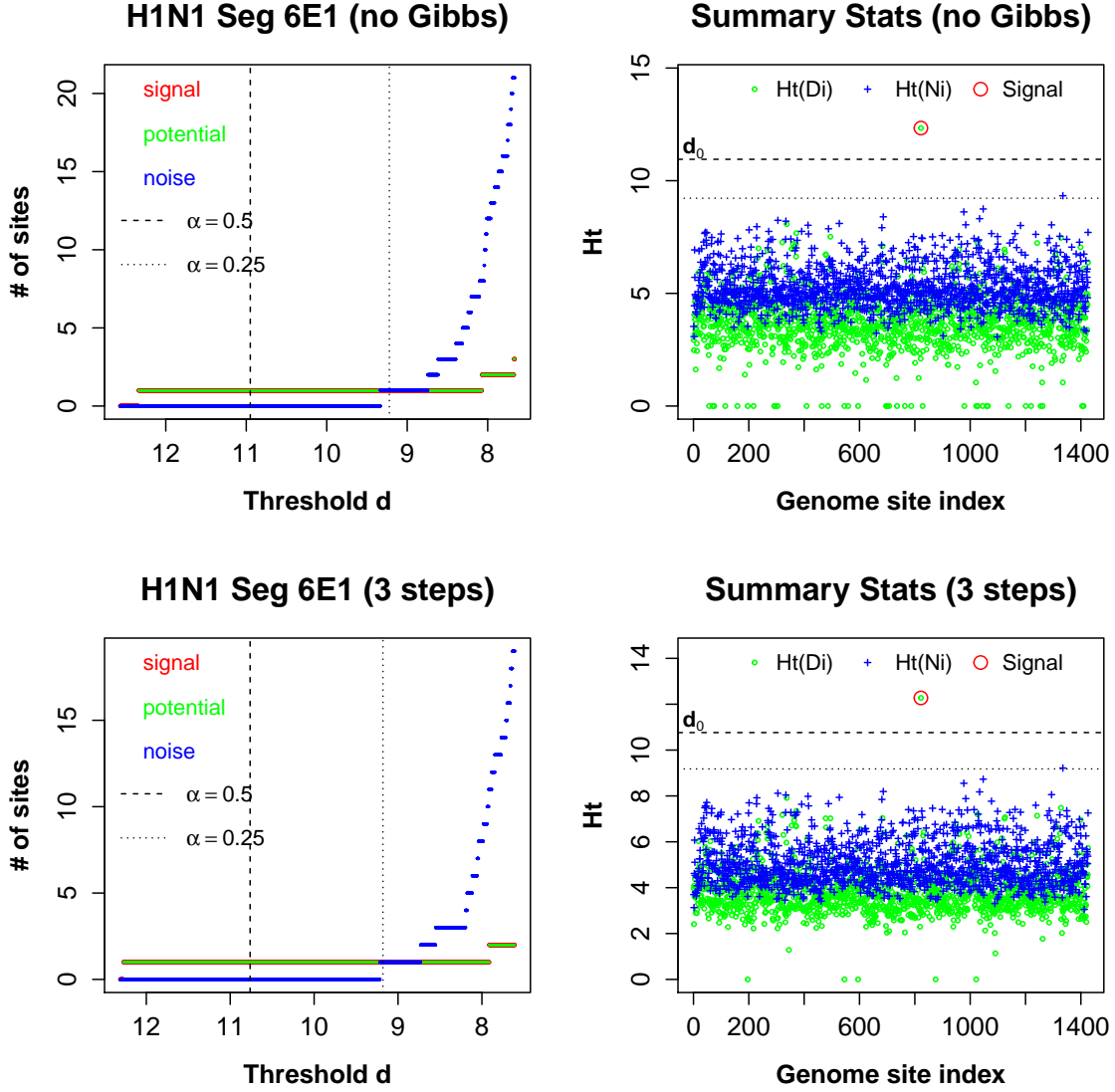


Figure 2.8: The results for H1N1 Seg6E1. Without (top panels) or with (bottom panels) the *Gibbs* step, our algorithm identified one signal site, site 822 (S6-822). It corresponds to a known oseltamivir-resistant mutation for H1N1. The inference result for H1N1 Seg6E1 is consistent even without the *Gibbs* step, and is robust to the choices of α parameter.

while the two on the bottom are based on E2. The controls are the left two panels; the treatment groups are the right two panels. The complete transition of the nucleotide type in the treated group and nearly no change in the control group indicates strong drug effect. The consistent behavior across replicates enables us to conclude that S6-822 is a substitution site due to the treatment. In fact, this is a known oseltamivir-resistant mutation, H274Y

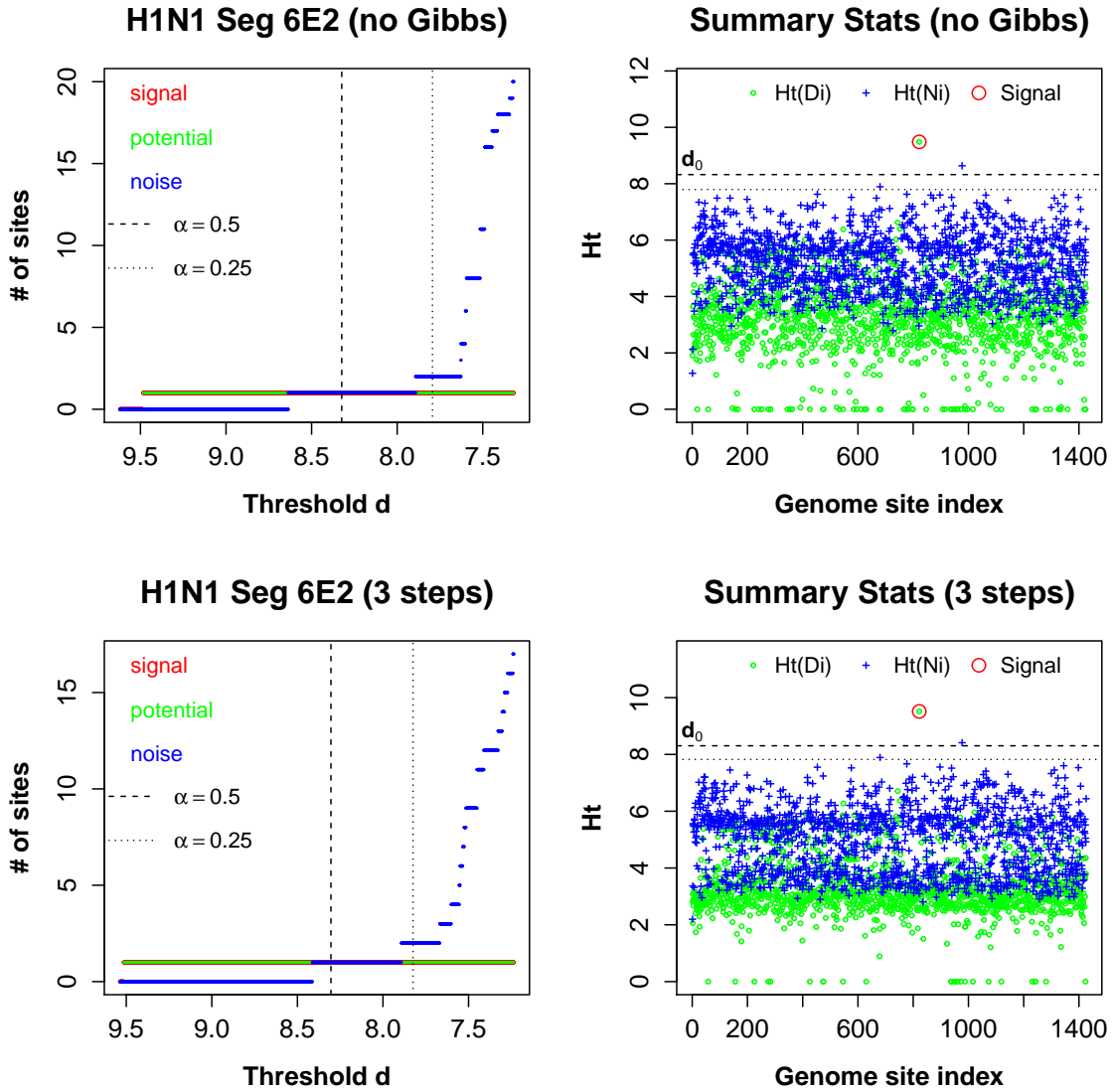


Figure 2.9: The result plots for H1N1 Seg6E2. Similar to Seg6E1, without (top panels) and with (bottom panels) the *Gibbs* step, our algorithm identified site 822 (S6-822). The inference result for H1N1 Seg6E1 is consistent even without the *Gibbs* step, and is robust to the choices of α parameter.

(Collins et al., 2008). The color tiles on the top of each panel indicates that the total read count at each time point varies.

In contrast, Segment 7 evinces a negative result. Figures 2.11 & 2.12 show result plots for E1 and E2, respectively. Top two panels were obtained without the *Gibbs* step. In the top right panel of Figure 2.11, site 503 (S7-503) was highlighted since it is above the threshold ($\Delta = 3$, $\alpha = 0.5$). However, it does not exceed the threshold in Figure 2.12. As

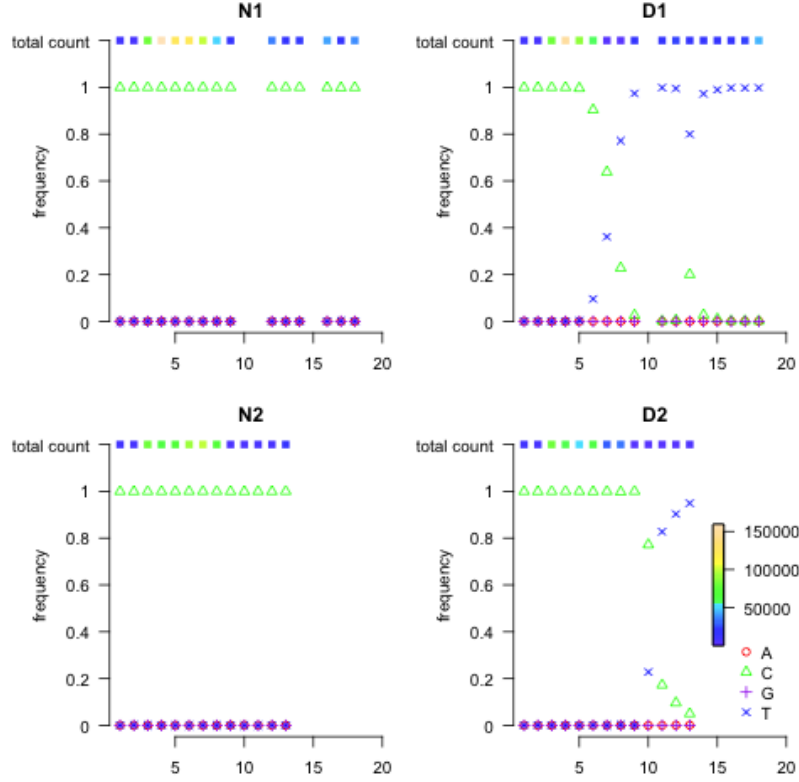


Figure 2.10: H1N1 nucleotide read count proportion and total count at position S6-822. The top and bottom rows are for Replicate I and Replicate II, respectively; the left and right panels are for control and treatment groups, respectively. For the treated groups, there is a complete transition from C to T due to the drug. The color tiles on the top of each panel indicates the total read count at each time point.

we are initially interested in substitution sites that are not replicate specific, we are only looking for signals observed in both biological replicates. The one red circle above threshold (top right panel of Figure 2.11) corresponds to S7-503. It appears to be a signal site based on E1, without the *Gibbs*. However, in Seg7E2, $Ht(D_{S7-503})$ is below the threshold (top right panel of Figure 2.12). Hence, we conclude that S7-503 is not a substitution site based on clustering result without the modification step. Conservatively, our conclusion is based on the intersection of the findings from each experiment. Of course it is possible each replicate could evolve along its own evolutionary path and hence differ between replicates. However, with the *Gibbs*, the signal set was adjusted to be empty for Seg7E1 (see bottom right panel of Figures 2.11 & 2.12) and we arrive at the same negative conclusion.

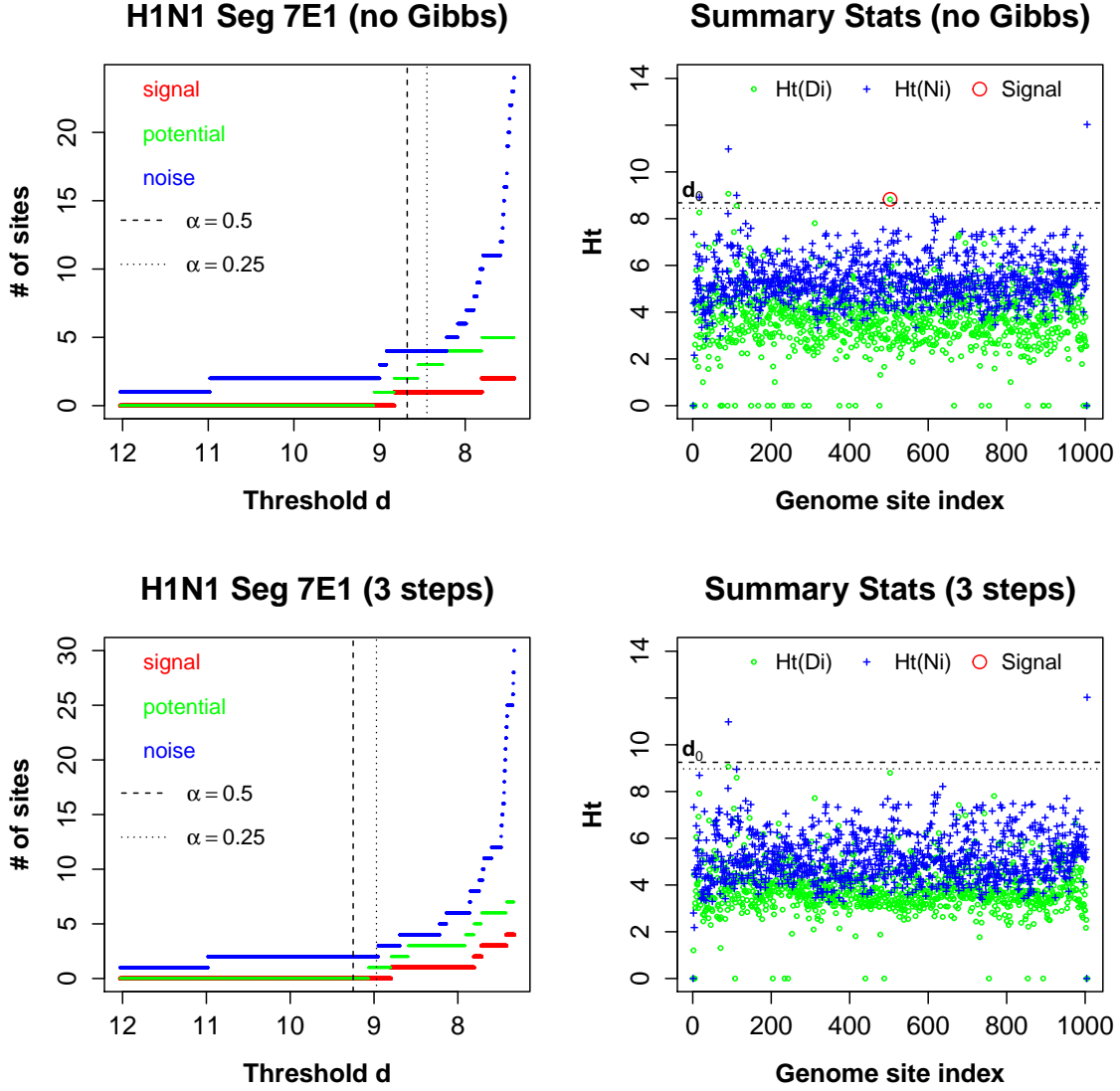


Figure 2.11: The results for H1N1 Seg7E1. Without the *Gibbs* step (top panels), S7-503 was highlighted. However, $Ht(D_{S7-503})$ does not exceed the threshold with the full algorithm (bottom right panel). Site 91 showed large summary statistic values $Ht(N_{S7-91})$ & $Ht(D_{S7-91})$ in both right panels. The control statistic value for S7-1005 is alarmingly high. We suspect that is the result of low alignment quality at the tail of the segment. The inference result for H1N1 Seg7E1 is robust to the choices of α parameter.

There is one site, S7-91, that consistently presented large $Ht(D_i)$ and $Ht(N_i)$ values across the two replicate. The proportions of its nucleotide read at each time point are shown in Figure 2.13. All four panels show complete transversion from G to C, with or without the treatment. The large values in $Ht(D_{S7-91})$ and $Ht(N_{S7-91})$ in both replicates precisely

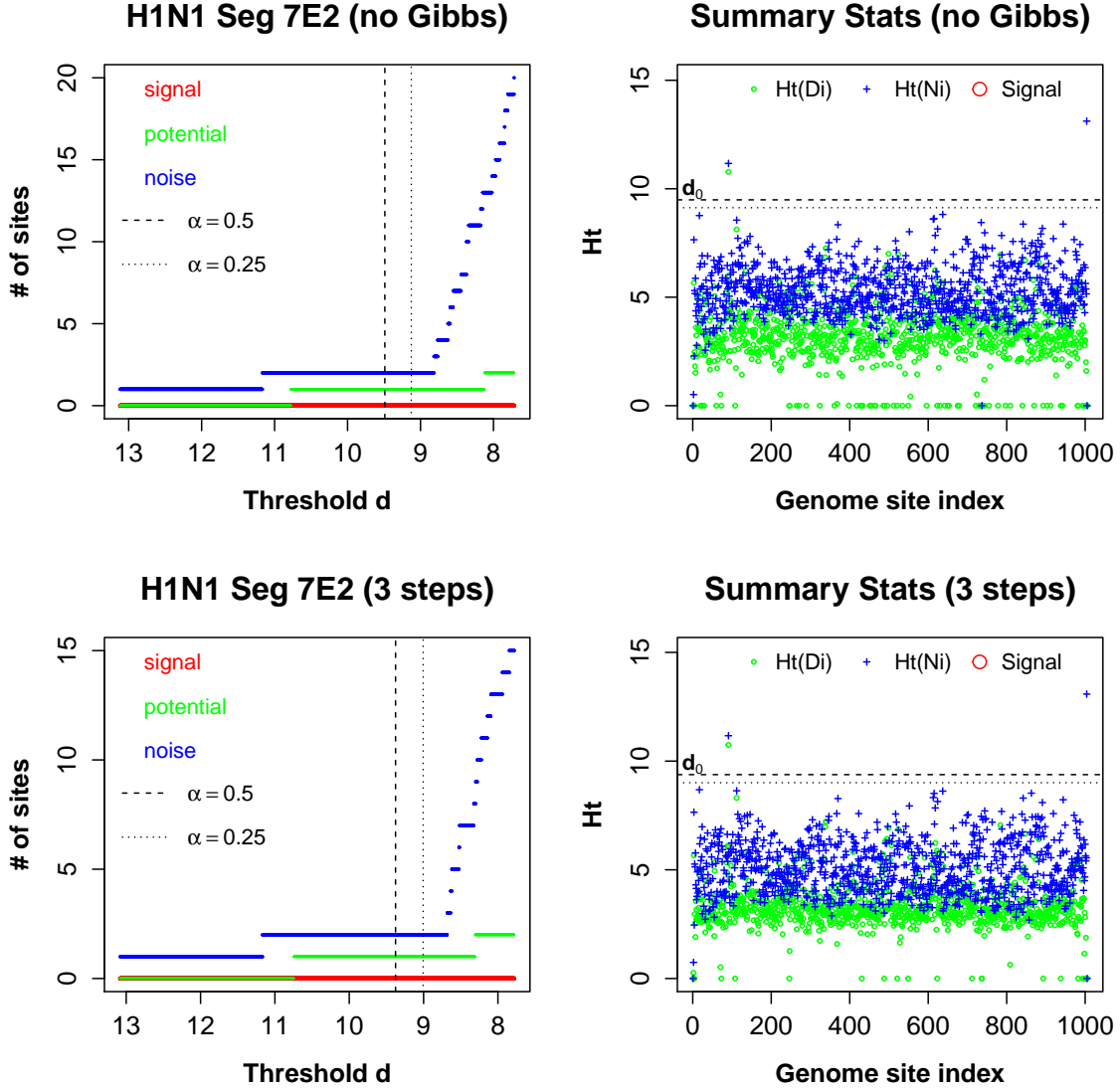


Figure 2.12: The result plots for H1N1 Seg7E2. The inference result for H1N1 Seg7E2 is consistent even without the *Gibbs* step, and is robust to the choices of α parameter. No site was identified as signal. Similar to Seg7E1, $Ht(N_{S7-91})$ & $Ht(D_{S7-91})$ exceeded the thresholds in both right panels. A site on the tail part of the segment, S7-1004, showed large control statistic values.

captures the read type switch that is likely due to genetic drift or adaptation to the host cells—not the drug.

With multiple biological replicates and many time point collections, the algorithm without the *Gibbs* step produced reasonable results with much greater computational efficiency. As discussed in Section 2.3, the modification step can take a long time due to the pitfall of

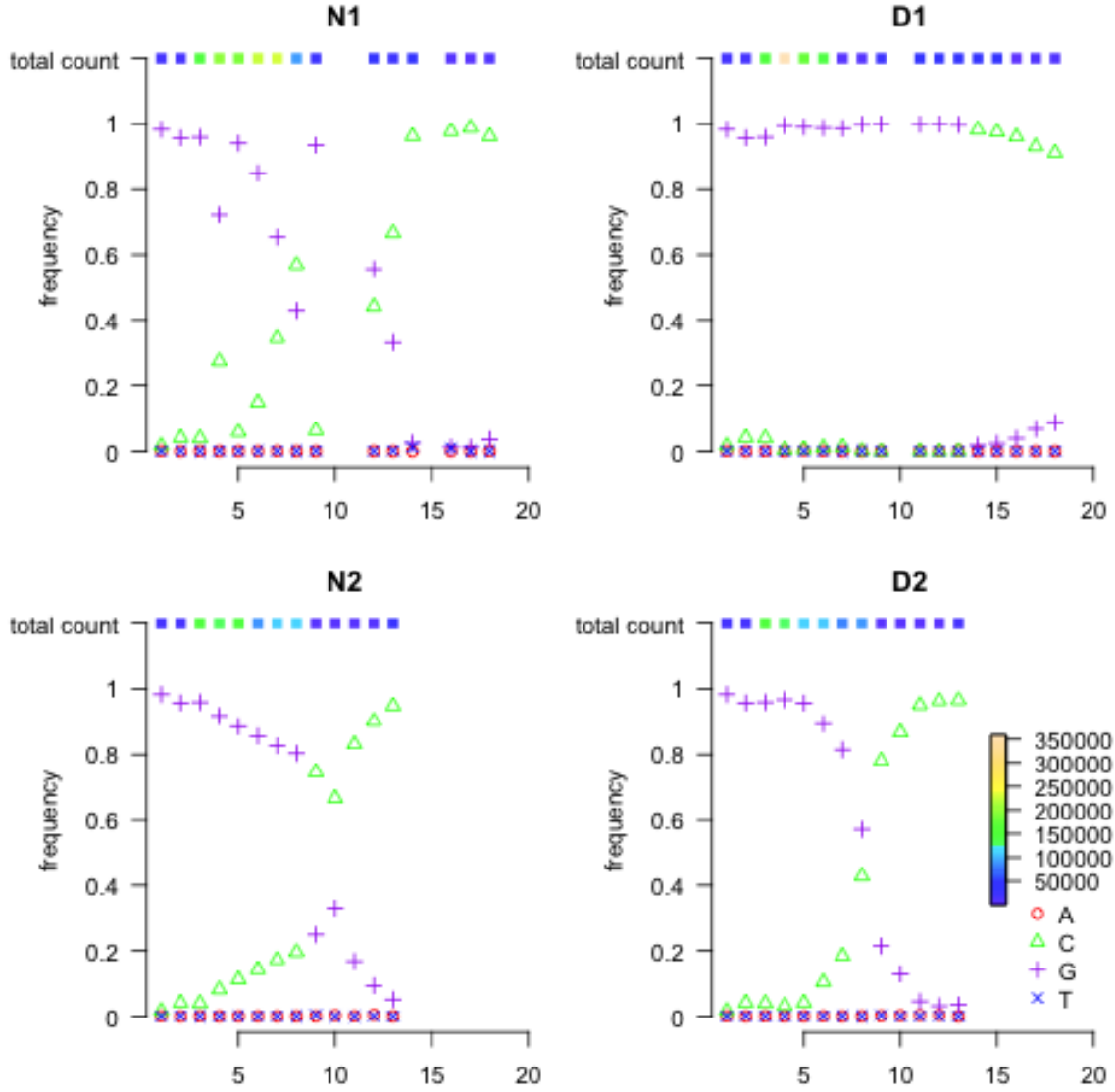


Figure 2.13: H1N1 nucleotide read count proportion and total count at position S7-91. All four panels show complete transversion from G to C.

standard Gibbs sampler, in which case, one may skip the *Gibbs* step at the cost of a slightly higher error rate.

For the rest of the segments, we performed the same analysis. Comparing the sites identified based on each biological replicate, we conclude that positions S6-822, and S8-80 are drug resistant sites. In addition, the following sites present evolutionary changes that are likely due to genetic drift or adaptation to the hosts: S1-2299, S1-2303, S3-2193, S4-1210, S5-1103, S7-91, S8-819. The complete list of signal and noise sets are provided in

Seg	E	Threshold d_0		Signal $S_3^{d_0}$		Noise $S_2^{d_0}$	
		w/ <i>Gibbs</i>	w/o <i>Gibbs</i>	w/ <i>Gibbs</i>	w/o <i>Gibbs</i>	w/ <i>Gibbs</i>	w/o <i>Gibbs</i>
1	1	8.756	10.59	1006, 1115, 1638, 1731, 1938, 2101	\emptyset	33, 281, 311, 404, 632, 839, 926, 1542, 1889, 2290, 2298, 2299, 2303	\emptyset
	2	9.401	9.483	\emptyset	\emptyset	2299, 2303	2299, 2303
2	1	9.782	9.86	\emptyset	\emptyset	224, 1118, 1499, 2066	224, 1118, 1499, 2066
	2	9.505	9.647	2099	2099	1036, 1675	1036, 1675
3	1	10.101	10.531	\emptyset	\emptyset	2193	2193
	2	9.655	8.937	174	174, 177	1850, 2193	79, 146, 153, 173, 176, 200, 203, 210, 1527, 1850, 2078, 2192, 2193, 2195
4	1	10.709	10.784	\emptyset	\emptyset	1210, 1394	729, 1210, 1394
	2	9.672	9.827	\emptyset	\emptyset	1210	638, 1210
5	1	10.352	9.98	\emptyset	300	1103, 1395	24, 389, 1103, 1395
	2	8.626	8.566	300	300	24, 389, 1103	24, 389, 1103
6	1	10.763	10.912	822	822	\emptyset	\emptyset
	2	8.304	8.319	822	822	977	977
7	1	9.249	8.57	\emptyset	503	91, 1005	17, 91, 112, 1005
	2	9.377	9.435	\emptyset	\emptyset	91, 1004	91, 1004
8	1	10.706	10.681	80	80, 848	385, 819, 848	385, 819, 848
	2	10.956	10.987	80	80	819	663, 819

Table 2.5: Result derived using Passages 1, 3, 9, 12, and the end time point. The table provides the thresholds and corresponding signal & noise sets for each segment according to each biological replicate with and without the *Gibbs* step. The sites identified as signal in both experiments are highlighted in red, the ones identified as noise in both experiments are highlighted in blue. The modification step took much more time comparing to the first two steps. The w/ *Gibbs* results was not finished for Seg1E1 and Seg2E1 in four weeks time with standard Gibbs. In comparison, the algorithm with or without the modification step produced similar final result after cross check the replicates.

Table 2.5. All of our findings are supported by the raw nucleotide read proportion plots (See Figures 2.10, 2.21, 2.22, 2.23, 2.24, 2.25, 2.28, 2.13, 2.29).

As mentioned earlier, for lengthy preprocessed data, direct Gibbs sampler can be computational expensive. Although our last clustering step, the *Gibbs*, only includes one scan of Gibbs sampling, it can also suffer from the same computational issue. For each segment, the one scan Gibbs sampler took at least two weeks real time on a high performance computing cluster while the first two steps only took a day or two. Our algorithm with and without the *Gibbs* step present consistent result generally. This is partly because that multiple time points were incorporated in the clustering procedure, yet only the last treated time was used to define the control statistics. Furthermore, because there are two biological replicates, taking the intersection of discoveries between the two helped to tease out some noise within each replicate. Skipping the modification step leads to a lightly higher error

rate, however, with multiple time points and replicates, the clustering result without the *Gibbs* leads to similar inference conclusion as basing on the full algorithm. When drawing inference without the modification step, we advise to double check the shift parameter α , as the default setting might not be the best for capturing the curvature of noise set size function.

We required that a "true" site be one that showed the same evolutionary behavior in both replicates. This approach is conservative as it requires that the same evolutionary path is taken by both viral populations, which may not necessarily be true. While at least two sites—including a known resistance variant—meet this strict criterion, there are several "signal" sites in each replicate that do not. These are potentially replicate specific adaptations. Moreover, it is possible the same amino acid can evolve through different nucleotide substitutions. For example, on segment 2 positions 31 and 32 evolved in Seg2E1 and Seg2E2 respectively. These neighboring changes both affect the amino acid lysine coded for by the 10th codon of the protein. Similar pattern is seen at sites 1004, 1005 on segment 7.

The first 12 passages of the dataset (Figure 2.14) were analyzed by Foll et al. from a population genetics and structural perspective (Foll et al., 2014). According to that study, the following sites are identified drug resistant: S2-32, S3-2193, S4-47, S4-1394, S6-581, S6-822, S7-146, S8-819; the sites with evolutionary changes without treatment are S2-1118, S4-1394, S5-1103, S5-1395.

For a fairer comparison to Foll et al, we applied our method to the joint data from Passages 1, 3, 9, 12 for both the control and treatment groups, i.e. $t_1, t_2, t_3, t_4, t_{3D}, t_{4D}$. The summary statistics used are

$$\begin{aligned} Ht(D_i) &= \min\{Ht(\pi_i^{t_1}, \pi_i^{t_{4D}}), Ht(\pi_i^{t_2}, \pi_i^{t_{4D}})\} \\ Ht(N_i) &= \max\{Ht(\pi_i^{t_1}, \pi_i^{t_2}), Ht(\pi_i^{t_1}, \pi_i^{t_j}), Ht(\pi_i^{t_2}, \pi_i^{t_j}), j = 3, 4\} \end{aligned} \quad (2.12)$$

The result from each biological replicate is shown in Table 2.6. Here we used the default parameters, $\Delta = 3$, $\alpha = 0.5$, and the summary results are for without the *Gibbs* step.

Taking the intersection of the findings from both replicates, we identify only S6-822 as a substitution site due to the treatment (although potentially S2-32, if the S2-31 and

Seg	E	d_0	$S_3^{d_0}$	$S_2^{d_0}$
1	1	8.713	\emptyset	33, 281, 404, 824, 839, 926, 1889, 2290, 2298, 2299 , 2303
	2	9.138	2072	311, 2299 , 2303
2	1	9.718	32	224, 1118, 1499, 2066
	2	9.43	31, 1663, 2099	1483, 1675, 2033
3	1	7.84	1613	89, 134, 173 , 174 , 176 , 177, 200 , 203 , 1556, 2078 , 2192 , 2193 , 2195
	2	7.944	180, 997	79, 146, 153, 173 , 174 , 176 , 200 , 203 , 210, 1527, 1850, 2078 , 2192 , 2193 , 2195
4	1	9.447	47, 1394	729
	2	9.274	\emptyset	638, 1210
5	1	8.855	\emptyset	24 , 389 , 1103 , 1395
	2	8.522	300	24 , 389 , 1103
6	1	8.496	581, 822	977 , 1047
	2	7.728	822	680, 977
7	1	9.099	146, 1005	\emptyset
	2	8.725	91, 1004	637
8	1	10.307	200, 819	385
	2	10.435	80	729, 819

Table 2.6: Our approach (w/o *Gibbs*) identifies only one true signal when data from only Passages 1, 3, 9, and 12 are used. The thresholds and corresponding signal & noise sets for each segment according to each biological replicate. The sites identified as signal in both experiments are highlighted in red, the ones identified as noise in both experiments are highlighted in blue. Fewer substitution sites were identified compared to previous table (see Table 2.5).

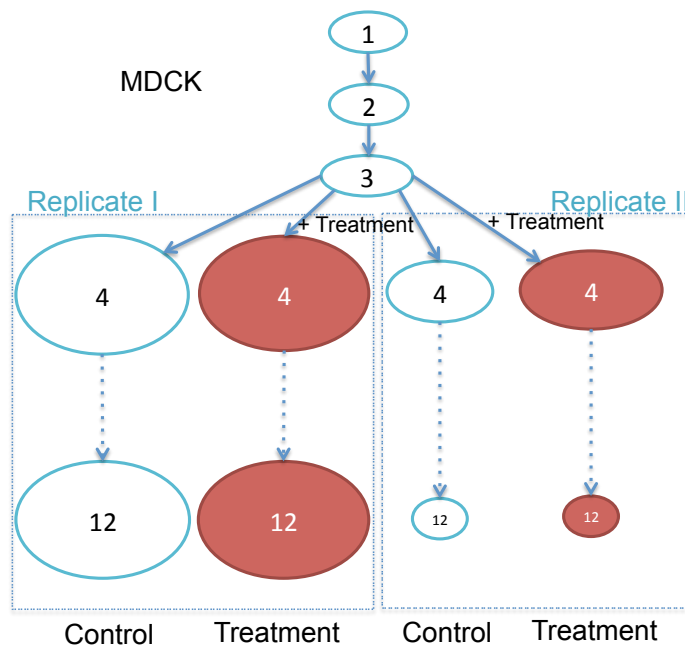


Figure 2.14: Only the first 12 passages were used in Foll et al (Foll et al., 2014). The complete dataset includes two biological replicates with one control group and one treatment group. Each ovals presents a passage. The colors white and red indicate absence and presence of the inhibitor. The sizes of the ovals indicate the average total read count per genome site. Note that the first replicate have much larger total reads than the second.

S2-32 sites of our analysis are treated as one). Several other sites, S1-2299, S2-2303, S3-173,174,176,200,203, 2078,2192,2193,2195, S5-24,389,1103, S6-977, were identified as locations with evolutionary changes not due to the treatment.

Intriguingly, most sites identified in Foll et al. (Table 2.6) appear in our analysis to only have signal in the first biological replicate. The exception, S6-822, has a strong signal in both replicates and regardless of end point generation analyzed (2.10). We speculate that the lack of consistent signal/false signal coming from the other sites is caused by the lower average read count per site for the second replicate compare to the first. The population genetic approach used in Foll et al. appears to be heavily influenced by the first replicate. This leads us to postulate that their result is adversely affected by the large imbalance in counts.

The additional sites identified in Table 2.5, S8-80, showed a more pronounced drug effect after the 12th passage in both replicate (Figures 2.21). We conclude that IVA may not have fully responded to the treatment by Passage 12, which was the final passage analyzed by

Foll et al. Thus our previous analysis, which included the last time point collection, is likely more reliable for the identification of substitution sites.

2.6 Discussion

We introduce a Dirichlet mixture model for detecting and clustering changes in allele frequencies in DNA or RNA sequence data from a population sampled at different time points. This annotation free approach is particularly useful for RNA viruses and other organisms where the secondary structure of the RNA can influence evolution in ways not predicted by standard molecular evolutionary analysis methods.

To identify significant changes in allele frequency, our clustering algorithm uses a combination of a hierarchical divisive clustering tree (*hierarchical SCMH*), a block Metropolis-Hasting (*block MH*), and a fixed scan Gibbs sampler (*Gibbs*) procedures. This approach does not require a prior distribution on the number of mixture components. The *hierarchical SCMH* step automatically produces an upper bound for the number of mixture components, K , and fine clusters for the *block MH* step. The hierarchical tree structure enables parallel computing and overcomes the computational difficulties any direct Markov chain Monte Carlo method presents. The *block MH* step improves the upper bound for K and combines similar clusters. Last but not least, the *Gibbs* step modifies the clustering result. The threshold for identifying substitution sites is derived based on the posterior distribution comparison for the time collections without treatment. It is chosen by examining the curvature in the graph of the number of members in the noise set instead of selecting an *ad hoc* cutoff.

With synthetic datasets we showed that our method with full clustering algorithm achieves results comparable to direct Gibbs without having to choose a K *ad hoc*. If the *Gibbs* step was skipped, we still achieved high perfect identification rate with even more gain in computation time. The *hierarchical SCMH* step enables parallel computing with partial data, which makes our clustering algorithm is much more efficient, even compared to the direct Gibbs with the true K .

The last cluster step of our algorithm, the *Gibbs*, can take a long time if the consolidated dataset is still very large. One may choose to skip this modification step at the price of a slightly higher error rate. It is advised to check the set size function plot and determine if the default parameters are appropriate.

As a positive control, we applied our method to a well described HIV-1 dataset. With minimal assumptions on gene annotation or the coding nature of the substitution, we successfully identified known drug resistance alleles previously reported (Jabara et al., 2011) and a list of sites with significant allelic changes within untreated population.

In the IVA dataset that motivated this study, we analyzed multiple time points and treatment-control simultaneously. We identified two sites, S6-822 & S8-80, with strong evidence of evolution in response to inhibitor treatment and six locations with high variability not due to the inhibitor. We compared our findings to a previous analysis of the same dataset based on a population genetic approach. Noticing that most of the sites identified using the latter method only appear in the biological replicate with larger sample size, we suspect that the population genetic based approach is biased due to this imbalance. Our algorithm performs analysis on each biological replicate individually first and then aggregate the results across replicates. Therefore, our inference technique is not sensitive to the unbalanced nature of the data.

In this chapter we have applied our method to high-throughput sequencing nucleotide read count data. It can also be applied to other count data, such as amino acids. As the model requires minimum assumption, it can be broadly applied. For example, this approach can be used to identify evolved sites in non-coding regions of the genome such as the regulator regions of genes or in RNA genes such as ribosomal RNA and other long non-coding RNAs.

2.7 Appendix

2.7.1 Proof of Theorem 2.1

Suppose there are only two genome positions to be clustered, $Y = [Y_1, Y_2]$. If they share the same probability parameter, then the likelihood of the two share the same parameter

is one when the numbers of observations at the two sites m_1 and m_2 are large. Fix a $J \in \{2, 3, 4, \dots\}$. Without loss of generality, assume $c_1 = c_2 = 1$, and then the marginal posterior likelihood ratio of splitting the two over current state on the log scale is the following:

$$\begin{aligned}
LR &= \log(\pi(c_1 = 1, c_2 = 2|Y)) - \log(\pi(c_1 = 1, c_2 = 1|Y)) \\
&= \sum_{j=1}^J \left[\log \Gamma(y_1^j + J^{-2}) + \log \Gamma(y_2^j + J^{-2}) \right] - \log \Gamma(m_1 + J^{-1}) - \log \Gamma(m_2 + J^{-1}) \\
&\quad - \sum_{j=1}^J \log \Gamma(y_1^j + y_2^j + J^{-2}) + \log \Gamma(m_1 + m_2 + J^{-1})
\end{aligned}$$

Claim: $LR \rightarrow -\infty$ a.s.

Proof. Recall that Stirling's formula provides the following approximation:

$$\log \Gamma(z) \approx \frac{1}{2} \log(2\pi) - \frac{1}{2} \log z + z \log z - z$$

Therefore,

$$\begin{aligned}
LR &\approx \sum_{j=1}^J \left[\frac{1}{2} \log(2\pi) - \frac{1}{2} (y_1^j + J^{-2}) - \frac{1}{2} (y_2^j + J^{-2}) + \frac{1}{2} (y_1^j + y_2^j + J^{-2}) + (y_1^j + J^{-2}) \log(y_1^j + J^{-2}) \right. \\
&\quad \left. + (y_2^j + J^{-2}) \log(y_2^j + J^{-2}) - (y_1^j + y_2^j + J^{-2}) \log(y_1^j + y_2^j + J^{-2}) - J^{-2} \right] - \frac{1}{2} \log(2\pi) \\
&\quad + \frac{1}{2} \log(m_1 + J^{-1}) + \frac{1}{2} \log(m_2 + J^{-1}) - \frac{1}{2} \log(m_1 + m_2 + J^{-1}) - (m_1 + J^{-1}) \log(m_1 + J^{-1}) \\
&\quad - (m_2 + J^{-1}) \log(m_2 + J^{-1}) + (m_1 + m_2 + J^{-1}) \log(m_1 + m_2 + J^{-1}) + J^{-1} \\
&= \frac{J-1}{2} \log(2\pi) + \sum_{j=1}^J \left[\left(y_1^j + J^{-2} - \frac{1}{2} \right) \log(y_1^j + J^{-2}) + \left(y_2^j + J^{-2} - \frac{1}{2} \right) \log(y_2^j + J^{-2}) \right. \\
&\quad \left. - \left(y_1^j + y_2^j + J^{-2} - \frac{1}{2} \right) \log(y_1^j + y_2^j + J^{-2}) \right] + \left(m_1 + m_2 + J^{-1} - \frac{1}{2} \right) \log(m_1 + m_2 + J^{-1}) \\
&\quad - \left(m_1 + J^{-1} - \frac{1}{2} \right) \log(m_1 + J^{-1}) - \left(m_2 + J^{-1} - \frac{1}{2} \right) \log(m_2 + J^{-1})
\end{aligned}$$

Under null hypothesis that Y_1 and Y_2 follow the same distribution, i.e. they share the same probability parameter. Denote the common probability parameter as $P =$

(p^1, \dots, p^J) . Then the normal approximation of the multinomial random variables are

$$y_i^j \approx m_i p^j + \sqrt{m_i} z_i^j + \mathcal{O}p(\sqrt{m_i}), \text{ for } i = 1, 2; j = 1, \dots, J,$$

where z_i^j 's are standard normal random variables and $\sum_{j=1}^J z_i^j = 0$ for $i = 1, 2$.

Hence,

$$\begin{aligned} & LR \\ \approx & \frac{J-1}{2} \log(2\pi) + \left(m_1 + m_2 + J^{-1} - \frac{1}{2}\right) \log(m_1 + m_2 + J^{-1}) - \left(m_1 + J^{-1} - \frac{1}{2}\right) \log(m_1 + J^{-1}) \\ & - \left(m_2 + J^{-1} - \frac{1}{2}\right) \log(m_2 + J^{-1}) + \sum_{j=1}^J \left[\left(m_1 p^j + \sqrt{m_1} z_1^j + J^{-2} - \frac{1}{2}\right) \log(m_1 p^j + \sqrt{m_1} z_1^j + J^{-2}) \right. \\ & + \left(m_2 p^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right) \log(m_2 p^j + \sqrt{m_2} z_2^j + J^{-2}) \\ & \left. - \left(m_1 p^j + \sqrt{m_1} z_1^j + m_2 p^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right) \log(m_1 p^j + \sqrt{m_1} z_1^j + m_2 p^j + \sqrt{m_2} z_2^j + J^{-2}) \right] \\ = & \frac{J-1}{2} \log(2\pi) + \left(m_1 + m_2 + J^{-1} - \frac{1}{2}\right) \left[\log(m_1 + m_2) + \log\left(1 + \frac{J^{-1}}{m_1 + m_2}\right) \right] \\ & - \left(m_1 + J^{-1} - \frac{1}{2}\right) \left[\log m_1 + \log\left(1 + \frac{J^{-1}}{m_1}\right) \right] - \left(m_2 + J^{-1} - \frac{1}{2}\right) \left[\log m_2 + \log\left(1 + \frac{J^{-1}}{m_2}\right) \right] \\ & + \sum_{j=1}^J \left\{ \left(m_1 p^j + \sqrt{m_1} z_1^j + J^{-2} - \frac{1}{2}\right) \left[\log(m_1 p^j) + \log\left(1 + \frac{\sqrt{m_1} z_1^j + J^{-2}}{m_1 p^j}\right) \right] \right. \\ & + \left(m_2 p^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right) \left[\log(m_2 p^j) + \log\left(1 + \frac{\sqrt{m_2} z_2^j + J^{-2}}{m_2 p^j}\right) \right] \\ & \left. - \left((m_1 + m_2) p^j + \sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right) \left[\log((m_1 + m_2) p^j) \right. \right. \\ & \left. \left. + \log\left(1 + \frac{\sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2}}{(m_1 + m_2) p^j}\right) \right] \right\} \\ = & \frac{J-1}{2} \log(2\pi) + \frac{J-1}{2} \log\left(\frac{1}{m_1} + \frac{1}{m_2}\right) + \left(m_1 + m_2 + J^{-1} - \frac{1}{2}\right) \log\left(1 + \frac{J^{-1}}{m_1 + m_2}\right) \\ & - \left(m_1 + J^{-1} - \frac{1}{2}\right) \log\left(1 + \frac{J^{-1}}{m_1}\right) - \left(m_2 + J^{-1} - \frac{1}{2}\right) \log\left(1 + \frac{J^{-1}}{m_2}\right) \\ & + \sum_{j=1}^J \left\{ \left(m_1 p^j + \sqrt{m_1} z_1^j + J^{-2} - \frac{1}{2}\right) \log\left(1 + \frac{\sqrt{m_1} z_1^j + J^{-2}}{m_1 p^j}\right) \right. \\ & + \left(m_2 p^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right) \log\left(1 + \frac{\sqrt{m_2} z_2^j + J^{-2}}{m_2 p^j}\right) \\ & \left. - \left((m_1 + m_2) p^j + \sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right) \log\left(1 + \frac{\sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2}}{(m_1 + m_2) p^j}\right) \right\} \end{aligned}$$

Note that, in general, by L'Hopital's rule, as $m_i \rightarrow \infty$,

$$\sqrt{m_i} \log \left(1 + \frac{\sqrt{m_i} z_i^j + J^{-2}}{m_i p^j} \right) = \frac{\log \left(1 + \frac{\sqrt{m_i} z_i^j + J^{-2}}{m_i p^j} \right)}{1/\sqrt{m_i}} \rightarrow \frac{z_i^j}{p^j},$$

for $i = 1, 2$; $j = 1, \dots, J$.

Under the assumption that m_1 and m_2 are increasing at the same rate, let $m_1 = m$ and $m_2 = cm$, for some $c > 0$. Then as $m \rightarrow \infty$,

$$\begin{aligned} & (\sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j) \log \left(1 + \frac{\sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2}}{(m_1 + m_2) p^j} \right) \\ &= (z_1^j + \sqrt{c} z_2^j) \sqrt{m} \log \left(1 + \frac{\sqrt{m} (z_1^j + \sqrt{c} z_2^j) + J^{-2}}{(1+c) m p^j} \right) \\ &\rightarrow \frac{(z_1^j + \sqrt{c} z_2^j)^2}{(1+c) p^j}, \text{ for } j = 1, \dots, J. \end{aligned}$$

Therefore, as $m \rightarrow \infty$, the log likelihood ratio

$$\begin{aligned} & LR \\ &\approx \frac{J-1}{2} \log(2\pi) + \frac{J-1}{2} \log \frac{1+c}{cm} + \left(m(1+c) + J^{-1} - \frac{1}{2} \right) \log \left(1 + \frac{J^{-1}}{m(1+c)} \right) \\ &\quad - \left(m + J^{-1} - \frac{1}{2} \right) \log \left(1 + \frac{J^{-1}}{m} \right) - \left(cm + J^{-1} - \frac{1}{2} \right) \log \left(1 + \frac{J^{-1}}{cm} \right) \\ &\quad + \sum_{j=1}^J \left\{ \left(mp^j + \sqrt{m} z_1^j + J^{-2} - \frac{1}{2} \right) \log \left(1 + \frac{\sqrt{m} z_1^j + J^{-2}}{mp^j} \right) \right. \\ &\quad + \left(cmp^j + \sqrt{cm} z_2^j + J^{-2} - \frac{1}{2} \right) \log \left(1 + \frac{\sqrt{cm} z_2^j + 1/25}{cmp^j} \right) \\ &\quad \left. - \left((1+c)mp^j + \sqrt{m} z_1^j + \sqrt{cm} z_2^j + J^{-2} - \frac{1}{2} \right) \log \left(1 + \frac{\sqrt{m} z_1^j + \sqrt{cm} z_2^j + J^{-2}}{(1+c)mp^j} \right) \right\} \\ &\rightarrow -\infty \end{aligned}$$

Therefore, Y_1 and Y_2 have the same cluster label almost surely.

□

2.7.2 Additional IVA Ht plots

In this subsection, we provide the Ht plots for some additional segments.

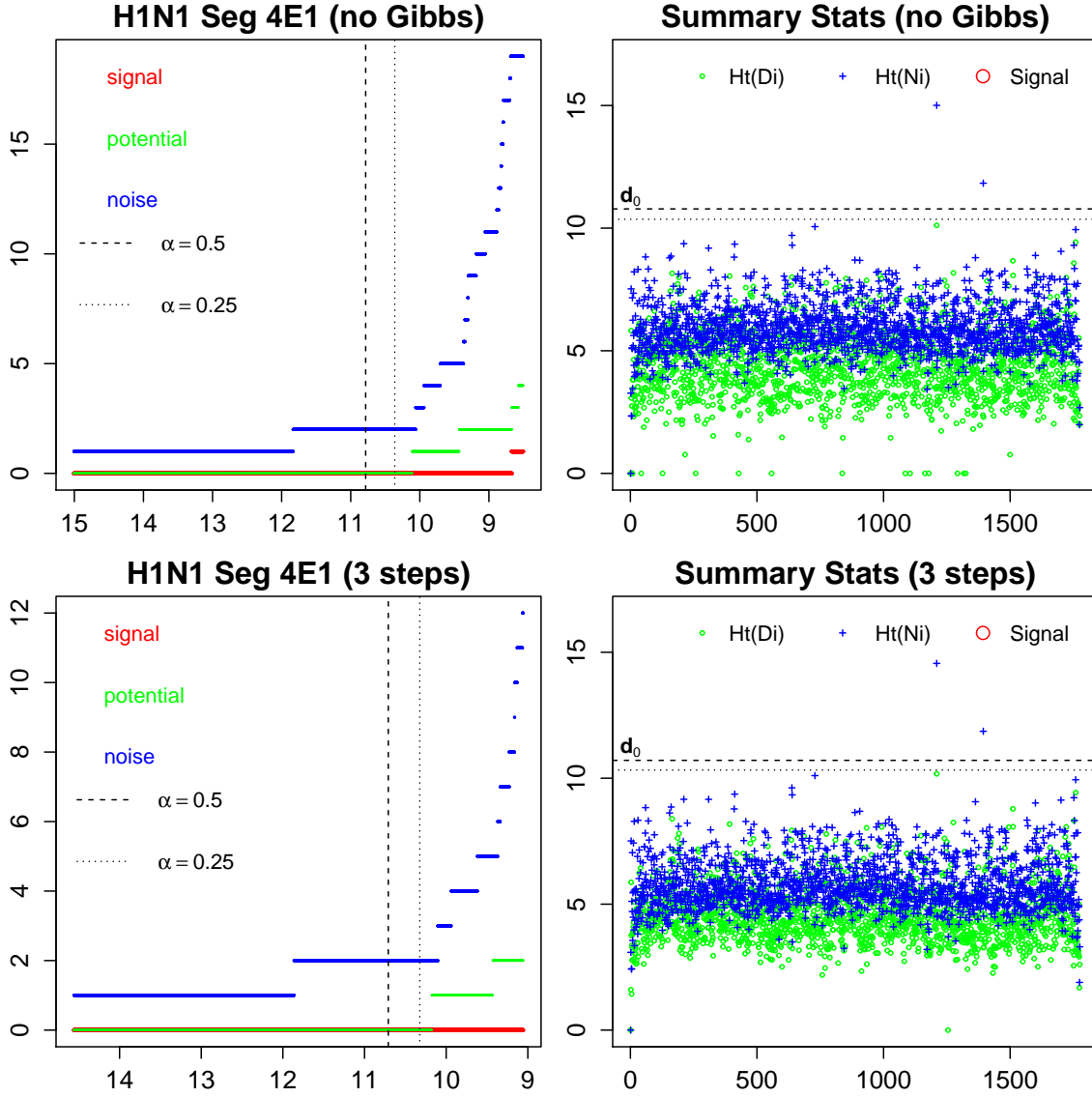


Figure 2.15: The two thresholds produce the same result. No drug resistant site is identified on this segment.

2.7.3 Raw data plot

Nucleotide read count proportion plot for identified IVA sites. The color tiles on the top of each panel indicates the total read count at each time point. The top and bottom rows are for each replicate; the left and right panels are for control and treatment groups, respectively.

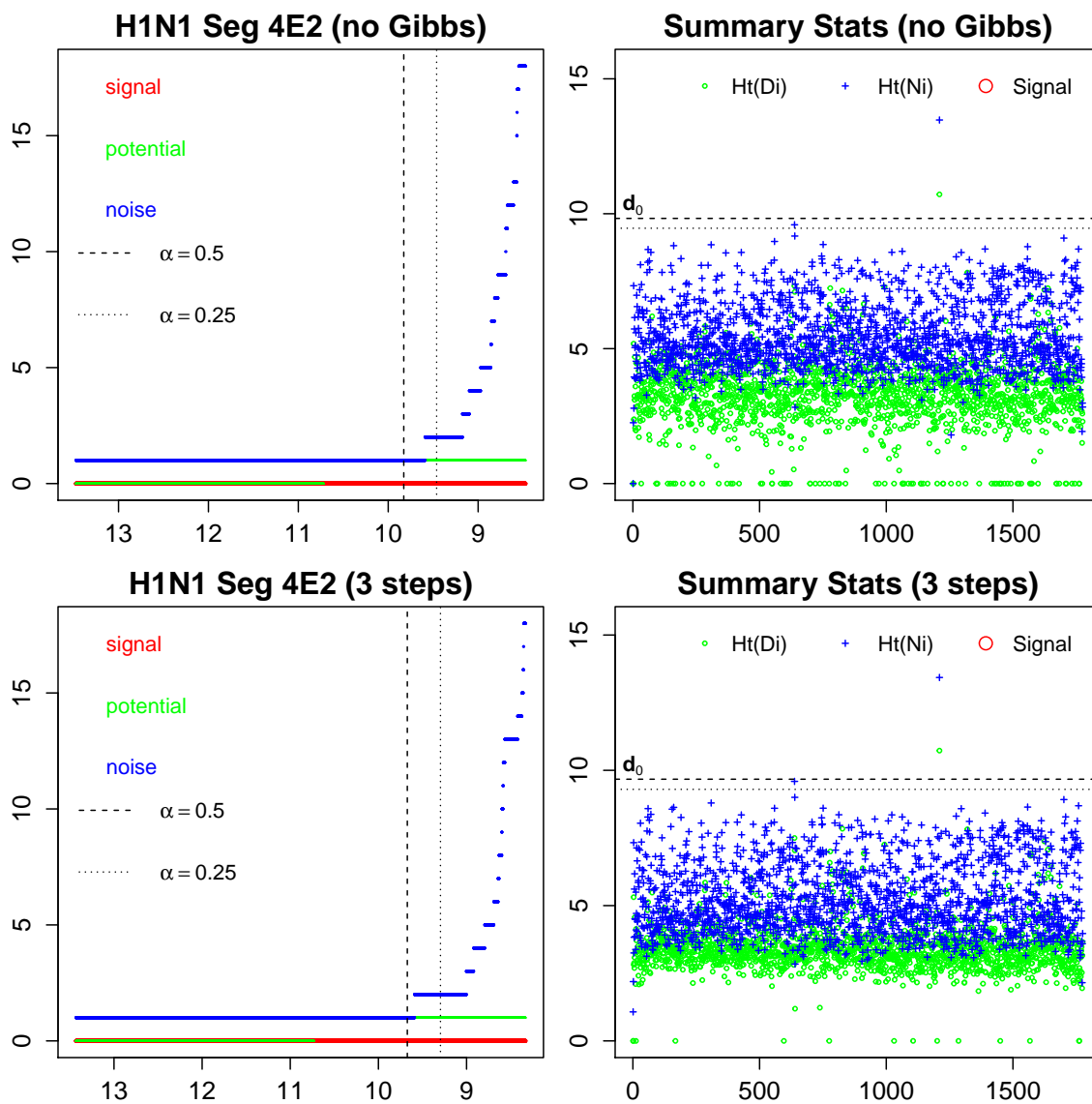


Figure 2.16: Similar to Seg4E1, no drug resistant site is identified.

The nucleotide read type proportions at each time point for the sites with high genetic variation that might be due to adaptation to the hosts: S1-2299, S3-2193, S4-1210, S5-24, S5-1103, S8-819.

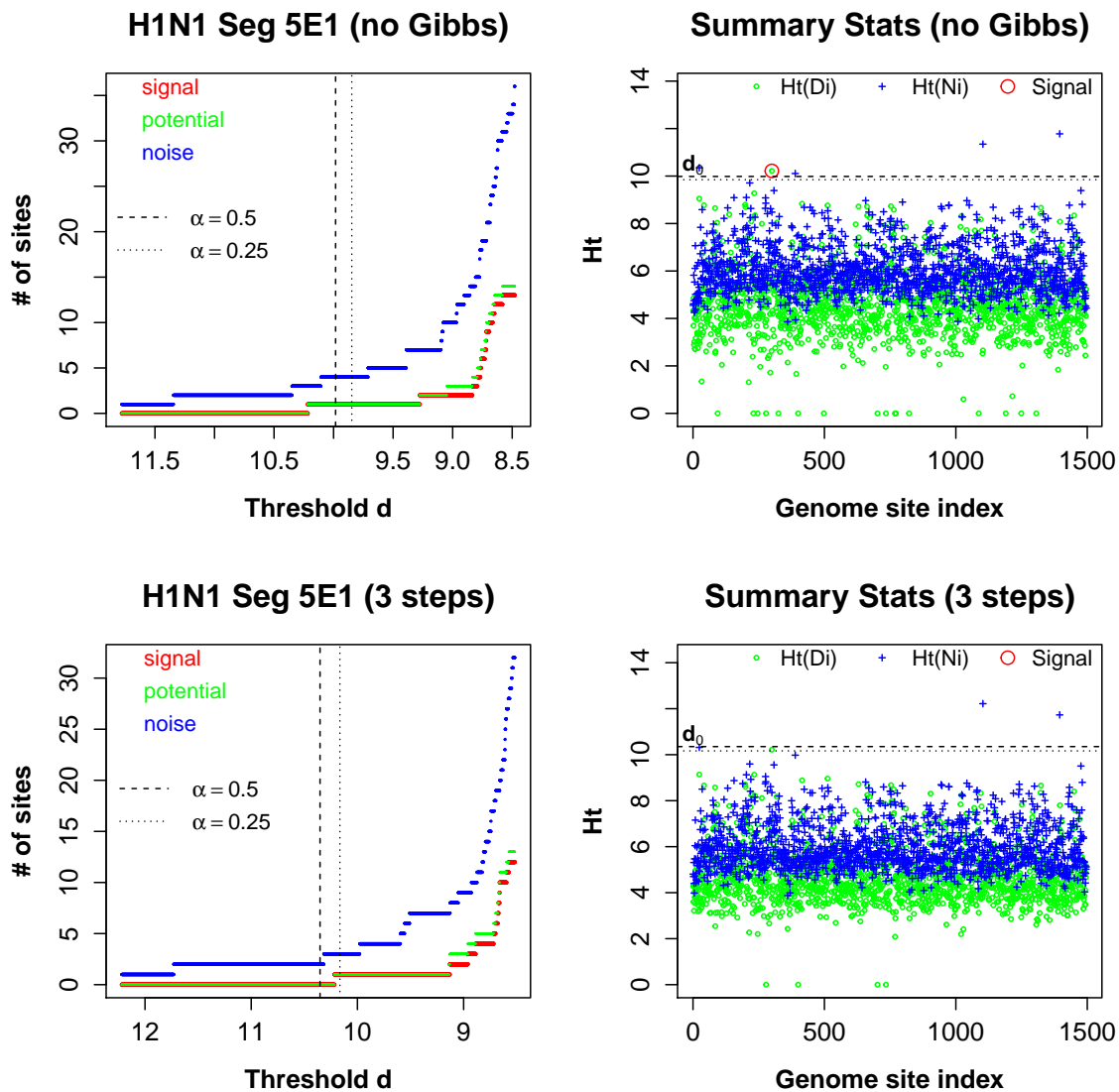


Figure 2.17: On this segment, without *Gibbs*, site S5-300 is highlighted, however, it is excluded from the signal set after the modification step.

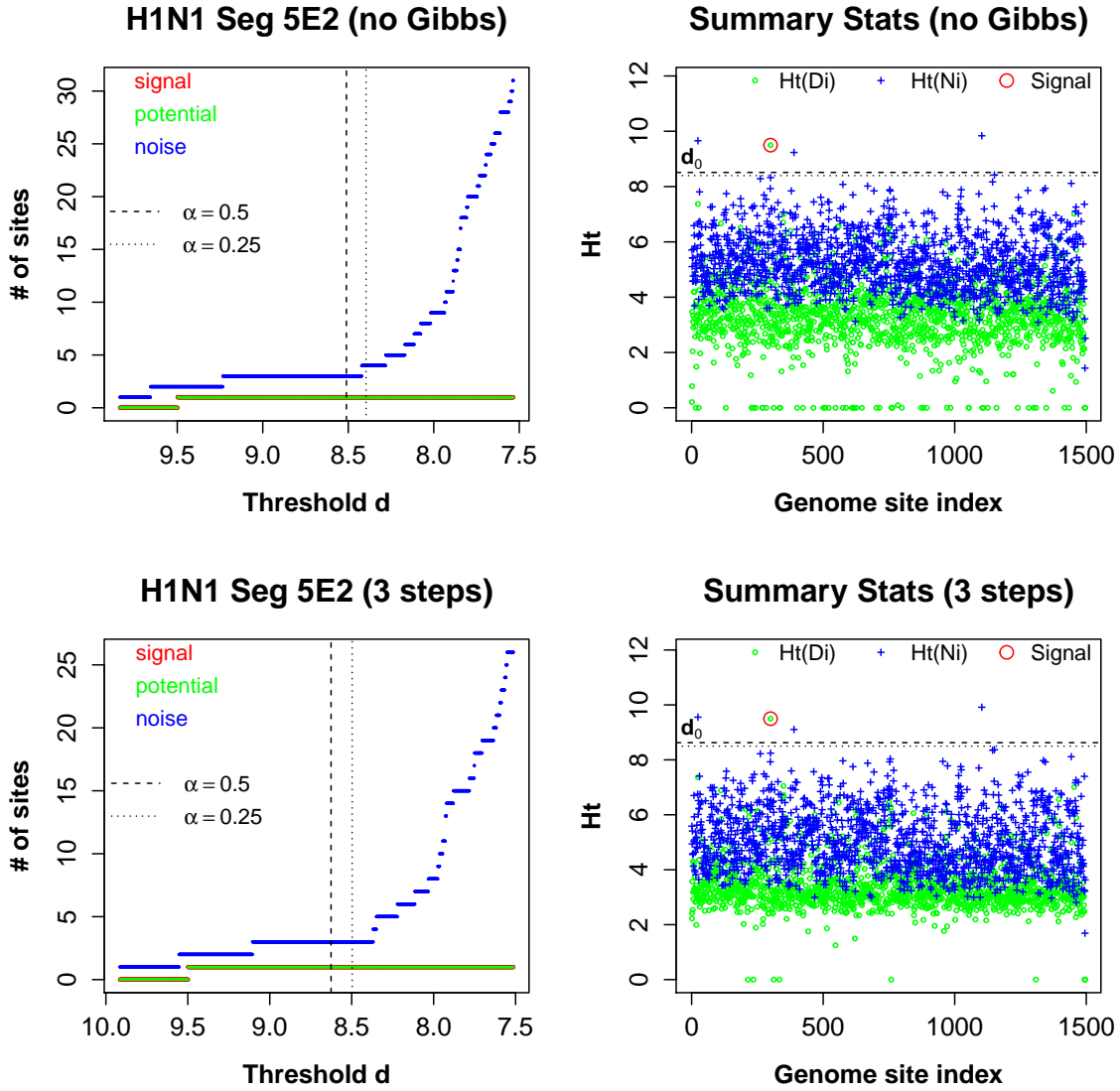


Figure 2.18: For Replicate II, both with and without the *Gibbs* reveals site S5-300. Since the site was not included in Replicate I, (with *Gibbs*), we exclude this location from the final result.

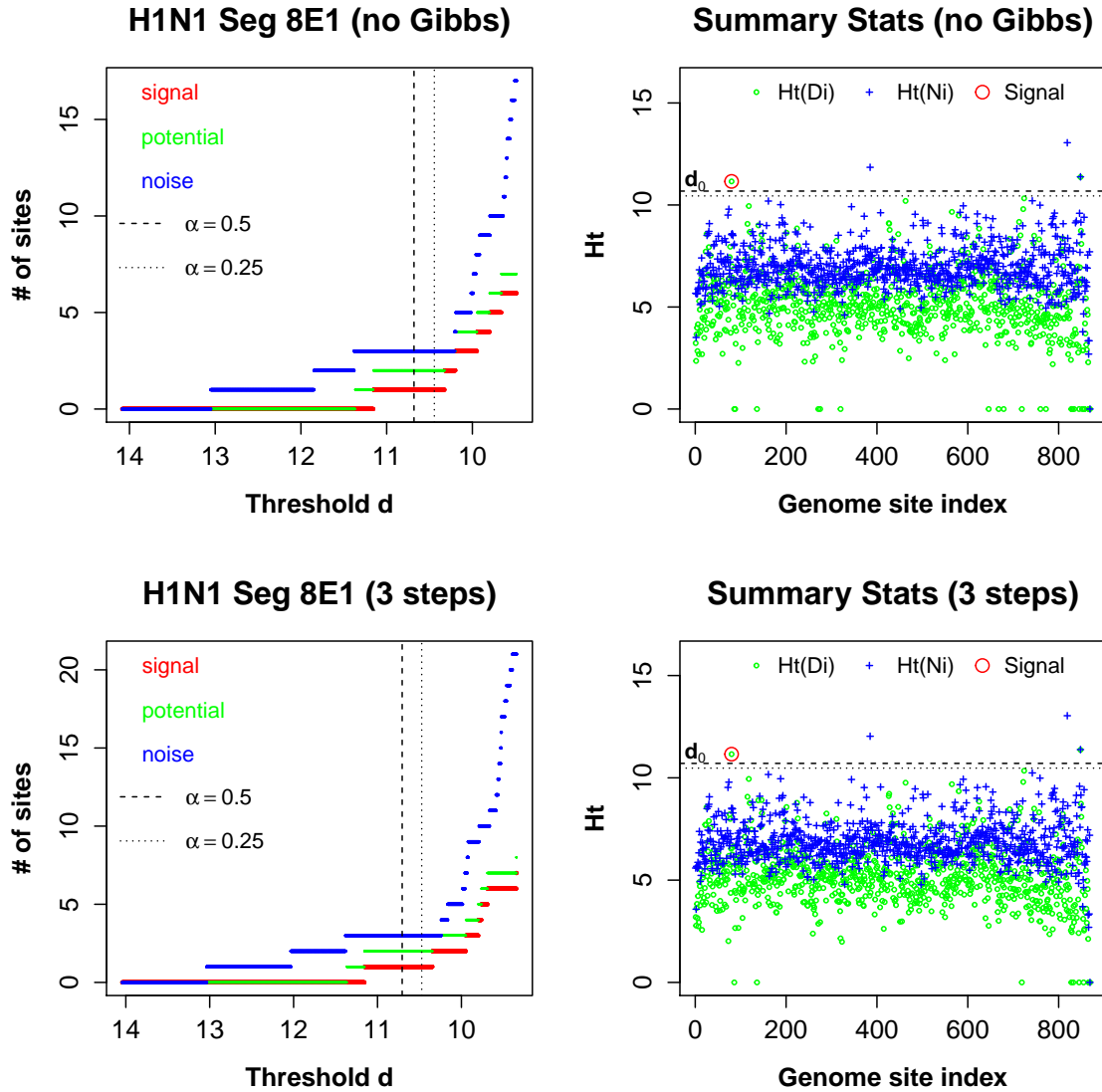


Figure 2.19: The result plots for H1N1 Seg8E1. The inference result for H1N1 Seg7E2 is consistent even without the *Gibbs* step, and is robust to the choices of α parameter. The highlighted site, S8-80, was identified as a drug resistant site with or without the *Gibbs* step. The noise set consists of S8-385, S8-819, S8-848.

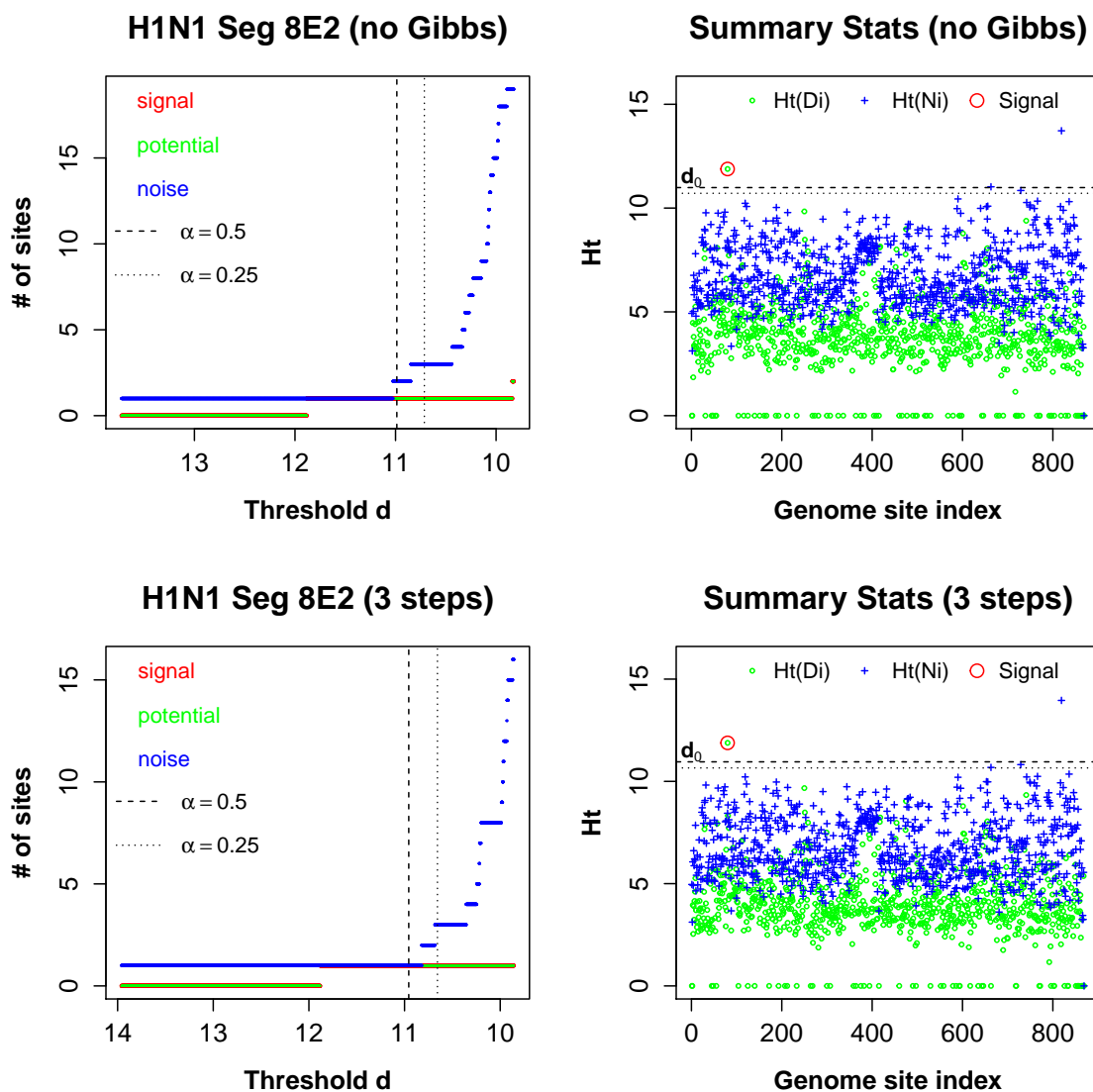


Figure 2.20: Similar to Seg8E1, the inference result is consistent with or without the *Gibbs* step. The highlighted S8-80 was identified as a signal site; S8-819 was a noise site that was also identified in Seg8E1.

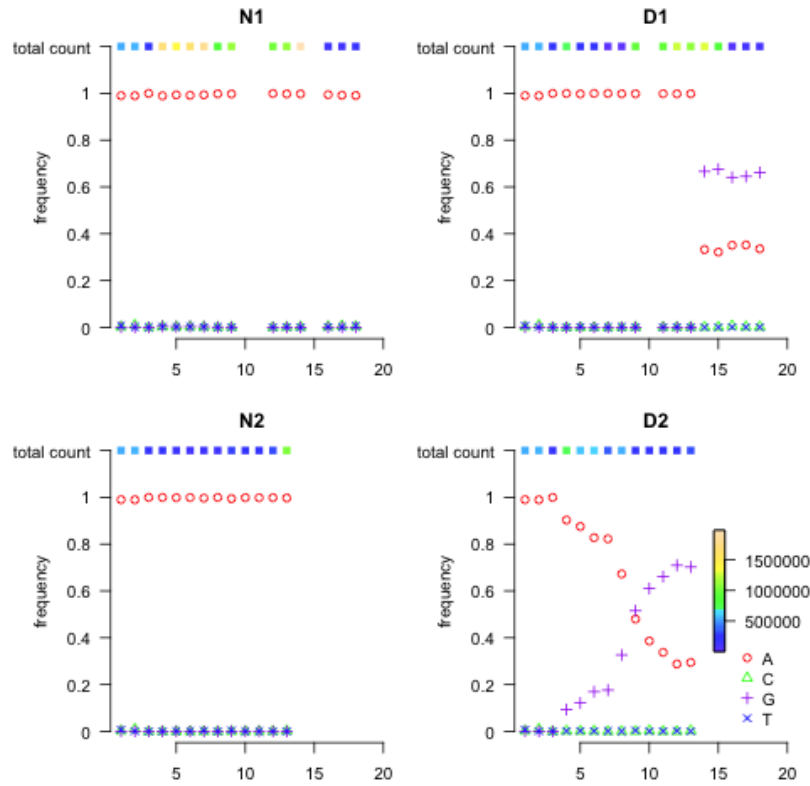


Figure 2.21: H1N1 nucleotide read count proportion and total count at position S8-80. Unlike the previous two figures, the read type does not switch completely. Instead, the starting read type A remains in more than 25% of the sample while the rest have read type G post treatment.

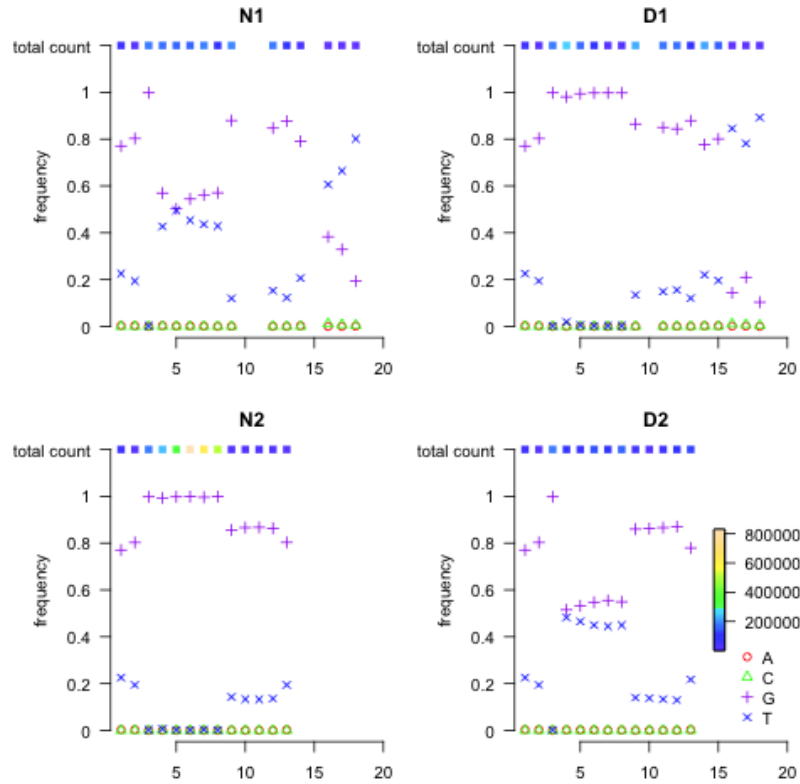


Figure 2.22: H1N1 nucleotide read count proportion and total count at position S1-2299. In all four panels, there is great variation between read type G and T. The alternating behavior happened in almost all panels in dictating that the genetic variation is not due to treatment and likely that the read types G and T together dominate the reads at equilibrium.

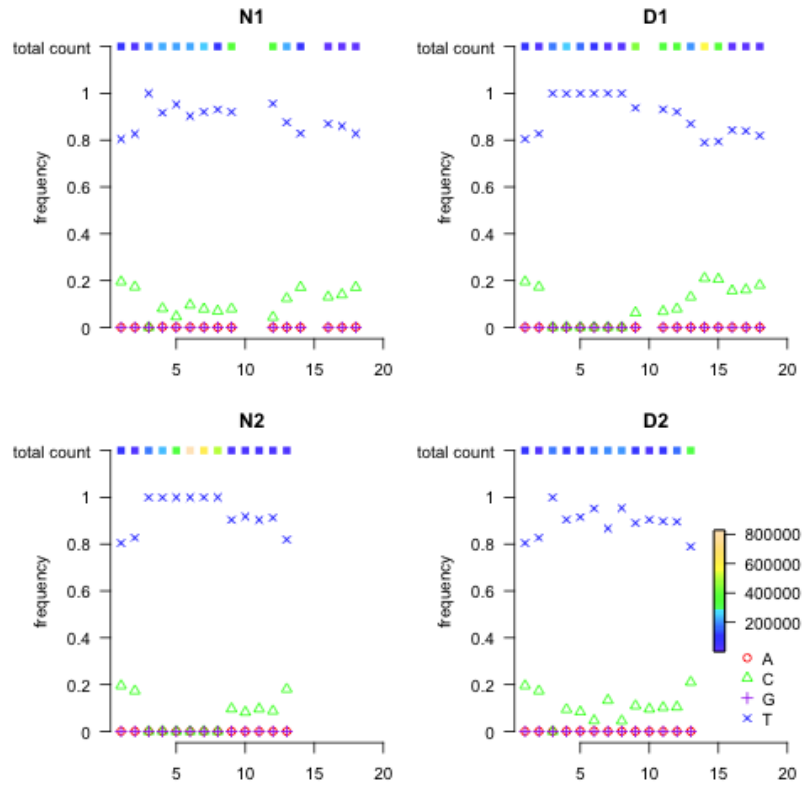


Figure 2.23: H1N1 nucleotide read count proportion and total count at position S1-2303. Regardless treatment or control, there was significant fluctuation in the mixture proportion of nucleotides C and T over time.

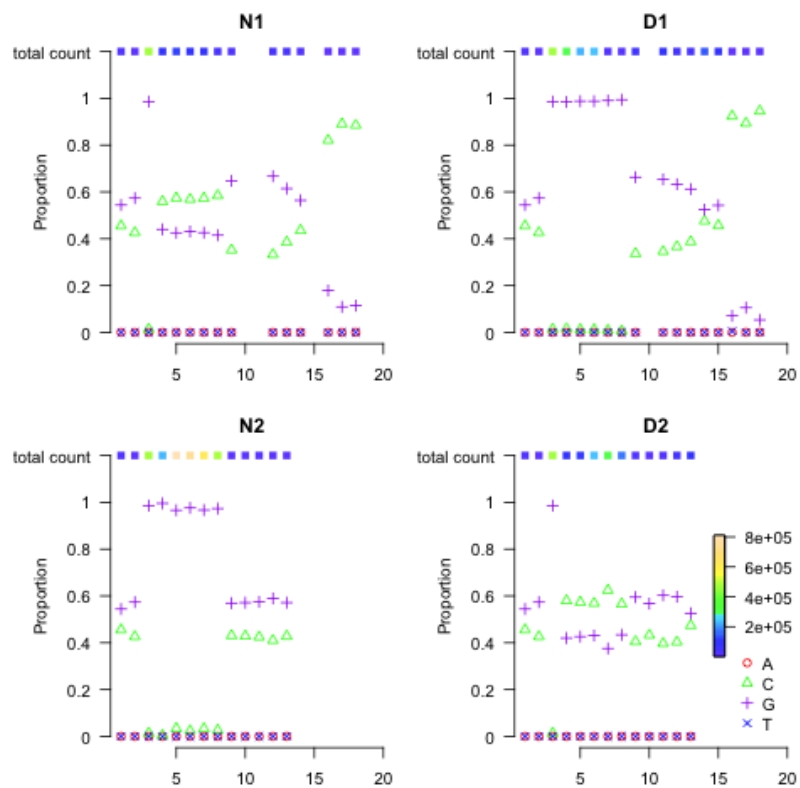


Figure 2.24: H1N1 nucleotide read count proportion and total count at position S3-2193. In all four panels, there is great variation between read type C and G with and without the treatment.

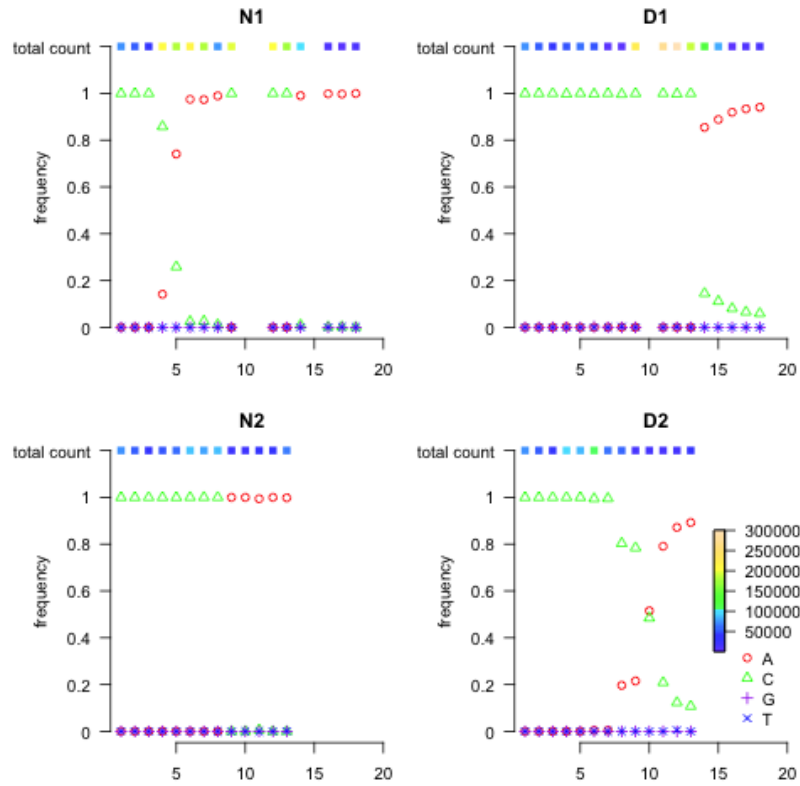


Figure 2.25: H1N1 nucleotide read count proportion and total count at position S4-1210. There is a complete transversion from C to A in all four panels, suggesting that this change might be due to adaptation to the hosts or genetic drifts.

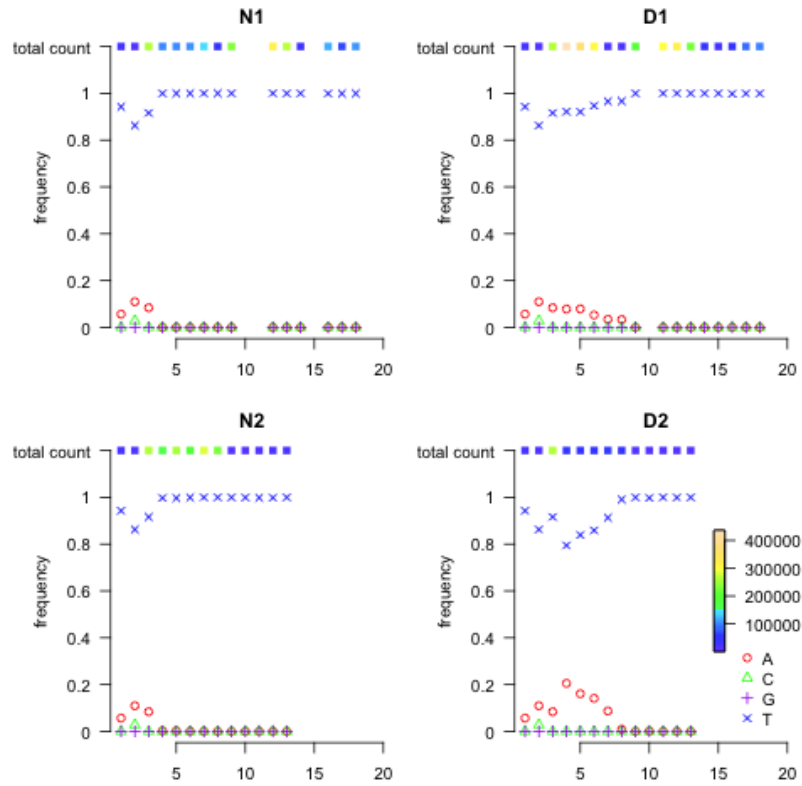


Figure 2.26: H1N1 nucleotide read count proportion and total count at position S5-24. The four panels show similar pattern. It appears that the treated groups (right two panels) took longer time to reach equilibrium.

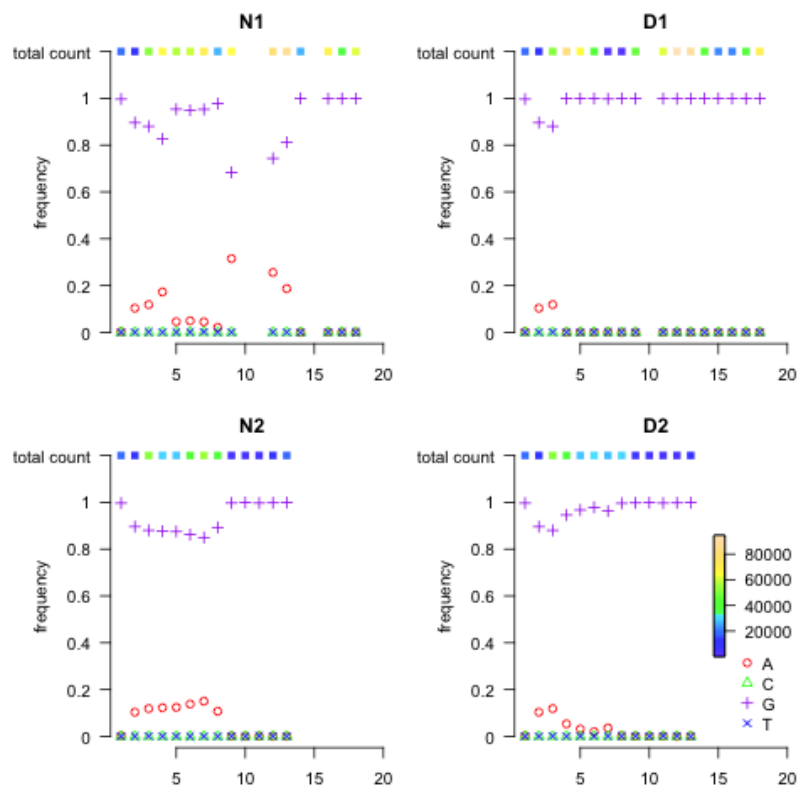


Figure 2.27: H1N1 nucleotide read count proportion and total count at position S5-389. The untreated groups appear to present greater variation over time.

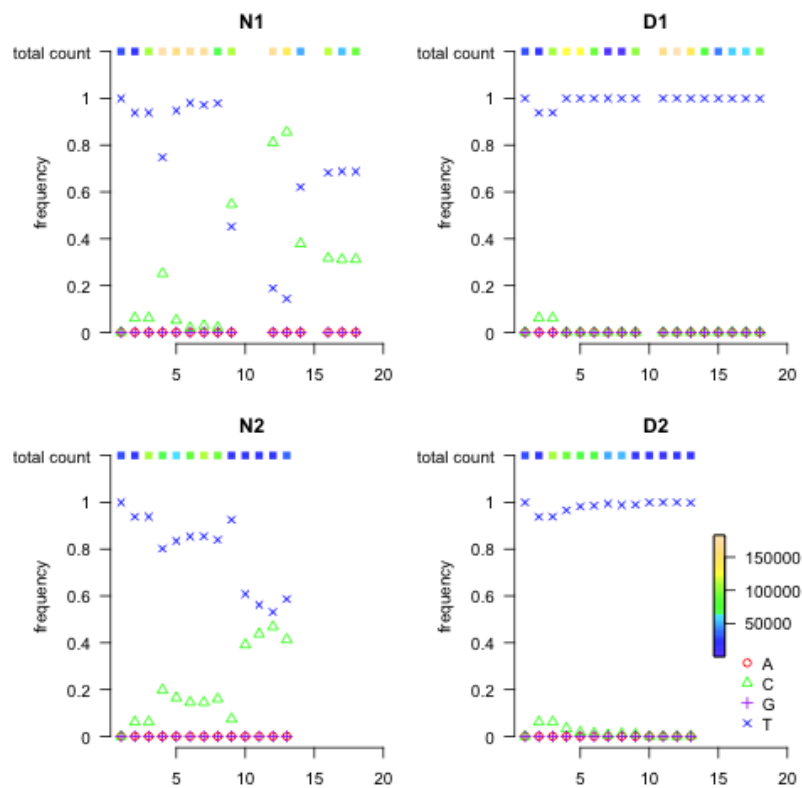


Figure 2.28: H1N1 nucleotide read count proportion and total count at position S5-1103. The untreated groups appear to present greater variation over time.

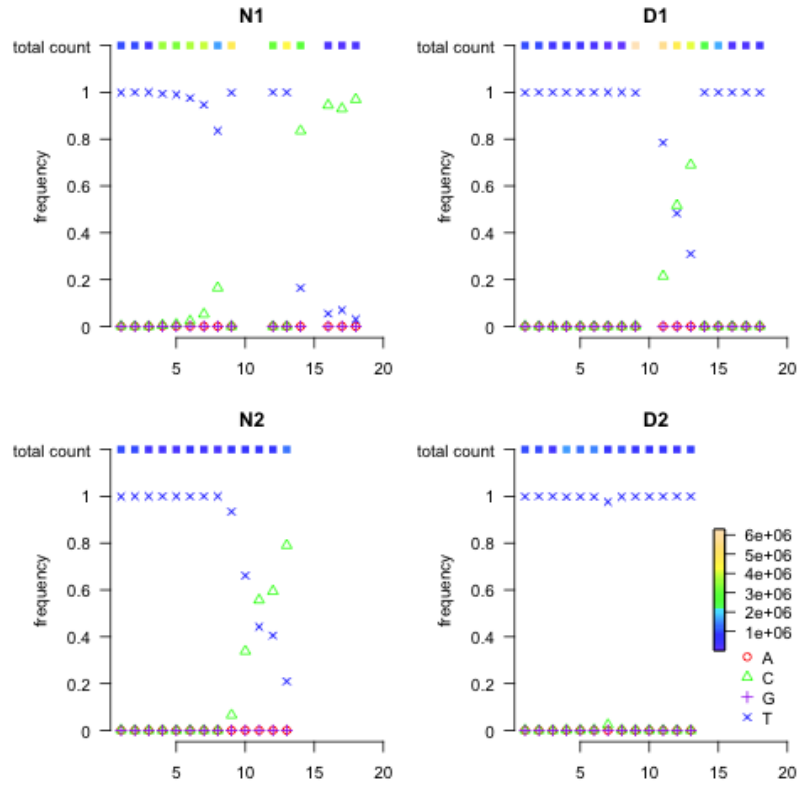


Figure 2.29: H1N1 nucleotide read count proportion and total count at position S8-819. The untreated groups present transition from G to C; one of the treated groups (top right panel) appears to have interchanged read type in the intermediate time points while the other treated group presents low variation.

CHAPTER 3

Covariance estimation via fiducial inference

3.1 Introduction

Estimating covariance matrices has always been an important task. We are particularly interested in a high-dimensional multivariate linear model setting with an atypical sparsity constraint. Instead of classic sparsity assumptions on the covariance matrix, we consider a type of experimental design that enforces sparsity on the covariate matrix. This phenomenon often arises in the studies of metabolomics and proteomics. One example of this setup is the modeling of the relationship between a set of gene expression levels and metabolomic data. The expression levels of the genes serve as the predictor variables while the response variables are a variety of metabolite levels, such as sugar and triglycerides. It is known that only a small subset of genes contribute to each metabolite level, and each gene can be responsible for just a few metabolite levels, hence the sparse structure in the covariate matrix. Another example is the analysis of protein-protein interaction networks. Detecting the cliques in the network can also be achieved through our fiducial covariance estimation framework.

In the world of covariance estimation, maximum likelihood based methods and Bayesian approaches are the most common tools. We took a different perspective by looking through the generalized fiducial inference glasses. The general approach of fiducial inference was first proposed by Ronald A. Fisher, whose intention was to overcome the need for priors and other problems with Bayesian methods at the time. The procedure of fiducial inference allows to obtain a measure on the parameter space without requiring priors and defines approximate pivots for parameters of interest. It is ideal when *a priori* information about the parameters is unavailable, which is often the case in biology. The key recipe of the fiducial argument is

the data generating equation. Roughly, the generalized fiducial likelihood is defined as the distribution of the functional inverse of the data generating mechanism.

In a high-dimensional and sparse covariate setting, we derived the generalized fiducial likelihood of the covariate matrix based on given observations and proved its asymptotic consistency as the sample size increases. Samples from the fiducial distribution of a covariate matrix can be generated using a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. Similar to the classic likelihood functions, fiducial distributions favor models with more parameters. Therefore, in the case where the exact sparsity structure of the covariate is unclear, a penalty term needs to be added. We chose a penalty function based on the minimum description length principle (MDL) (Rissanen, 1978). To obtain a family of covariance estimators, we adapted a zeroth-order method and develop an efficient RJMCMC algorithm that samples from the penalized fiducial distribution.

One great advantage of the fiducial approach to covariance matrix estimation is that, without specifying a prior, it produces a family of matrices that are close to the true covariance matrix with a probabilistic characterization using the fiducial likelihood function. This attractive property enables a meaningful definition for matrix confidence regions.

The rest of this chapter is organized as follows. In Section 3.2 we first provide a brief background and recent developments on fiducial inference. Then we introduce the fiducial model for covariance estimation and derive the Generalized Fiducial Distribution (GFD) for the covariate and covariance matrices in Section 3.3. We explore the asymptotic properties of the GFD of the covariance matrix under minor assumption, and show that it satisfies the Fiducial Bernstein von-Mises Theorem (Sonderegger and Hannig, 2012) in Section 3.4. Section 3.5 presents an adaptive Reversible Jump Markov Chain Monte Carlo method (RJMCMC) developed for sampling from the GFD, followed by simulation studies in Section 3.6. We then conclude the chapter with a discussion (Section 3.7).

3.2 Generalized fiducial inference

3.2.1 Brief background

The idea of fiducial inference was first proposed by Ronald Aylmer Fisher in 1930 when he first introduced the concept of a fiducial distribution of a parameter. In the case of a single parameter family of distributions, Fisher gave the following definition for a *fiducial density* $f(\theta|x)$ of the parameter based on a single observation x for the case where the cumulative distribution function $F(x|\theta)$ is a monotonic decreasing function of θ :

$$f(\theta|x) \propto -\frac{\partial F(x|\theta)}{\partial \theta} \quad (3.1)$$

A fiducial distribution can be viewed as a Bayesian posterior distribution without hand picking priors. In many single parameter distribution families, Fisher's fiducial intervals coincide with classical confidence interval. For families of distributions with multiple parameters, the fiducial approach leads to confidence set. Seeing the advantage of avoiding choosing prior ad hoc, Fisher proposed the use of fiducial inference in replace of Bayesian framework, which led to major discussions among prominent statisticians in 1930's, 40's, and 50's (Dempster, 1966, 1968; Fraser, 1961b,a, 1966, 1968; Jeffereys, 1940; Lindley, 1958; Stevens, 1950). Many focused on non-exactness of the confidence sets and non-uniqueness of fiducial distributions. In the latter part of 20th century, only a handful of publications (Barnard, 1995; Dawid and Stone, 1982; Salome, 1998; Wilkinson, 1977) regarding Fisher's idea has been seen as the fiducial approach fell into disfavor. In recent years, the work of Tsui and Weerahndi (Tsui and Weerahandi, 1989, 1991) and Weerahndi (Weerahandi, 1993, 1994, 1995) on *generalized confidence intervals* and the work of Chiang (Chiang, 2001) on the *surrogate variable method* for obtaining confidence intervals for variance components led to the realization that there was a connection between these new procedures and fiducial inference. This realization evolved through a series of works by Hannig, Iyer and their collaborators (Hannig and Lee, 2009a; Hannig et al., 2006b; Iyer et al., 2004; Patterson et al., 2004) where the definition of fiducial inference has been generalized. The strengths and limitations of fiducial approach has been better understood (Hannig and Lee, 2009a; Hannig, 2012). In particular, the asymptotic exactness of fiducial confidence seats, under fairly

general conditions, was established in (Hannig, 2012; Hannig et al., 2006b; Sonderegger and Hannig, 2012). The generalized fiducial approach has been applied to a variety of models, both parametric and nonparametric, both continuous and discrete. These applications include bioequivalence (Hannig et al., 2006a), variance components (Cisewski and Hannig, 2012; E et al., 2008), problems of metrology (Hannig et al., 2007, 2003; Wang et al., 2012; Wang and Iyer, 2005, 2006a,b), inter laboratory experiments and international key comparison experiments (Iyer et al., 2004), maximum mean of a multivariate normal distribution (Wandler and Hannig, 2011), multiple comparisons (Wandler and Hannig, 2012a), extreme value estimation (Wandler and Hannig, 2012b), mixture of normal and Cauchy distributions (Glagovskiy, 2006), wavelet regression (Hannig and Lee, 2009b), logistic regression and LD₅₀ (E et al., 2009).

3.2.2 Generalized fiducial distribution

The idea underlying GFI is built upon a *data generating equation* $G(\cdot, \cdot)$ expressing the relationship between the data X and the parameters θ :

$$X = G(U, \theta), \quad (3.2)$$

where U is the random component of this data generating equation whose distribution is known. The data X are assumed to be created by generating a random variable U and plugging it into the data generating equation (3.2). A set-value function of inverse notion of G is defined as

$$Q(x, u) = \{\theta : x = G(u, \theta)\}, \quad (3.3)$$

where u is an implicit function of θ , x is a fixed realization of X .

A *weak fiducial distribution* of θ is defined as a conditional distribution of

$$V(Q(x, U^*)) | \{Q(x, U^*) \neq \emptyset\}, \quad (3.4)$$

where U^* is an independent copy of U , and V is a function often taken to be identity. A weak fiducial distribution is well-defined provided that $\mathbb{P}(Q(x, U^*) \neq \emptyset) > 0$. However, if $\mathbb{P}(Q(x, U^*) \neq \emptyset) = 0$, definition (3.4) leads to a potential source of non-uniqueness due to

conditioning on events with zero probability, known as Borel Paradox (Casella and Berger, 2002). Since it is reasonable to assume the data have been discretized by a measuring device and storage on a computer, this potential hazard can be resolved by simply perturbing the observation and looking at a small neighborhood of the observation (Hannig, 2009; Hannig et al., 2007). We then arrive at a formal definition of *Generalized Fiducial Distribution* (GFD):

$$\lim_{\epsilon \rightarrow 0} [Q_{x,\epsilon}(U^*) | \{Q_{x,\epsilon}(U^*) \neq \emptyset\}] \quad (3.5)$$

where $Q_{x,\epsilon}(U^*) = \{\theta : \|x - G(U^*, \theta)\| < \epsilon\}$ and $\|v\|$ is a norm of the vector v . Equation (3.5) can be rewritten as

$$\lim_{\epsilon \rightarrow 0} [\operatorname{argmin}_{\theta} \|x - G(U^*, \theta)\| | \min_{\theta} \|x - G(U^*, \theta)\| < \epsilon]. \quad (3.6)$$

While definition (3.6) is conceptually appealing and very general, it not immediately clear how to compute the limit in many practical situations. In a less general setup using l^∞ norm, Hannig derives a closed form of the limit in (3.6) applicable to many practical situations (Hannig, 2012). In particular, assume that the parameter $\theta \in \Theta \subset \mathbb{R}^p$ is p -dimensional and that the inverse to (3.2) $G^{-1}(x, \theta) = u$ exists. Then under some differentiability assumptions, Hannig (Hannig, 2012) has shown that the GFD is absolutely continuous with density

$$r(\theta|x) = \frac{f(x, \theta)J(x, \theta)}{\int_{\Theta} f(x, \theta')J(x, \theta')d\theta'}, \quad (3.7)$$

where

$$J(x, \theta) = \binom{n}{p}^{-1} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p) \\ 1 \leq i_1 < \dots < i_p \leq n}} \left| \det \left(\frac{d}{d\theta} G(u, \theta) \Big|_{u=G^{-1}(x, \theta)} \right)_{\mathbf{i}} \right| \quad (3.8)$$

In the above $f(x, \theta)$ is the likelihood the the sum goes over all p -tuples of indices $\mathbf{i} = (1 \leq i_1 < \dots < i_p \leq n) \subset \{1, \dots, n\}$. For any matrix M , $(M)_{\mathbf{i}}$ stands for a $p \times p$ matrix consisting of rows $\mathbf{i} = (i_1, \dots, i_p)$ of M .

3.3 A fiducial approach to covariance estimation

In this section, we introduce the data generating equation under our framework. For two special cases and the general case, we derive the GFD for the covariance matrix of a multivariate normal random variable.

3.3.1 Data generating equation

Let Q^T denote the transpose of a matrix Q . For a collection of n observed p dimensional objects $\mathbf{Y} = \{Y_i, i = 1, \dots, n\}$. Consider the following data generating equation:

$$Y_i = AZ_i, i = 1, \dots, n; \quad (3.9)$$

where A is a $p \times p$ matrix of full rank, $\mathbf{Z} = \{Z_i = (z_{i1}, \dots, z_{ip})^T, i = 1, \dots, n\}$ are independent and identically distributed (*i.i.d*) $p \times 1$ random vectors following multivariate normal distribution $N(0, I)$. Hence, Y_i 's are *i.i.d* random vectors centered at 0 with variance AA^T ,

$$\text{i.e. } Y_i \stackrel{\text{iid}}{\sim} N(0, \Sigma), \text{ where } \Sigma = AA^T. \quad (3.10)$$

Consequently, we have

$$f(\mathbf{Y}, A) = (2\pi)^{-\frac{np}{2}} |\det(A)|^{-n} \exp \left[-\frac{1}{2} \text{tr}\{nS_n(AA^T)^{-1}\} \right], \quad (3.11)$$

where $S_n = \frac{1}{n} \sum_{i=1}^n Y_i' Y_i$ is the corresponding sample covariance matrix and $\text{tr}\{\cdot\}$ is the trace operator.

We propose to estimate the covariance matrix Σ through the GFD of covariate matrix A :

$$r(A|\mathbf{Y}) \propto J(\mathbf{Y}, A) f(\mathbf{Y}, A) \quad (3.12)$$

Since A is assumed to be invertible, we have $Z_i = A^{-1}Y_i$ for all i . For any $\mathbf{i} = (i_1, \dots, i_p)$, $1 \leq i_1 < \dots < i_p \leq n$, denote the stacked observation vector $W_{\mathbf{i}} = (Y_{i_1}^T, \dots, Y_{i_p}^T)^T = (w_1^{\mathbf{i}}, \dots, w_p^{\mathbf{i}})^T$, and let a_{ij} be the ij th entry of matrix A , i.e. $A =$

$[a_{ij}]_{1 \leq i, j \leq p}$. The corresponding Jacobian $J(\mathbf{Y}, A)$ derived from (3.8) is then

$$J(\mathbf{Y}, A) = \binom{n}{p}^{-1} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p) \\ 1 \leq i_1 < \dots < i_p \leq n}} \left| \det \left(\frac{\partial W_{\mathbf{i}}}{\partial A} \right) \right|, \quad (3.13)$$

and $\frac{\partial W_{\mathbf{i}}}{\partial A}$ is defined to be

$$\frac{\partial W_{\mathbf{i}}}{\partial A} = \begin{pmatrix} \frac{\partial w_1^{\mathbf{i}}}{\partial A} \\ \frac{\partial w_2^{\mathbf{i}}}{\partial A} \\ \vdots \\ \frac{\partial w_{p^2}^{\mathbf{i}}}{\partial A} \end{pmatrix};$$

for all $q \in 1, \dots, p^2$,

$$\frac{\partial w_q^{\mathbf{i}}}{\partial A} = \begin{pmatrix} \frac{\partial w_q^{\mathbf{i}}}{\partial a_{11}} & \frac{\partial w_q^{\mathbf{i}}}{\partial a_{12}} & \dots & \frac{\partial w_q^{\mathbf{i}}}{\partial a_{ij}} & \dots & \frac{\partial w_q^{\mathbf{i}}}{\partial a_{pp}} \end{pmatrix}.$$

Note that if a_{kl} is fixed at zero, then the term $\frac{\partial w_q^{\mathbf{i}}}{\partial a_{kl}}$ will be excluded in $\frac{\partial w_q^{\mathbf{i}}}{\partial A}$.

3.3.2 Jacobian

With a complex form of Jacobian (3.13), finding the GFD of A is not trivial. We start with two simple cases and close this subsection with discussions on how to work with the general case.

- Special Case I: No element of A is fixed at zero, i.e. the parameter space is $\mathbb{R}^{p \times p}$.

$$\frac{\partial W_{\mathbf{i}}}{\partial A} = \begin{matrix} & B_1 & B_2 & \cdots & B_p \\ \begin{matrix} R_1 \\ R_2 \\ \vdots \\ R_p \\ \vdots \\ R_{(p-1)p+1} \\ R_{(p-1)p+2} \\ \vdots \\ R_{p^2} \end{matrix} & \begin{pmatrix} (A^{-1}Y_{i_1})^T & & & \\ & (A^{-1}Y_{i_1})^T & & \\ & & \ddots & \\ & & & (A^{-1}Y_{i_1})^T \\ & \vdots & \vdots & \vdots & \vdots \\ (A^{-1}Y_{i_p})^T & & & & \\ & (A^{-1}Y_{i_p})^T & & & \\ & & \ddots & & \\ & & & (A^{-1}Y_{i_p})^T \end{pmatrix} \end{matrix} = Q_{\mathbf{i}}. \quad (3.14)$$

Here the matrix $\frac{\partial W_{\mathbf{i}}}{\partial A} = Q_{\mathbf{i}}$ consists of p blocks, B_1, \dots, B_p , each of dimension $p^2 \times p$. Every row of $Q_{\mathbf{i}}$, R_1, \dots, R_{p^2} , has non-zero entries in only one block. By swapping rows in the matrix $Q_{\mathbf{i}}$, we can achieve matrix $P_{\mathbf{i}}$:

$$P_{\mathbf{i}} = \begin{matrix} & B'_1 & \cdots & B'_p \\ \begin{pmatrix} (A^{-1}Y_{i_1})^T \\ \vdots \\ (A^{-1}Y_{i_p})^T \\ & \ddots & \\ & & (A^{-1}Y_{i_1})^T \\ & & \vdots \\ & & (A^{-1}Y_{i_p})^T \end{pmatrix} \end{matrix} = \begin{pmatrix} B'_1 & \cdots & B'_p \\ U_{\mathbf{i}} & & \\ & \ddots & \\ & & U_{\mathbf{i}} \end{pmatrix}, \quad (3.15)$$

where $U_{\mathbf{i}} = (A^{-1}Y_{i_1}; \dots; A^{-1}Y_{i_p})^T = V_{\mathbf{i}}(A^{-1})^T$, $V_{\mathbf{i}} = (Y_{i_1}; \dots; Y_{i_p})^T$.

Since swapping rows does not change the absolute value of the determinant of a matrix, the Jacobian (3.13) can be expressed with matrix $P_{\mathbf{i}}$:

$$J(\mathbf{Y}, A) = \binom{n}{p}^{-1} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p) \\ 1 \leq i_1 < \dots < i_p \leq n}} |\det(P_{\mathbf{i}})|. \quad (3.16)$$

Therefore, in the case when all entries of A are not identically zero, we have

$$J(\mathbf{Y}, A) = \binom{n}{p}^{-1} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p) \\ 1 \leq i_1 < \dots < i_p \leq n}} |\det(U_{\mathbf{i}})|^p = C(\mathbf{Y}) |\det(A)|^{-p}, \quad (3.17)$$

where

$$C(\mathbf{Y}) = \binom{n}{p}^{-1} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p) \\ 1 \leq i_1 < \dots < i_p \leq n}} |\det(V_{\mathbf{i}})|^p, \quad (3.18)$$

By (3.12), the GFD is proportional to

$$r(A|\mathbf{Y}) \propto C(\mathbf{Y}) (2\pi)^{-\frac{np}{2}} |\det(A)|^{-(n+p)} \exp \left[-\frac{1}{2} \text{tr} \{ n S_n (A A^T)^{-1} \} \right]. \quad (3.19)$$

Transforming GFD of A , we conclude that the GFD of $\Sigma = A A^T$ has the inverse Wishart distribution with n degrees of freedom and parameter $n S_n$.

- Special Case II: Clique model.

Now suppose that A is a block diagonal matrix. Assume that the coordinates of \mathbf{Y} are broken into cliques; i.e. coordinates i and j are correlated if i, j belong to the same clique and independent otherwise. Equivalently, $a_{i,j} = 0$ if i and j are not in the same clique. The Minimum Description Length (MDL) (Rissanen, 1978) for model A with k cliques with sizes g_1, \dots, g_k respectively is

$$q(A) = \frac{1}{2} \left(\sum_{i=1}^k g_i^2 \right) \log n + (p+1) \log k \quad (3.20)$$

Denote S_n^i the $g_i \times g_i$ sample covariance matrix of the i th clique and $C_i(\mathbf{Y})$ the constant in the Jacobian function computed only using the coordinates in clique i . Applying MDL penalty and using the facts from the Wishart distribution to integrate out (3.19), we have the GFD of the model is proportional to

$$\frac{\prod_{i=1}^k \left[\Gamma_{g_i} \left(\frac{n}{2} \right) C_i(\mathbf{Y}) (2\pi)^{\frac{g_i(g_i-1)}{2}} |\det S_n^i|^{-\frac{n}{2}} \right]}{\exp \left\{ \frac{1}{2} \left(\sum_{i=1}^k g_i^2 \right) \log n + (p+1) \log k \right\}} \quad (3.21)$$

where $\Gamma_{g_i}(\frac{n}{2})$ is the multivariate gamma function.

- General Case: Possible fixed zero elements in A .

The assumption of some fixed zero elements in A enables covariance estimation in large dimension low sample size scenarios. There are many possible biological applications for the set-up. One application is to model the relationship between metabolism and gene expression levels.

Denote the ij th entry of A as A_{ij} . Let B'_{ij} be the j th column of block B'_i . Define a set of paired indices S_i by

$$S_i = \{(i, j) : A_{ij} \equiv 0, j = 1, \dots, p\}, i = 1, \dots, p. \quad (3.22)$$

The set S_i indicates which entries of A in the i th row are identically zero. The union of all S_i 's, $S_0 = \cup_{i=1}^p S_i$, consists of the indices of all fixed zeros in the covariate matrix A . Denote that total number of non-fixed-zeros in A as $p_A = p^2 - |S_0|$. Let R be a set of column vectors such that

$$R = \{B'_{ij} : (i, j) \in S_0\} = \{B'_{ij} : A_{ij} \equiv 0\}. \quad (3.23)$$

Then equation (3.13) becomes

$$J(\mathbf{Y}, A) = \binom{n}{p}^{-1} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p), 1 \leq i_1 < \dots < i_p \leq n, \\ \mathbf{r}=(r_1, \dots, r_{p_A}), 1 \leq r_1 < \dots < r_{p_A} \leq p^2.}} \left| \det (P_{\mathbf{i}, [-R]})_{\mathbf{r}} \right|, \quad (3.24)$$

where $P_{\mathbf{i}, [-R]}$ is the largest submatrix of $P_{\mathbf{i}}$ without the columns included in R .

In order to have nonzero $\det (P_{\mathbf{i}, [-R]})_{\mathbf{r}}$, the index vector \mathbf{r} has to include p_i number of nonzero rows of block B'_i . The integer p_i is the number of nonzero entries in the i th row of A . Hence, the Jacobian (3.24) can be reduced to

$$J(\mathbf{Y}, A) = \binom{n}{p}^{-1} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p), 1 \leq i_1 < \dots < i_p \leq n, \\ \mathbf{r}_i=(r_{i,1}, \dots, r_{i,p_i}), 1 \leq r_{i,1} < \dots < r_{i,p_i} \leq p.}} \prod_{i=1}^p \left| \det (U_{\mathbf{i}, i})_{\mathbf{r}_i} \right|. \quad (3.25)$$

Here each $U_{\mathbf{i},i}$ is the largest submatrix of $U_{\mathbf{i}}$ without the columns in S_i , i.e. $U_{\mathbf{i},i} = U_{\mathbf{i},[:, -S_i]}$.

When the dimension of A and sample size become large, direct calculation of the Jacobian can be infeasible with the formula (3.25). Notice that for each \mathbf{i} , we sum over total $\prod_{i=1}^p \binom{p}{p_i}$ determinant products. It can be shown by induction that above equation is equivalent to the following:

$$J(\mathbf{Y}, A) = \binom{n}{p}^{-1} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p) \\ 1 \leq i_1 < \dots < i_p \leq n}} \prod_{i=1}^p \binom{p}{p_i} \overline{|\det(U_{\mathbf{i},i})_{\mathbf{r}_i}|}, \quad (3.26)$$

where $\overline{|\det(U_{\mathbf{i},i})_{\mathbf{r}_i}|}$ denote the average absolute determinant of all possible expressions of $(U_{\mathbf{i},i})_{\mathbf{r}_i}$ for a fixed index vector \mathbf{i} and a row number i .

Therefore, in the general case, the GFD is proportional to

$$\frac{\exp[-\frac{1}{2}\text{tr}\{nS_n(AA^T)^{-1}\}]}{|\det(A)|^n \binom{n}{p}} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p) \\ 1 \leq i_1 < \dots < i_p \leq n}} \prod_{i=1}^p \binom{p}{p_i} \overline{|\det(U_{\mathbf{i},i})_{\mathbf{r}_i}|} \quad (3.27)$$

3.4 Theoretic results

In general, there is no one-to-one correspondence between the covariance matrix Σ and the covariate matrix A . This leads to the identifiability issue from Σ to A . However, if A is assumed to be sparse with the sparse locations known, then the identifiability problem often vanishes. In this section we show that, if there is one-to-one correspondence between Σ and A , then the GFD defined by (3.27) achieves the Fiducial Bernstein-von Mises Theorem, which provides theoretical guarantees of asymptotic normality and asymptotic efficiency for GFD (Sonderegger and Hannig, 2012).

The results here are derived based on FM-distance, a metric for comparing covariance matrices suggested by Förstner and Moonen (Förstner and Moonen, 1999). For two symmetric positive definite matrices M and N , with the eigenvalues $\lambda_i(M, N)$ from $\det(\lambda H - C) = 0$,

the FM-distance between the two matrices M and N is

$$\mathbf{d} = \sqrt{\sum_{i=1}^n \ln^2 \lambda_i(M, N)}. \quad (3.28)$$

This distance measure is a metric and invariant with respect to both affine transformations of the coordinate system and an inversion of the matrices (Förstner and Moonen, 1999). More details on FM-distance can be found in the appendix.

The particular choice of the distance measure for covariance matrices is not crucial in proving the consistency of GFD.

Definition 3.1. *For a fixed covariate matrix A_0 and $\delta \geq 0$, define the δ -neighborhood of A_0 as the set $B(A_0, \delta) = \{A : \mathbf{d}(AA^T, A_0A_0^T) \leq \delta\}$.*

Before presenting the theorem on consistency of the GFD, we need to establish some regularity condition on the likelihood function and Jacobian formula (Propositions 3.1, 3.2, 3.3). The proofs can be found in the appendix.

Proposition 3.1. *For any $\delta > 0$ there exists $\epsilon > 0$ such that*

$$P_{A_0} \left\{ \sup_{A \notin B(A_0, \delta)} \frac{1}{n} (L_n(A) - L_n(A_0)) \leq -\epsilon \right\} \rightarrow 1,$$

where $L_n(A) = \log f(\mathbf{Y}, A) = \sum_{i=1}^n \log f(Y_i, A)$.

Proposition 3.2. *Let $L_n(\cdot)$ be as above. Then for any $\delta > 0$*

$$\inf_{A \notin B(A_0, \delta)} \frac{\min_{\substack{\mathbf{i}=\{i_1, \dots, i_p\} \\ 1 \leq i_1 < \dots < i_p \leq n}} \log f(A, \mathbf{Y}_{\mathbf{i}})}{|L_n(A) - L_n(A_0)|} \xrightarrow{A_0} 0.$$

Proposition 3.3. *Let $\mathbf{Y}_0 = (Y_1, Y_2, \dots, Y_p)$ and $\pi(A) = E_{A_0} J(\mathbf{Y}_0, A)$. Assume that there is a one-to-one correspondence between A and $\Sigma = AA^T$. Then the Jacobian function $J(\mathbf{Y}, A) \xrightarrow{a.s.} \pi(A)$ uniformly on compacts in A .*

The Bernstein-von Mises Theorem provides conditions under which the Bayesian posterior distribution is asymptotically normal (van der Vaart 1998, Ghosh 2003). The fiducial Bernstein-von Mises Theorem is an extension that includes a list of conditions under which the GFD is asymptotically normal (Sonderegger and Hannig, 2012). Those conditions can be divided into three parts to ensure each of the following:

- (a) the Maximum Likelihood Estimator (MLE) is asymptotically normal
- (b) the Bayesian posterior distribution becomes close to that of the MLE
- (c) the fiducial distribution is close to the Bayesian posterior

It is clear that the MLE of $f(\mathbf{Y}, A)$ is asymptotically normal. Under our model, the conditions for (b) holds due to Proposition (3.1) and the construction of the Jacobian formula; the conditions for (c) are satisfied by Propositions (3.2, 3.3). Closely following (Sonderegger and Hannig, 2012), we arrive at Theorem (3.1).

Theorem 3.1. *(Consistency) Let \mathcal{R}_A be an observation from the fiducial distribution $r(A|\mathbf{Y})$ and denote the density of $B = \sqrt{n}(\mathcal{R}_A - \hat{A}_n)$ by $\pi^*(B, \mathbf{Y})$, where \hat{A}_n is a maximum likelihood estimator. Let $I(A)$ be the Fisher information matrix. Under the assumption that there is one-to-one correspondence from the covariance matrix Σ to the covariate matrix A ,*

$$\int_{\mathbb{R}^{p \times p}} \left| \pi^*(B, \mathbf{Y}) - \frac{\sqrt{\det I(A_0)}}{\sqrt{2\pi}} \exp\{-\mathbf{Y}^T I(A_0) \mathbf{Y} / 2\} \right| dB \xrightarrow{P_{A_0}} 0 \quad (3.29)$$

Detailed proof can be found in Section 3.8.

3.5 Reversible jump Markov chain Monte Carlo

With the GFD derived in the two special cases, Gibbs sampler can be applied to estimate the covariance matrix $\Sigma = AA^T$. The GFD approach works well on finding sparse structure in the clique model. However, outside the special models, the GFD is not based on the inverse Wishart distribution. Although the GFD for the general situation is rather complicated, the simplification for the Jacobian from (3.13) to (3.25) and the approximation (3.26) make sampling methods feasible, even for p large. We propose to utilize an adaptive reversible jump Markov chain Monte Carlo (RJMCMC) method to efficiently sample from the GFD, with sparsity assumption on A and under a high dimension parameter space frame work.

RJMCMC is an extension of standard Markov chain Monte Carlo methods that allows simulation of the target distribution on spaces of varying dimensions. It was first introduced by Peter J. Green (Green, 1995). The “jumps” refers to moves between models with

possibly different parameter spaces. To maintain detailed balance of a irreducible and aperiodic chain that converges to the correct target distribution, the moves are required to be reversible. Among many others, Richardson and Green (Richardson and Green, 1997), Dellaportas *et al.* (Dellaportas et al., 2002), Robers *et al.* (Robers et al., 2001), Troughton and Godsill (Troughton and Godsill, 1998), Insua and Müller (Insua and Müller, 1998), Barbieri and O’Hagon (Barbieri and O’Hagan, 1996) and Huerta and West (Huerta and West, 1999) applied RJMCMC to mixture models, variable selection, curve fitting, autoregressive models, neural networks, autoregressive moving average models and component structure in autoregressive models, respectively. Since the number of fixed zeros in the matrix A , the property of jumping between parameter spaces with different dimension is desired for estimate Σ .

3.5.1 Algorithm flow

Let the set of possible models be $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \dots\}$. Write model $\mathcal{M}_k = (A_k, d_k)$, where d_k is the dimension of model \mathcal{M}_k , i.e. the number of non fixed zero elements in the covariate matrix A_k . Denote the GFD of model \mathcal{M}_k as $r(\mathcal{M}_k)$. Similar to maximum likelihood estimation, GFI tends to favor models with more parameters over one with fewer parameters. Therefore an outside penalty accounting for our preference toward parsimony needs to be incorporated in the model. Denote a suitable penalty as $h(\mathcal{M})$. Then the RJMCMC method involve Metropolis-Hastings (MH) algorithm that move a simulation analysis between $\mathcal{M}_k = (A_k, d_k)$ and $\mathcal{M}_{k'} = (A_{k'}, d_{k'})$ can be described as follows:

1. Propose a visit to model $\mathcal{M}_{k'}$ from model \mathcal{M}_k with probability $p_{d_k \rightarrow d_{k'}}$.
2. Sample u from a proposal density $q(u|A_k, d_k, d_{k'})$.
3. Set $(A_{k'}, u') = g_{d_k, d_{k'}}(A_k, u)$, where the jump map $g_{d_k, d_{k'}}(\cdot)$ is a bijection between (A_k, u) and $(A_{k'}, u')$, where u and u' play the role of matching the dimensions of models \mathcal{M}_k and $\mathcal{M}_{k'}$.
4. Accept model $A_{k'}, d_{k'}$ with rate $\alpha(A_k, A_{k'})$, which is

$$\min \left\{ 1, \frac{r(\mathcal{M}_{k'})h(\mathcal{M}_{k'})}{r(\mathcal{M}_k)h(\mathcal{M}_k)} \frac{p_{d_{k'} \rightarrow d_k} q(u'|A_{k'}, d_k, d_{k'})}{p_{d_k \rightarrow d_{k'}} q(u|A_k, d_k, d_{k'})} \left| \frac{\partial g_{d_k, d_{k'}}(A_k, u)}{\partial (A_k, u)} \right| \right\}.$$

Here $r(\cdot)$ denotes the GFD of a model, $p_{d_a \rightarrow d_b}$ is the probability of moving from a parameter space d_a to a model space with d_b . Looping through the four steps generates a Markov chain that enables the estimation of covariance matrix $\Sigma = AA^T$. Similar to the clique model, we apply an MDL type penalty:

$$h(\mathcal{M}_k) = \sum_{i=1}^p \left[\frac{1}{2} p_i \log(np) + \log \binom{p}{p_i} \right] \quad (3.30)$$

where A_k is a $p \times p$ matrix with p_i many non-fixed-zero elements in its i th row, and n is the number of observations.

3.5.2 Jump map

The choice of Jump map is rather tricky. The ideal jumps would be constrained upon minimum changes in $A_k A_k^T$, sparsity, and computational efficiency. We defined the jump map $g_{d_k, d_{k'}}(\cdot, \cdot)$ by allowing only the following three types of moves:

- *Update*

The *update* move is essentially the same as the moves in standard MCMC, where the Markov chain moves between two spaces of the same dimension. Here we refer it to changing one non-zero entry in A_k to another non-zero value while other entries remain the same.

- *Birth*

If *birth* move occurs, the dimensionality of the parameter space, i.e. the number of non-zeros in A_k , increases by one. One of the zero entries of matrix A_k gets to be replaced by a non-zero entry while other entries remain the same.

- *Death*

Death move decreases the dimension of parameter space by one. It changes a non-zero entry to zero in matrix A_k while other entries remain the same.

At each iteration of the Markov chain, a type of move is randomly chosen taking account of the assumed dimension restrictions. The probability of choosing a move is taken to be

independent of which element of matrix A gets proposed to be updated. Table 3.1 shows the assigned probabilities for the moves at each state.

Move	$d_k=\text{MinDim}$	$d_k=\text{MaxDim}$	$\text{MinDim} < d_k < \text{MaxDim}$
Update	1/2	1/2	1/2
Birth	1/2	0	1/4
Death	0	1/2	1/4

Table 3.1: Probabilities of assigning each move, $p_{d_k \rightarrow d_{k'}}$. MinDim and MaxDim are the minimum and maximum dimension allowed for covariate matrix A (model \mathcal{M}), respectively.

The probability of choosing “update” is always $\frac{1}{2}$. If the model is at maximum dimension allowed, *death* is also proposed each with probability $\frac{1}{2}$; if the model is at minimum dimension allowed, *birth* is proposed with probability $\frac{1}{2}$; if neither, *birth* and *death* each has $\frac{1}{4}$ of a chance to be selected. When an entry a_{ij} is picked to be updated, a new value \tilde{a}_{ij} is generated according to the move type. In the case of *death*, $\tilde{a}_{ij} = 0$; for *update* and *birth*, \tilde{a}_{ij} is sampled from $N(a_{ij}, s_1^2)$ and $N(a'_{ij}, s_2^2)$, respectively. Consequently, the term $\left| \frac{\partial g_{d_k, d_{k'}}(A_k, u)}{\partial(A_k, u)} \right|$ reduces to 1 for all moves. The parameters s_1^2 , a'_{ij} and s_2^2 can be selected adaptively to achieve good acceptance ratio and efficiency. Currently, s_1^2 is chosen a priori; a'_{ij} is the optimizer that maximizes the fiducial likelihood; s_2^2 is chosen adaptively using the zeroth-order method in (Brooks et al., 2003). See Section 3.5.3 for further discussion on zeroth-order method. The proposal procedure is summarized in Table 3.2.

Move	Entry to update ($a_{ij,k}$)	Proposed value (\tilde{a}_{ij})
Update	randomly choose one from Update Set	$\tilde{a}_{ij} \sim N(a_{ij,k}, s_1^2)$
Birth	randomly choose one from Birth Set	$\tilde{a}_{ij} \sim N(a'_{ij}, s_2^2)$
Death	choose one from Death Set according to its likelihood	$\tilde{a}_{ij} = 0$

Table 3.2: Covariate proposal detail. Update Set, Birth Set, and Death Set are lists keeping track of which entries in A_k allow update, birth, and death moves, respectively. The variances s_1^2 and s_2^2 are predetermined. The center for birth, a'_{ij} , is chosen to be the one either maximizes the normal likelihood of proposed covariate matrix.

3.5.3 Zeroth-order method

The zeroth-order method is a simple and easy-to-implement method for automatically choosing proposal scales (Brooks et al., 2003). It ensures that the acceptance probability equals one for centered jumps between A_k and $\tilde{A}_{k'}$, where $\tilde{A}_{k'}$ is almost identical to A_k with a_{ij}

replaced by its center a'_{ij} ,

$$\text{i.e.} \quad \alpha(A_k, \tilde{A}_{k'}) \equiv 1 \quad (3.31)$$

Denote the dimensions of A_k and $\tilde{A}_{k'}$ as d_k and $d_{k'}$, respectively. The proposal standard deviation s_2 is the solution to the equation 3.31 when $d_{k'} = dk + 1$. More precisely,

$$\begin{aligned} & \frac{r(A_k)h(\mathcal{M}_k)p_{d_k \rightarrow d_{k'}}q(a'_{ij}|A_k, d_k, d_{k'})}{r(\tilde{A}_{k'})h(\mathcal{M}_{k'})p_{d_{k'} \rightarrow d_k}} = 1 \\ \Rightarrow \quad s_2 &= \frac{r(A_k)h(\mathcal{M}_k)p_{d_k \rightarrow d_{k'}}}{r(\tilde{A}_{k'})h(\mathcal{M}_{k'})p_{d_{k'} \rightarrow d_k}} (2\pi)^{-\frac{1}{2}} \end{aligned}$$

The zeroth-order method adaptively proposes a value nearby the likelihood optimizer, enhances the acceptance rate for the proposals, and increases the efficiency of the RJMCMC.

3.6 Implementation

3.6.1 Special Case I: No fixed zero entries in A

With the assumption that none of the entries in A is fixed at zero, the GFD of Σ follows the inverse Wishart distribution with n degrees of freedom and parameter nS_n (see Section 3.3). Sampling from the GFD becomes straight forward and it can be done through one of the inverse Wishart random generation functions, e.g. `InvWishart` (`MCMCpack`, R) or `iwishrnd` (Matlab).

When p is small and n is large, the estimation of Σ can always be done through this setting, regardless if there are zero entries in A . The concept of having entries of A fixed at zero is to impose sparsity structure and allow estimation under a high dimensional setting without requiring large number of observations.

3.6.2 Special Case II: Clique model

Estimation of cliques is closely related to applications in network analysis, such as cliques of people in social networks and gene regulatory network. Under the clique model introduced

in Section 3.3,

$$r(M) \propto \frac{\prod_{i=1}^k \left[\Gamma_{g_i} \left(\frac{n}{2} \right) C_i(\mathbf{Y}) (2\pi)^{\frac{g_i(g_i-1)}{2}} |\det S_n^i|^{-\frac{n}{2}} \right]}{\exp \left\{ \frac{1}{2} \left(\sum_{i=1}^k g_i^2 \right) \log n + (p+1) \log k \right\}}$$

Assuming that both the number of cliques k and the clique sizes g_k 's are unknown, the clique structure can be estimated via Gibbs sampler. Example 1, 2, 3, below show the simulation results for a small, a medium, and a large covariance matrix. For each example, the ij th entry of Σ satisfies that

$$\Sigma_{ij} = \begin{cases} 1 & , i = j \\ 0.5 & , i \text{ \& } j \text{ belong to the same clique} \\ 0 & , \text{ otherwise} \end{cases}$$

Total 10 Gibbs sampler Markov chains with random initial states were implemented simultaneously. Results from each parallel Markov chain Monte Carlo (MCMC) were consolidated post verification of convergence.

- Example 1 Small Σ : $k = 3, p = 10, n = 50$.

In the first clique model example we consider a block 10×10 covariance matrix. From top down, left to right, Figure 3.1 shows the trace plot for $r(M)$ without normalizing constant, and the heat maps for true covariance Σ , sample covariance S_n , and the fiducial probability of the estimated cliques based on the 10 chains. Besides each heat map, the interpretation of the gray scale is provided. The trace plot indicates good mixing. In the last panel, the three true cliques present fiducial probabilities approximate to 1. The gray stripes around the two large cliques correspond to relatively high value in the same entries of sample covariance matrix.

- Example 2 Medium Σ : $k = 5, p = 50, n = 100$.

The second clique example has similar setup as before. The trace plot in the top left of Figure 3.2 indicates good mixing. There are five cliques. In the bottom right panel, the heat map of fiducial probability of pairwise coordinates belong to the same clique, The general shape of the five cliques matches the truth. The largest clique appears to be more sparse. The smaller values are likely caused by the small entries in S_n .

- Example 3 Large Σ : $k = 10, p = 100, n = 200$. In this example, the cyan chain might have been stuck at a local optima. The 10 chains do not mix well as previous clique models. However, the estimated clique structure matches the true Σ while underestimated the probabilities for some coordinate pairs in the larger cliques.

3.6.3 General case with sparsity known

In the general case,

$$r(A) \propto \frac{\exp \left[-\frac{1}{2} \text{tr} \{ n S_n (A A^T)^{-1} \} \right]}{|\det(A)|^n \binom{n}{p}} \sum_{\substack{\mathbf{i}=(i_1, \dots, i_p) \\ 1 \leq i_1 < \dots < i_p \leq n}} \prod_{i=1}^p \binom{p}{p_i} \left| \det (U_{\mathbf{i}, i})_{\mathbf{r}_i} \right|$$

Assuming that there are fixed zeros in A , then for a $p \times p$ matrix A , the number needed to be estimated is less than p^2 . If there are many fixed zeros, then this number is much smaller, hence the estimation is feasible even if the number of observations n is less than p . In other words, the sparsity assumption on A allows estimations under a large p small n setting. Suppose the zero entry locations of A are known. The rest of A can be obtain via standard MCMC techniques, such as Metroplis.

To access the sampling result, we examine the following six statistics: GFD of A without normalizing constant (figure titled GFD), number of nonzeros in A (Dim), distance between each estimated covariance $\hat{\Sigma}$ and the sample covariance (D2Sn), distance from $\hat{\Sigma}$ to true covariance (D2Sig), determinant of Σ on the log scale (LogD), largest eigenvalue (Eig1), ratio between the largest two eigenvalues (Eig1/Eig2), angle between the leading eigenvector of $\hat{\Sigma}$ and the leading eigenvector of Σ (Eigvec angle), and the condition number of $\hat{\Sigma}$ (Cond).

With a family of samples from each Markov chain, we plotted the confidence curve for each statistic. Suppose the empirical cumulative density function is f_{ecdf} , then the confidence curve function f_{cc} is defined as the following:

$$f_{cc}(x) = \begin{cases} 1 - f_{ecdf}(x), & \text{if } f_{ecdf}(x) \geq 0.5, \\ 2f_{ecdf}(x), & \text{otherwise.} \end{cases} \quad (3.32)$$

The confidence curve provides a more direction visualization for confidence regions.

The simulations shown here each consist of five different initial states: a square-root of sample covariance matrix with entries replaced by 0 if the corresponding entry in A is zero (**SnPa**, in blue), a diagonal matrix with the diagonal of lower Cholesky decomposition on the diagonal (**dcho**, in red), a diagonal matrix with the diagonal of square-root of sample covariance matrix (**diag**, in yellow), a random matrix with the same sparsity structure as A (**rand**, in magenta), and A (**true**, in green). In the plots for GFD, Dim, D2Sn, LogD, Eig1, Eig1/Eig2, Eigvec angle, and cond, oracle is shown in cyan vertical lines. For D2Sig, the oracle is 0. The cyan line plotted indicates the distance between sample covariance and true covariance.

Example 1 Small Σ : $p = 4, n = 20$, (Figure 3.4).

All six chains have converged. The peaks of the confidence curves in GFD, D2Sn, LogD, Eig1, Eigvec angle panels are close to the oracle; all estimated covariate have the correct number of nonzeros as shown in the Dim panel; the peaks of distance to Σ are slightly larger than the distance between S_n and Σ ; the ratios of the largest two eigenvalues peak on the slight left of the oracle; the condition number of estimated covariance matrices are better than the truth.

Example 2 Medium Σ : $p = 50, n = 50$, (Figure 3.5).

Although the chain **rand** has reached the correct dimension for A like the others, it has not converged. This is not surprising since the random starting points is likely far away from the truth. With a larger parameter space, the time needed to reach convergence can be very long. For the other five Markov chains, the confidence curves in panels GFD, D2Sn, Eig1 concentrate around the oracles. Panel D2Sig shows that, including rand, the samples from all six chains recorded are closer to Σ than S_n . The determinant of estimated covariance and the ratios of the leading eigenvalues peak on the slight left of the oracle. Once again, the samples collected have higher condition number than the true covariance matrix.

Example 3 Large Σ : $p = 100, n = 100$, (Figure 3.6).

Similar to previous example, all chains but **rand** has converged. All the samples recorded are closer to Σ comparing to S_n as shown in the D2Sig panel and the Eigvec

angle panel, while the chains either all over-estimated or all under-estimated in panels LogD, Eig1, and Eig1/Eig2. In the Eigvec angel panel, all confidence curves are on the left hand side of the cyan vertical line, indicating that comparing to the leading vector of S_n , the estimated leading eigenvectors have closer direction to the leading vector of Σ .

3.6.4 General case with sparse locations unknown

In the general case with sparse locations unknown, we further assume that there is a maximum number of nonzeros per column allowed, denoted as $maxC$. This additional constraint can be viewed as each predictor only contribute to few tuples of the multivariate response. This assumption has been implemented to reduce the search space for RJMCMC.

For each example below, we consider five Markov chains started with: a sample covariance matrix with only the smallest $p - maxC$ entries per column replaced by 0 (**MaxC**, in blue), a lower Cholesky decomposition matrix with the furthest $p - maxC$ off-diagonal entries replaced by 0 (**chol**, in red), and as before, **dcho** (in yellow), **diag** (in magenta), and **true** (in green).

- Example 1 Small Σ : $p = 4, n = 20, maxC = 2$, (Figure 3.7).

The GFD panel shows that the confidence curves of all chains concentrate on the left hand side of the truth; The estimated dimensions are slightly larger than the truth; the estimated covariances have similar distance to S_n and Σ ; the comparisons in the LogD, Eig1, Eig1/Eig2. Eigvec angle, and Cond panels indicate that the estimates present similar statistics in these categories comparing to oracle/ S_n .

- Example 2 Moderate Σ : $p = 15, n = 30, maxC = 3$, (Figure 3.8).

Similar to the previous example, in the GFD panel the estimates concentrate on the slight left of the truth and the estimated dimensions are slightly larger than the truth. Panels D2Sn, D2Sig, and Eigvec angle show that comparing to S_n the estimated covariances behave more similar to Σ . In the LogD and the Eig1/Eig2 panels, the peaks of the curves are on slight right of the truth. In the last panel, we see that the estimated covariances have better condition number than Σ .

- Example 3 Medium Σ : $p = 50, n = 50, \max C = 5$, (Figure 3.9)

The estimated covariate matrices have smaller GFD than the truth. The estimators have larger dimension than the truth. Base on the D2Sn and the D2Sig panels, the estimated covariance matrices are closer to Σ than S_n to Σ in terms of the FM-distance.

3.7 Discussion

In this chapter we approach covariance estimation via the generalized fiducial inference perspective. For two special cases and the general scenario, we derive the GFD for the covariance matrix. Considering a sparse covariate structure, we explore the asymptotic property of the GFD and showed that it satisfies the Fiducial Bernstein von-Mises Theorem. Notice that, like maximum likelihood estimation, the GFD tend to favor models with larger dimensions. We chose a MDL type penalty to discourage the models in a larger parameter space.

To sample from the GFD, we suggest to use MCMC methods. When the sparsity structure of the covariate matrix is unknown, a RJMCMC procedure is needed. With larger parameter search space, the standard algorithm can be infeasible. We propose an adaptive RJMCMC algorithm that incorporates the zeroth-order method to improve the efficiency. Our simulation results show that in general, the MDL penalty might be slightly liberal, but the resulting fiducial estimators behave at least as good as the sample covariance matrix. As the dimension of the parameter space grows, comparing to the sample covariance, the fiducial samples are much closer to the truth.

The family of estimators enable the definition of confidence regions. We will continue this project towards to producing meaningful confidence regions for covariance matrices.

3.8 Proofs

3.8.1 Proof of Proposition 3.1

Let $\Sigma = AA^T$, $\Sigma_0 = A_0A_0^T$. Denote S_n as the sample covariance matrix as before, $n \in \mathbb{N}$.

Since S_n is the maximum likelihood estimator, we have

$$S_n \xrightarrow{P_{A_0}} \Sigma_0,$$

$$\text{i.e. } \forall r > 0, P_{A_0}(\{\omega : \mathbf{d}(S_n(\omega), \Sigma_0) \geq r\}) \rightarrow 0.$$

Define $L_{\delta,n} = \{\omega : \mathbf{d}(S_n(\omega), \Sigma_0) < \delta/2\}$. For an arbitrary $\omega \in L_{\delta,n}$, assume that λ_i^\dagger 's and λ_i^* 's are the eigenvalues of $S_n(\omega)\Sigma^{-1}$ and $S_n(\omega)\Sigma_0^{-1}$, respectively. Suppose that $A \notin B(A_0, \delta)$, then

$$\delta < \mathbf{d}(\Sigma, \Sigma_0) \leq \mathbf{d}(\Sigma, S_n(\omega)) + \mathbf{d}(S_n(\omega), \Sigma_0) < \mathbf{d}(\Sigma, S_n(\omega)) + \delta/2$$

$$\Rightarrow \mathbf{d}(\Sigma, S_n(\omega)) = \sqrt{\sum_{i=1}^p \ln^2 \lambda_i^\dagger} > \delta/2$$

So there exists $k \in \{1, 2, \dots, p\}$, such that $\ln^2 \lambda_k^\dagger > \frac{\delta^2}{4p}$, then

$$\ln \lambda_k - \lambda_k < \max \left\{ \frac{\delta}{2\sqrt{p}} - \exp \left(\frac{\delta}{2\sqrt{p}} \right), -\frac{\delta}{2\sqrt{p}} - \exp \left(-\frac{\delta}{2\sqrt{p}} \right) \right\} := m_\delta,$$

due to the fact that the function $g(\lambda) = \ln \lambda - \lambda$ is concave with unique maxima $\lambda = 1$;
 $g(1) = -1$.

Meanwhile,

$$\begin{aligned} & \frac{1}{n}(L_n(A) - L_n(A_0))(\omega) \\ &= -\ln |\det(A)| - \frac{1}{2}\text{tr}\{S_n(\omega)\Sigma^{-1}\} + \ln |\det(A_0)| + \frac{1}{2}\text{tr}\{S_n(\omega)\Sigma_0^{-1}\} \\ &= \frac{1}{2}\ln(S_n(\omega)\Sigma^{-1}) - \frac{1}{2}\text{tr}\{S_n(\omega)\Sigma^{-1}\} - \frac{1}{2}\ln(S_n(\omega)\Sigma_0^{-1}) + \frac{1}{2}\text{tr}\{S_n(\omega)\Sigma_0^{-1}\} \\ &= \frac{1}{2} \left\{ \sum_{i=1}^p (\ln \lambda_i^\dagger - \lambda_i^\dagger) - \sum_{i=1}^p (\ln \lambda_i^* - \lambda_i^*) \right\} \\ &< \frac{1}{2} \{-(p-1) + m_\delta + p\} \\ &= \frac{1}{2}(m_\delta + 1) \end{aligned}$$

$$\Rightarrow \sup_{A \notin B(A_0, \delta)} \frac{1}{n} (L_n(A) - L_n(A_0))(\omega) \leq \frac{1}{2}(m_\delta + 1) < 0.$$

Let $\epsilon = -\frac{1}{2}(m_\delta + 1)$, $U_{\delta, n} = \left\{ \omega : \sup_{A \notin B(A_0, \delta)} \frac{1}{n} (L_n(A) - L_n(A_0))(\omega) \leq -\epsilon \right\}$. Then $L_{\delta, n} \subseteq U_{\delta, n}$. Notice that

$$1 = \lim_{n \rightarrow \infty} P_{A_0}(L_{\delta, n}) = \liminf_{n \rightarrow \infty} P_{A_0}(L_{\delta, n}) \leq \liminf_{n \rightarrow \infty} P_{A_0}(U_{\delta, n}) \leq \limsup_{n \rightarrow \infty} P_{A_0}(U_{\delta, n}) \leq 1,$$

Therefore, $\lim_{n \rightarrow \infty} P_{A_0}(U_{\delta, n}) = 1$.

3.8.2 Proof of Proposition 3.2

Note that

$$\inf_{A \notin B(A_0, \delta)} \frac{\min_{\substack{\mathbf{i}=\{i_1, \dots, i_p\} \\ 1 \leq i_1 < \dots < i_p \leq n}} \log f(A, \mathbf{Y}_{\mathbf{i}})}{|L_n(A) - L_n(A_0)|} \leq \frac{\inf_{A \notin B(A_0, \delta)} \min_{\substack{\mathbf{i}=\{i_1, \dots, i_p\} \\ 1 \leq i_1 < \dots < i_p \leq n}} \log f(A, \mathbf{Y}_{\mathbf{i}})}{\inf_{A \notin B(A_0, \delta)} |L_n(A) - L_n(A_0)|}$$

For any $A \notin B(A_0, \delta)$, denote $\Sigma = AA^T$, $\Sigma_0 = A_0A_0^T$ and let $t > 0$, we have

$$\begin{aligned} & P_{A_0} \left(\min_{\substack{\mathbf{i}=\{i_1, \dots, i_p\} \\ 1 \leq i_1 < \dots < i_p \leq n}} \log f(A, \mathbf{Y}_{\mathbf{i}}) \leq -t \log n \right) \\ & \leq P_{A_0} \left(\min_{i=1, \dots, n} \log f(A, Y_i) \leq -\frac{t \log n}{p} \right) \\ & = 1 - \left[1 - P_{A_0} \left(-\log f(A, Y_i) \geq -\frac{t \log n}{p} \right) \right]^n \\ & \leq 1 - \left[1 - \frac{p \mathbb{E}_{A_0}(-\log f(A, Y_i))}{t \log n} \right]^n \quad (\text{Markov inequality}) \\ & = 1 - \left[1 - \frac{p(\log(2\pi) + \log \det(\Sigma) + \text{tr}\{\Sigma^{-1}\Sigma_0\})}{2t \log n} \right]^n \\ & \rightarrow 0, \quad \text{as } n \rightarrow \infty. \end{aligned}$$

So the numerator goes to $-\infty$ at most as fast as $-t \log n$. Meanwhile, for a fixed n and any

$$\omega \in L_{\delta, n} = \{\omega : \mathbf{d}(S_n(\omega), \Sigma_0) < \delta/2\},$$

$$\inf_{A \notin B(A_0, \delta)} |L_n(A) - L_n(A_0)| = - \sup_{A \notin B(A_0, \delta)} L_n(A) - L_n(A_0) \geq \epsilon n$$

By Proposition (3.1),

$$\lim_{n \rightarrow \infty} P_{A_0} \left(\inf_{A \notin B(A_0, \delta)} |L_n(A) - L_n(A_0)| \geq \epsilon n \right) = 1,$$

i.e. the denominator goes to infinity at least as fast as ϵn .

3.8.3 Proof of Proposition 3.3

This proposition states that the Jacobian function is a U-statistic. By Theorem 1 of Yeo and Johnson (Yeo and Johnson, 2001), it suffices to show the following:

Set

$$J_j((y_1, \dots, y_j), A) = E_{A_0} J((y_1, \dots, y_j, Y_{j+1}, \dots, Y_p), A)$$

$$j = 1, \dots, p.$$

(3.3a) There is an integrable and symmetric kernel $g(\cdot)$ and compact space $\bar{B}(A_0, \delta)$ such that, for all A , and $\mathbf{y}_0 = (y_1, \dots, y_p) \in \mathbb{R}^{p \times p}$,

$$|J(\mathbf{y}_0, A)| \leq g(\mathbf{y}_0).$$

(3.3b) There is a sequence S_M^p of measurable sets such that

$$P \left(\mathbb{R}^{p \times p} - \bigcup_{M=1}^{\infty} S_M^p \right) = 0$$

(3.3c) For each M and for all $j = 1, \dots, p$, $J_j((y_1, \dots, y_j), A)$ is equicontinuous in A for $(y_1, \dots, y_j) \in S_M^j$, where $S_M^p = S_M^j \times S_M^{p-j}$.

Denote

$$W_j = \begin{pmatrix} (A_0^{-1} y_1)^T \\ \vdots \\ (A_0^{-1} y_j)^T \\ (A_0^{-1} Y_{j+1})^T \\ \vdots \\ (A_0^{-1} Y_p)^T \end{pmatrix} = \begin{pmatrix} z_1^T \\ \vdots \\ z_j^T \\ Z_{j+1}^T \\ \vdots \\ Z_p^T \end{pmatrix}.$$

Then

$$U_0 = \begin{pmatrix} (A^{-1} y_1)^T \\ \vdots \\ (A^{-1} y_p)^T \end{pmatrix} = W_0 (A^{-1} A_0)^T.$$

For simplicity, let $\sum_{\mathbf{r}_i}$ be short for $\sum_{\substack{\mathbf{r}_i=(r_{i,1},\dots,r_{i,p_i}), \\ 1 \leq r_{i,1} < \dots < r_{i,p_i} \leq p.}}$, which sums over all p_i -tuples of ordered indices between 1 and p , \mathbf{r}_i .

$$\begin{aligned} \Rightarrow J(\mathbf{y}_0, A) &= \prod_{i=1}^p \left\{ \sum_{\mathbf{r}_i} \left| \det (U_{0,i})_{\mathbf{r}_i} \right| \right\} \\ &= \prod_{i=1}^p \left\{ \sum_{\mathbf{r}_i} \left| \det ([W_0(A^{-1}A_0)^T]_{(\cdot, -R_i)})_{\mathbf{r}_i} \right| \right\} \\ &= \prod_{i=1}^p \left\{ \sum_{\mathbf{r}_i} \left| \det ([W_0(A^{-1}A_0)^T]_{(\mathbf{r}_i, -R_i)}) \right| \right\} \end{aligned}$$

Using Cauchy-Binet formula,

$$\det ([W_0(A^{-1}A_0)^T]_{(\mathbf{r}_i, -R_i)}) = \sum_{\tilde{\mathbf{r}}_i} \det ([W_0]_{(\mathbf{r}_i, \tilde{\mathbf{r}}_i)}) \det ([A^{-1}A_0]_{(-R_i, \tilde{\mathbf{r}}_i)}).$$

Therefore,

$$\begin{aligned} &J(\mathbf{y}_0, A) \\ &= \prod_{i=1}^p \left\{ \sum_{\mathbf{r}_i} \left| \sum_{\tilde{\mathbf{r}}_i} \det ([W_0]_{(\mathbf{r}_i, \tilde{\mathbf{r}}_i)}) \det ([A^{-1}A_0]_{(-R_i, \tilde{\mathbf{r}}_i)}) \right| \right\} \\ &\leq \prod_{i=1}^p \left\{ \sum_{\tilde{\mathbf{r}}_i} \left[\left| \det ([A^{-1}A_0]_{(-R_i, \tilde{\mathbf{r}}_i)}) \right| \sum_{\mathbf{r}_i} \left| \det ([W_0]_{(\mathbf{r}_i, \tilde{\mathbf{r}}_i)}) \right| \right] \right\} \\ &\leq \prod_{i=1}^p \left\{ \binom{p}{p_i} \max \left\{ \left| \det ([W_0]_{(\mathbf{r}_i, \mathbf{r}'_i)}) \right| \right\} \sum_{\tilde{\mathbf{r}}_i} \left| \det ([A^{-1}A_0]_{(-R_i, \tilde{\mathbf{r}}_i)}) \right| \right\} \end{aligned}$$

Given an ordered index vector $\mathbf{r} = (r_1, \dots, r_l)$, let $E_{\mathbf{r}} = (e_{r_1}; \dots; e_{r_l})$, where each e_{r_j} is a $1 \times p$ vector with 1 in the r_j th tuple and 0 everywhere else. Define $I_{\mathbf{r}} = E_{\mathbf{r}} y_0 E_{\mathbf{r}}^T$ to be the matrix similar to an identity matrix but with the kk th entry being 0 if $k \notin \{r_1, \dots, r_l\}$.

$$\begin{aligned} &\sum_{\tilde{\mathbf{r}}_i} \left| \det ([A^{-1}A_0]_{(-R_i, \tilde{\mathbf{r}}_i)}) \right| \\ &= \sum_{\tilde{\mathbf{r}}_i} \left| \det (E_{-R_i}^T (A^{-1}A_0) E_{\tilde{\mathbf{r}}_i}) \right| \\ &= \sum_{\tilde{\mathbf{r}}_i} \sqrt{\det [E_{-R_i}^T (A^{-1}A_0) I_{\tilde{\mathbf{r}}_i} (A^{-1}A_0)^T E_{-R_i}]} \end{aligned}$$

By Hadamard's inequality,

$$\begin{aligned}
& \det [E_{-R_i}^T (A^{-1} A_0) I_{\tilde{\mathbf{r}}_i} (A^{-1} A_0)^T E_{-R_i}] \\
& \leq \prod_{k \notin R_i} [(A^{-1} A_0) I_{\tilde{\mathbf{r}}_i} (A^{-1} A_0)^T]_{kk} \\
& \leq \prod_{k \notin R_i} [(A^{-1} A_0) (A^{-1} A_0)^T]_{kk}
\end{aligned}$$

Furthermore, recall that $A \in \bar{B}(A_0, \delta)$, i.e. $\mathbf{d}(A, A_0) \leq \delta$. Then $\forall k$,

$$\begin{aligned}
[(A^{-1} A_0) (A^{-1} A_0)^T]_{kk} &= e_k^T (A^{-1} A_0) (A^{-1} A_0)^T e_k \\
&= e_k^T A^T (\Sigma^{-1} \Sigma_0) (A^{-1})^T e_k \\
&\leq \lambda_1 \\
&\leq e^{\sqrt{\ln^2 \lambda_1}} \\
&\leq e^\delta
\end{aligned}$$

where λ_1 is the leading eigenvalue of $\Sigma^{-1} \Sigma_0$.

$$\begin{aligned}
& \Rightarrow J(\mathbf{y}_0, A) \\
&= \prod_{i=1}^p \left\{ \sum_{\mathbf{r}_i} |\det ([W_0 (A^{-1} A_0)^T]_{(\mathbf{r}_i, -R_i)})| \right\} \\
&\leq \prod_{i=1}^p \left\{ \binom{p}{p_i} \max \left\{ |\det ([W_0]_{(\mathbf{r}_i, \mathbf{r}'_i)})| \right\} \sum_{\tilde{\mathbf{r}}_i} |\det ([A^{-1} A_0]_{(-R_i, \tilde{\mathbf{r}}_i)})| \right\} \\
&\leq \prod_{i=1}^p \left\{ \binom{p}{p_i}^2 e^{\delta p_i / 2} \max \left\{ |\det ([W_0]_{(\mathbf{r}_i, \mathbf{r}'_i)})| \right\} \right\} \\
&:= g(\mathbf{y}_0)
\end{aligned}$$

It is clear that $g(\mathbf{y}_0)$ is integrable and symmetric.

Note that

$$J_j((y_1, \dots, y_j), A) = E_{A_0} J((y_1, \dots, y_j, Y_{j+1}, \dots, Y_p), A)$$

$$\begin{aligned}
&= \mathbb{E} \left\{ \prod_{i=1}^p \left\{ \sum_{\mathbf{r}_i} |\det ([W_j(A^{-1}A_0)^T]_{(\mathbf{r}_i, -R_i)})| \right\} \right\} \\
&= \mathbb{E} \left\{ \prod_{i=1}^p \left\{ \sum_{\mathbf{r}_i} |\det (E_{\mathbf{r}_i} W_j(A^{-1}A_0)^T E_{-R_i})| \right\} \right\}
\end{aligned}$$

can be viewed as a polynomial function of entries of A^{-1} with coefficients being polynomial functions of y_1, \dots, y_j .

Let $S_M^p = \{(y_1, \dots, y_p) : |y_i| \leq M, \forall i\}$, where M is a positive integer. It is clear that S_M^p 's are measurable. By construction, $P(\mathbb{R}^{p \times p} - \bigcup_{M=1}^{\infty} S_M^p) = 0$. For each fixed M , if $(y_1, \dots, y_j) \in S_M^j$, then the coefficients of $J_j((y_1, \dots, y_j), A)$ are bounded, hence $J_j((y_1, \dots, y_j), A)$ is equicontinuous in A .

3.8.4 Proof of Theorem 3.1

Proposition 3.3 and the uniform strong law of large numbers for U-statistics imply that $\pi(A)$ is continuous,

$$\sup_{A \in B(A_0, \delta)} |J(\mathbf{Y}, A) - \pi(A)| \rightarrow 0 \quad a.s. \ P_{A_0}$$

$$\begin{aligned}
\pi^*(B, \mathbf{Y}) &= \frac{J\left(\mathbf{Y}, \hat{A}_n + \frac{B}{\sqrt{n}}\right) f\left(\mathbf{Y} | \hat{A}_n + \frac{B}{\sqrt{n}}\right)}{\int_{\mathbb{R}^{p \times p}} J\left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}}\right) f\left(\mathbf{Y} | \hat{A}_n + \frac{C}{\sqrt{n}}\right) dC} \\
&= \frac{J\left(\mathbf{Y}, \hat{A}_n + \frac{B}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{A}_n + \frac{B}{\sqrt{n}}\right) - L_n(\hat{A}_n)\right]}{\int_{\mathbb{R}^{p \times p}} J\left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{A}_n + \frac{C}{\sqrt{n}}\right) - L_n(\hat{A}_n)\right] dC}
\end{aligned}$$

Notice that

$$H = -\frac{1}{n} \frac{\partial^2}{\partial A \partial A}(\hat{A}_n) \rightarrow I(A_0) \quad a.s. \ P_{A_0}.$$

It suffices to show that

$$\begin{aligned}
\int_{\mathbb{R}^{p \times p}} \left| J\left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{A}_n + \frac{C}{\sqrt{n}}\right) - L_n(\hat{A}_n)\right] \right. \\
\left. - \pi(A_0) \exp\left[\frac{-C^T I(A_0) C}{2}\right] \right| dC \xrightarrow{P_{A_0}} 0
\end{aligned} \tag{3.33}$$

Let C_x be the ij th entry of C , where $x = i + (p-1)j$. By Taylor Theorem,

$$L_n\left(\hat{A}_n + \frac{C}{\sqrt{n}}\right) = L_n(\hat{A}_n) + \sum_{x=1}^{p^2} \left(\frac{C_x}{\sqrt{n}}\right) \frac{\partial}{\partial A_x} L_n(\hat{A}_n)$$

$$\begin{aligned}
& + \frac{1}{2} \sum_{x=1}^{p^2} \sum_{y=1}^{p^2} \left(\frac{C_x C_y}{(\sqrt{(n)})^2} \right) \frac{\partial^2}{\partial A_x \partial A_y} L_n(\hat{A}_n) \\
& + \frac{1}{6} \sum_{x=1}^{p^2} \sum_{y=1}^{p^2} \sum_{z=1}^{p^2} \left(\frac{C_x C_y C_z}{(\sqrt{(n)})^3} \right) \frac{\partial^3}{\partial A_x \partial A_y \partial A_z} L_n(A') \\
& = L_n(\hat{A}_n) - \frac{C^T H C}{2} + R_n
\end{aligned}$$

for some $A' \in [\hat{A}_n, \hat{A}_n + \frac{C}{\sqrt{n}}]$. Notice that $R_n = \mathcal{O}p(n^{-3/2} \|C\|)$. Given any $0 < \delta < \delta_0$ and $t > 0$, the parameter space $\mathbb{R}^{p \times p}$ can be partitioned into three regions:

$$S_1 = \{C : \|C\| < t \log \sqrt{n}\}$$

$$S_2 = \{C : t \log \sqrt{n} < \|C\| < \delta \sqrt{n}\}$$

$$S_3 = \{C : \|C\| > \delta \sqrt{n}\}$$

On $S_1 \cup S_2$,

$$\begin{aligned}
& \int_{S_1 \cup S_2} \left| J\left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}}\right) \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) \right] \right. \\
& \quad \left. - \pi(A_0) \exp \left[\frac{-C^T I(A_0) C}{2} \right] \right| dC \\
& \leq \int_{S_1 \cup S_2} \left| J\left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}}\right) - \pi\left(\hat{A}_n + \frac{C}{\sqrt{n}}\right) \right| \\
& \quad \times \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) \right] dC \\
& \quad + \int_{S_1 \cup S_2} \left| \pi\left(\hat{A}_n + \frac{C}{\sqrt{n}}\right) \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) \right] \right. \\
& \quad \left. - \pi(A_0) \exp \left[\frac{-C^T I(A_0) C}{2} \right] \right| dC
\end{aligned}$$

Since $\pi(\cdot)$ is a proper prior on the region $S_1 \cup S_2$, the second term goes to zero by the Bayesian Bernstein-von Mises Theorem (see the proof of Theorem 1.4.2 in (Ghosh and Ramamoorthi, 2003)).

Next we notice that

$$\int_{S_1 \cup S_2} \left| J\left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}}\right) - \pi\left(\hat{A}_n + \frac{C}{\sqrt{n}}\right) \right|$$

$$\begin{aligned}
& \times \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) \right] dC \\
& \leq \sup_{C \in S_1 \cup S_2} \left| J \left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}} \right) - \pi \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) \right| \\
& \quad \times \int_{S_1 \cup S_2} \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) \right] dC
\end{aligned}$$

Since $\sqrt{n}(\hat{A}_n - A_0) \xrightarrow{\mathcal{D}} N(0, I(A_0)^{-1})$, we have

$$P_{A_0} \left[\left\{ \hat{A}_n + \frac{C}{\sqrt{n}}; C \in S_1 \cup S_2 \right\} \subset B(A_0, \delta_0) \right] \rightarrow 1.$$

Furthermore,

$$L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) = -\frac{C^T H C}{2} + R_n,$$

so the integral converges in probability to 1. Since $\max_{C \in S_1 \cup S_2} \leq \delta$ and $J_n \rightarrow \pi$, the term goes to 0 in probability.

Turning our attention to S_3 , notice that

$$\begin{aligned}
& \int_{S_3} \left| J \left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}} \right) \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) \right] \right. \\
& \quad \left. - \pi(A_0) \exp \left[\frac{-C^T I(A_0) C}{2} \right] \right| dC \\
& \leq \int_{S_3} J \left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}} \right) \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) \right] dC \\
& \quad + \int_{S_3} \pi(A_0) \exp \left[\frac{-C^T I(A_0) C}{2} \right] dC
\end{aligned}$$

The last integral goes to zero in P_{A_0} because $\min_{A_3} \|C\| \rightarrow \infty$. As for the first integral,

$$\begin{aligned}
& \int_{S_3} J \left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}} \right) \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) \right] dC \\
& = \frac{1}{n} \sum_{i=1}^n \int_{S_3} J \left(Y_i, \hat{A}_n + \frac{C}{\sqrt{n}} \right) \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) \right] dC \\
& = \frac{1}{n} \sum_{i=1}^n \int_{S_3} J \left(Y_i, \hat{A}_n + \frac{C}{\sqrt{n}} \right) f \left(Y_i | \hat{A}_n + \frac{C}{\sqrt{n}} \right) \\
& \quad \times \exp \left[L_n \left(\hat{A}_n + \frac{C}{\sqrt{n}} \right) - L_n(\hat{A}_n) - \log f \left(Y_i | \hat{A}_n + \frac{C}{\sqrt{n}} \right) \right] dC
\end{aligned}$$

By Proposition 3.2, the exponent goes to $-\infty$. Because $J(\cdot)$ is a probability measure, the integral converges to 0 in probability. Having shown Eq 3.33, we now follow Ghosh and Ramamoorthi (Ghosh and Ramamoorthi, 2003) and let

$$D_n = \int_{\mathbb{R}^{p \times p}} \left| J\left(\mathbf{Y}, \hat{A}_n + \frac{C}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{A}_n + \frac{C}{\sqrt{n}}\right) - L_n(\hat{A}_n)\right] \right| dC$$

Then the main result to be proven (Eq 3.29) becomes

$$\begin{aligned} D_n^{-1} \left\{ \int_{\mathbb{R}^{p \times p}} \left| J\left(\mathbf{Y}, \hat{A}_n + \frac{B}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{A}_n + \frac{B}{\sqrt{n}}\right) - L_n(\hat{A}_n)\right] \right. \right. \\ \left. \left. - D_n \frac{\sqrt{\det(I(A_0))}}{\sqrt{2\pi}} \exp\left(-\frac{B^T I(A_0) B}{2}\right) \right| \right\} dB \xrightarrow{P_{A_0}} 0 \end{aligned} \quad (3.34)$$

Because

$$\begin{aligned} & \int_{\mathbb{R}^{p \times p}} J(\mathbf{Y}, \hat{A}_n) \exp\left(-\frac{B^T I(A_0) B}{2}\right) dB \\ &= J(\mathbf{Y}, \hat{A}_n) \int_{\mathbb{R}^{p \times p}} \exp\left(-\frac{B^T I(A_0) B}{2}\right) dB \\ &= J(\mathbf{Y}, \hat{A}_n) \frac{\sqrt{2\pi}}{\sqrt{\det(H)}} \\ &\xrightarrow{a.s.} \pi(A_0) \frac{\sqrt{2\pi}}{\sqrt{\det(H)}} \end{aligned}$$

and Eq 3.33 implies that $C_n \xrightarrow{P} \pi(A_0) \frac{\sqrt{2\pi}}{\sqrt{\det(H)}}$, it is enough to show that the integral in Eq 3.34 goes to 0 in probability. This integral is less than $I_1 + I_2$, where

$$\begin{aligned} I_1 &= \int_{\mathbb{R}^{p \times p}} \left| J\left(\mathbf{Y}, \hat{A}_n + \frac{B}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{A}_n + \frac{B}{\sqrt{n}}\right) - L_n(\hat{A}_n)\right] \right. \\ &\quad \left. - J\left(\mathbf{Y}, \hat{A}_n\right) \exp\left(-\frac{B^T I(A_0) B}{2}\right) \right| dB \end{aligned}$$

and

$$\begin{aligned} I_2 &= \int_{\mathbb{R}^{p \times p}} \left| J\left(\mathbf{Y}, \hat{A}_n\right) \exp\left(-\frac{B^T H B}{2}\right) \right. \\ &\quad \left. - D_n \frac{\sqrt{\det(I(A_0))}}{\sqrt{2\pi}} \exp\left(-\frac{B^T I(A_0) B}{2}\right) \right| dB \end{aligned}$$

Eq 3.33 shows that $I_1 \rightarrow 0$ in probability.

Since

$$\begin{aligned} J(\mathbf{Y}, \hat{A}_n) &\xrightarrow{P} \pi(A_0) \\ D_n &\xrightarrow{P} \pi(A_0) \frac{\sqrt{2\pi}}{\sqrt{\det(I(A_0))}} \end{aligned}$$

we have

$$\begin{aligned} I_2 &= \left| J(\mathbf{Y}, \hat{A}_n) - D_n \frac{\sqrt{\det(I(A_0))}}{\sqrt{2\pi}} \right| \int_{\mathbb{R}^{p \times p}} \exp\left(-\frac{B^T H B}{2}\right) dB \\ &\xrightarrow{P} 0. \end{aligned}$$

3.9 Appendix

3.9.1 Förstner-Moonen distance (FM-distance)

As a basic task in mensuration design, the idea of comparing covariance matrices dates back to 1973, when Baarda compared the variances of arbitrary functions $f = \mathbf{e}^T \mathbf{x}$ on one hand determined with a given covariance matrix C and on the other hand determined with a reference or criterion matrix H . One requirement would be the variance $\sigma_f^{2(C)}$ of f when calculated with C to be always smaller than the variance $\sigma_f^{2(H)}$ of f when calculated with H . In other words,

$$\mathbf{e}^T C \mathbf{e} \leq \mathbf{e}^T H \mathbf{e} \quad \text{for all } \mathbf{e} \neq 0,$$

or the Raleigh ratio

$$0 \leq \lambda(\mathbf{e}) = \frac{\mathbf{e}^T C \mathbf{e}}{\mathbf{e}^T H \mathbf{e}} \leq 1 \quad \text{for all } \mathbf{e} \neq 0.$$

The maximum λ from $1/2 \partial \lambda(\mathbf{e}) / \partial \mathbf{e} = 0 \leftrightarrow \lambda H \mathbf{e} - C \mathbf{e} = (\lambda H - C) \mathbf{e} = 0$ results in the maximum eigenvalue $\lambda_{\max}(CH^{-1})$ from the generalized eigenvalue problem

$$\det(\lambda H - C) = 0. \tag{3.35}$$

Note that

$$\lambda \mathbf{e}^T H \mathbf{e} - \mathbf{e}^T C \mathbf{e} = \mathbf{e}^T (\lambda H - C) \mathbf{e} = 0 \quad \text{for } \mathbf{e} \neq 0 \text{ only if (3.35) holds.}$$

The eigenvalues of (3.35) are non-negative if the two matrices C and H are positive semidefinite. Follow this idea Förstner and Moonen suggested a metric for covariance matrices comparison in 1999 (Förstner and Moonen, 1999). For two symmetric positive definite matrices M and N , with the eigenvalues $\lambda_i(M, N)$ from $\det(\lambda H - C) = 0$, the matrix distance between the two matrices M and N is

$$\mathbf{d} = \sqrt{\sum_{i=1}^n \ln^2 \lambda_i(M, N)}.$$

This distance measure is a metric and invariant with respect to both affine transformations of the coordinate system and an inversion of the matrices (Förstner and Moonen, 1999). We will use this metric to compare estimated matrices of the RJMCMC and validate the estimated covariance matrices.

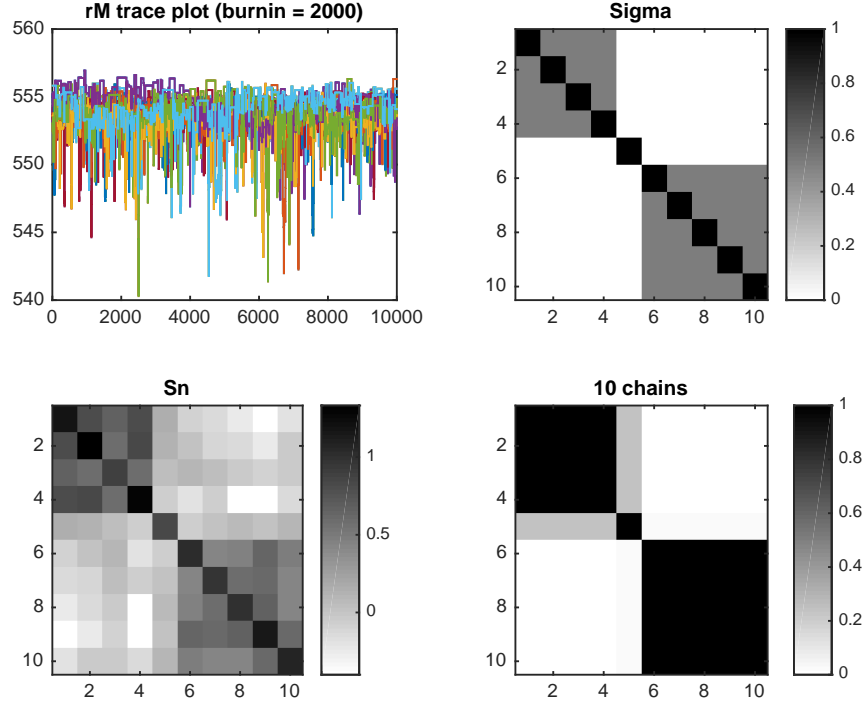


Figure 3.1: Clique result for $k = 3, p = 10, n = 50$. The panels are trace plot for $r(M)$ without normalizing constant and the heat maps for $Sigma$, S_n , and fiducial probability of pairwise coordinates belong to the same clique, from top to bottom, left to right. The fiducial probabilities of the coordinates in the same true clique that belong to a clique are all close to 1.

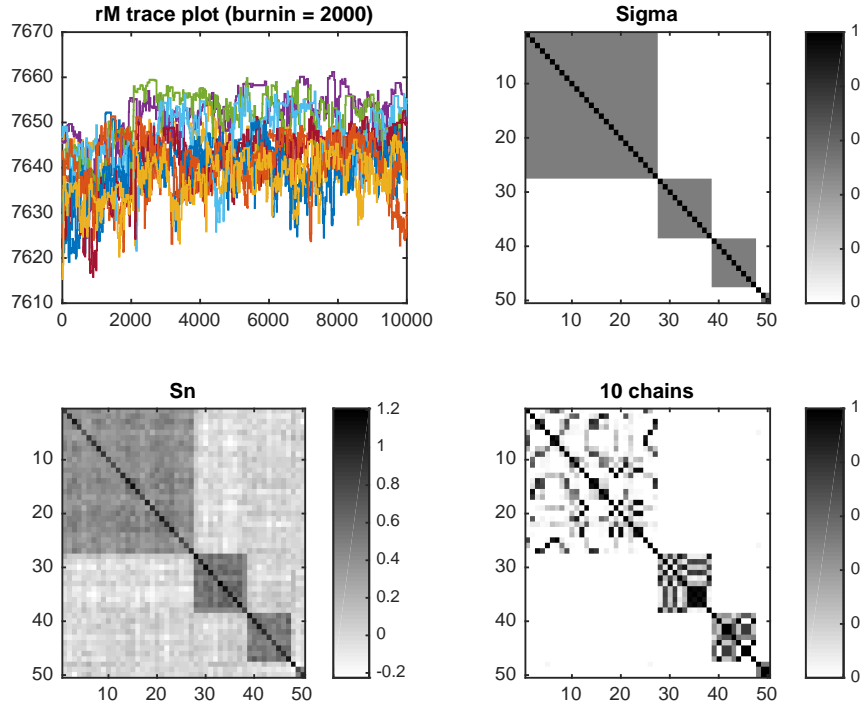


Figure 3.2: Clique result for $k = 5, p = 50, n = 100$. The fiducial probabilities of the coordinates in the same true clique that belong to a clique are generally consistent with the sample covariance matrix. Outside of the diagonal blocks, almost all other entries are at zero. The largest diagonal block appears to be sparse. The low fiducial probabilities are likely caused by the small S_n values at the same locations.

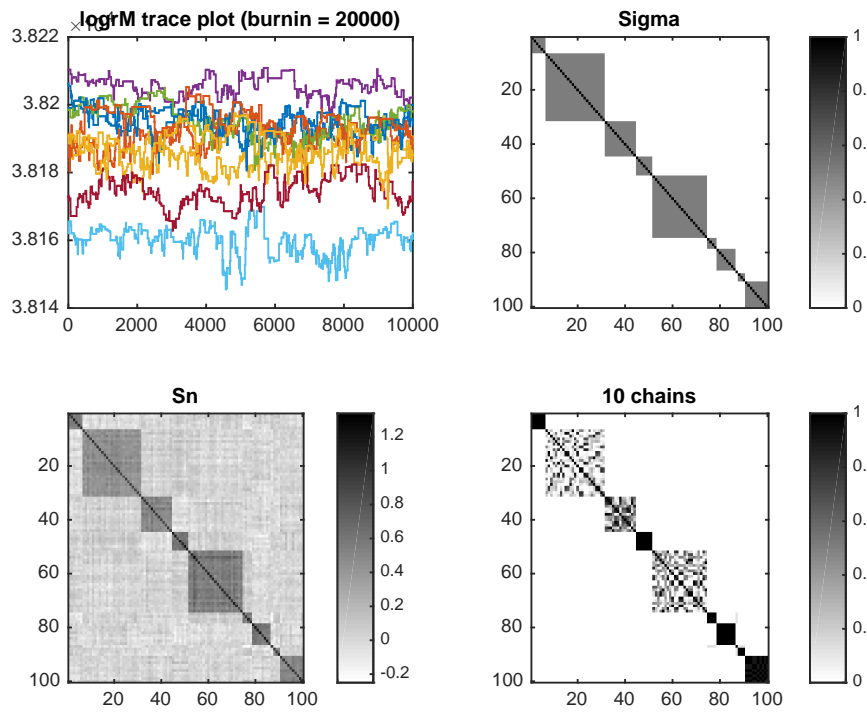


Figure 3.3: Clique result for $k = 10, p = 100, n = 200$. The fiducial probabilities of the coordinates in the same true clique that belong to a clique are generally consistent with the sample covariance matrix.

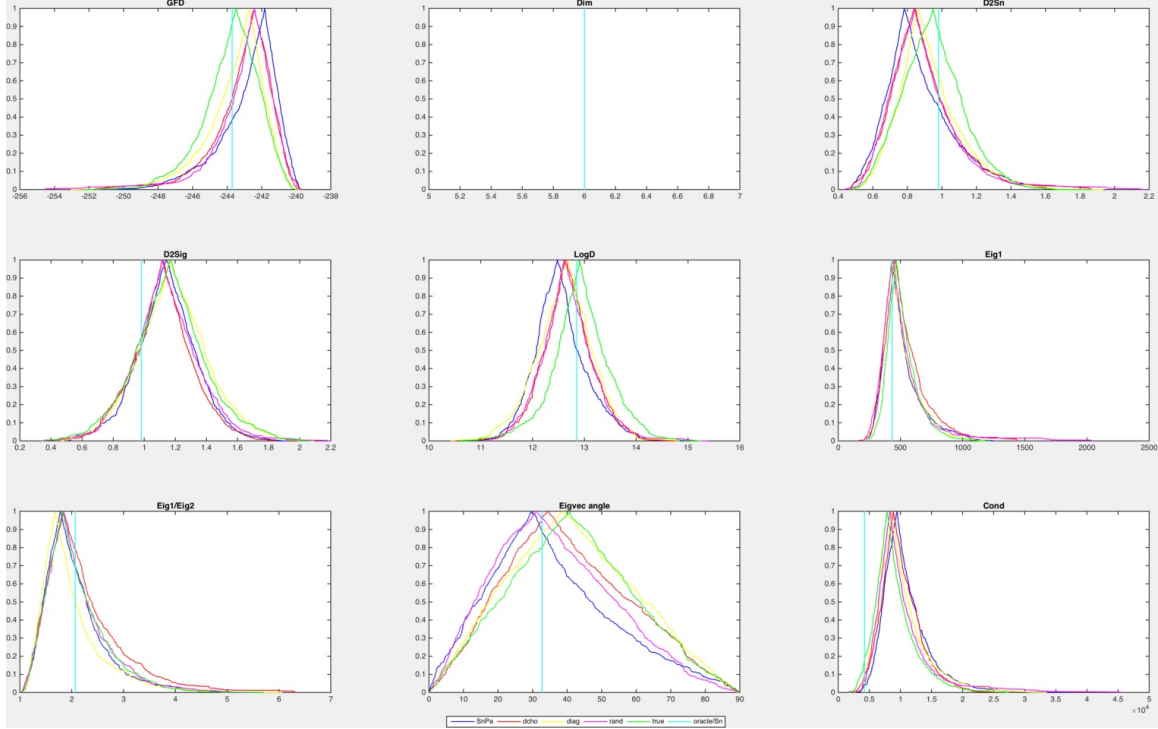


Figure 3.4: Confidence curve plots for $p = 4$, $n = 20$, $burnin = 1000$. All six chains provide good estimations comparing to the oracles.

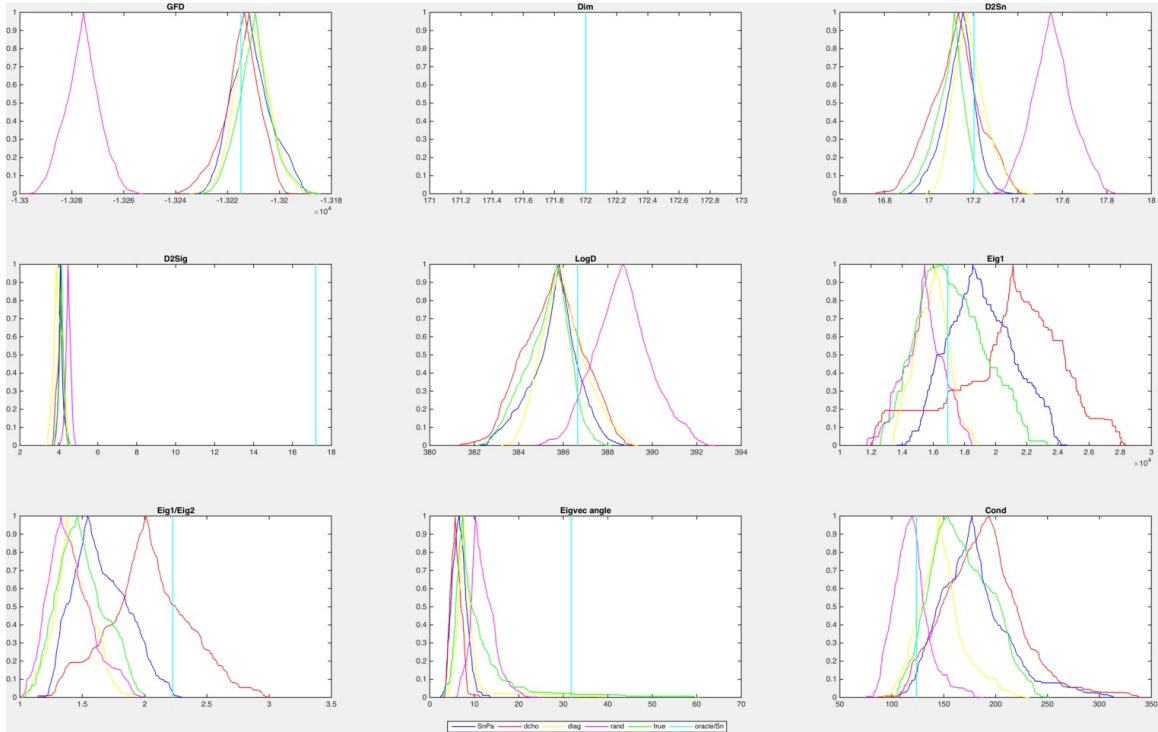


Figure 3.5: Confidence curve plots for $p = 50$, $n = 50$, $burnin = 100000$. All six chains but **rand** provide good estimations comparing to the oracles. The chain **rand** has not yet converged.

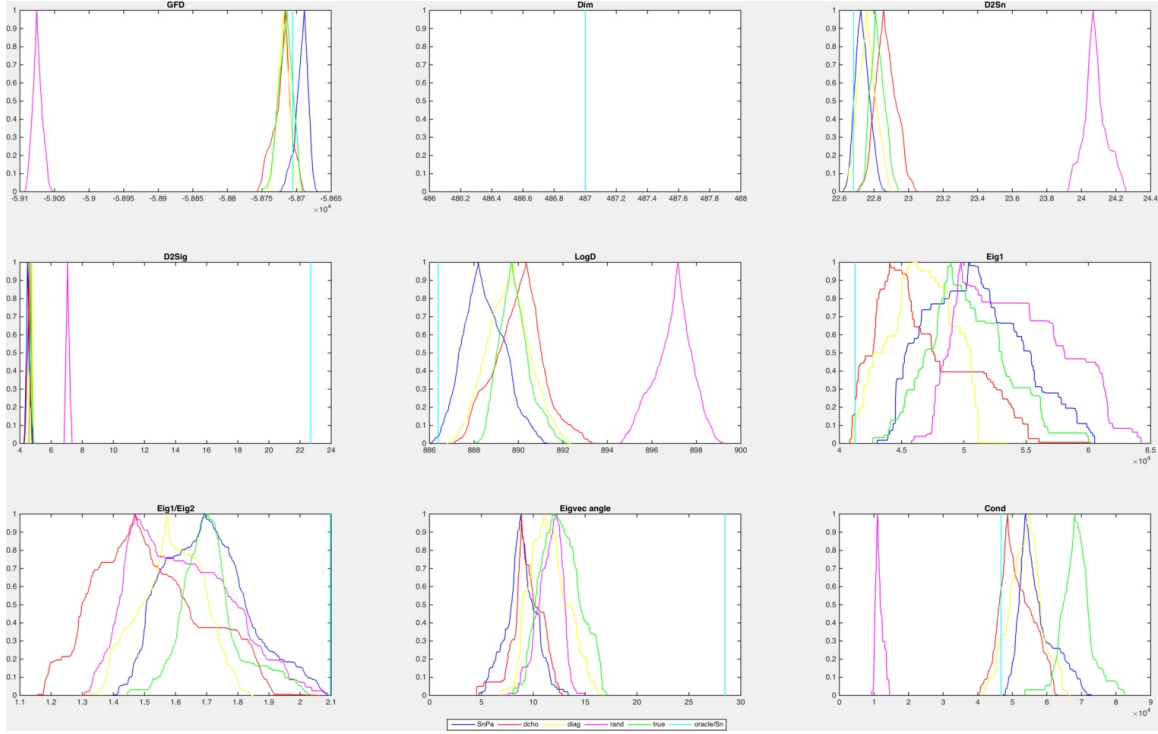


Figure 3.6: Confidence curve plots for $p = 100$, $n = 100$, $burnin = 200000$. Similar to previous example, all chains but **rand** has converged. All the samples recorded are closer to Σ comparing to S_n , while the chains either all over-estimated or all under-estimated in panels LogD, Eig1, and Eig1/Eig2.

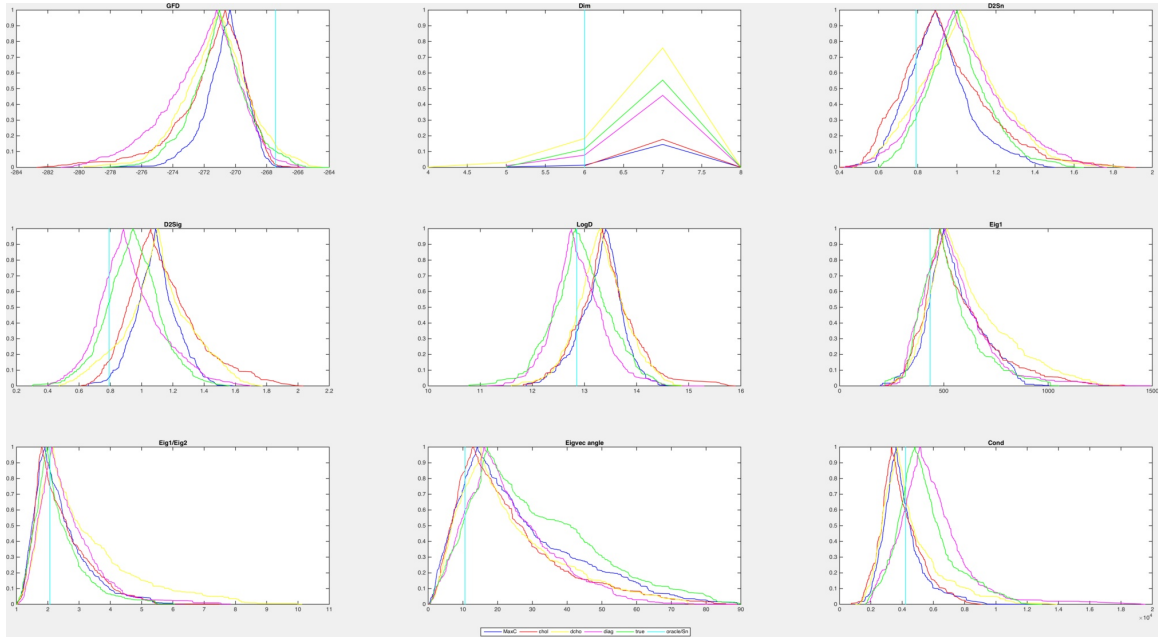


Figure 3.7: Confidence curve plots for RJMCMC with $p = 4$, $n = 20$, $maxC = 2$, $burnin = 1000$. Overall, the sampled covariances concentrate near the oracle/ S_n .

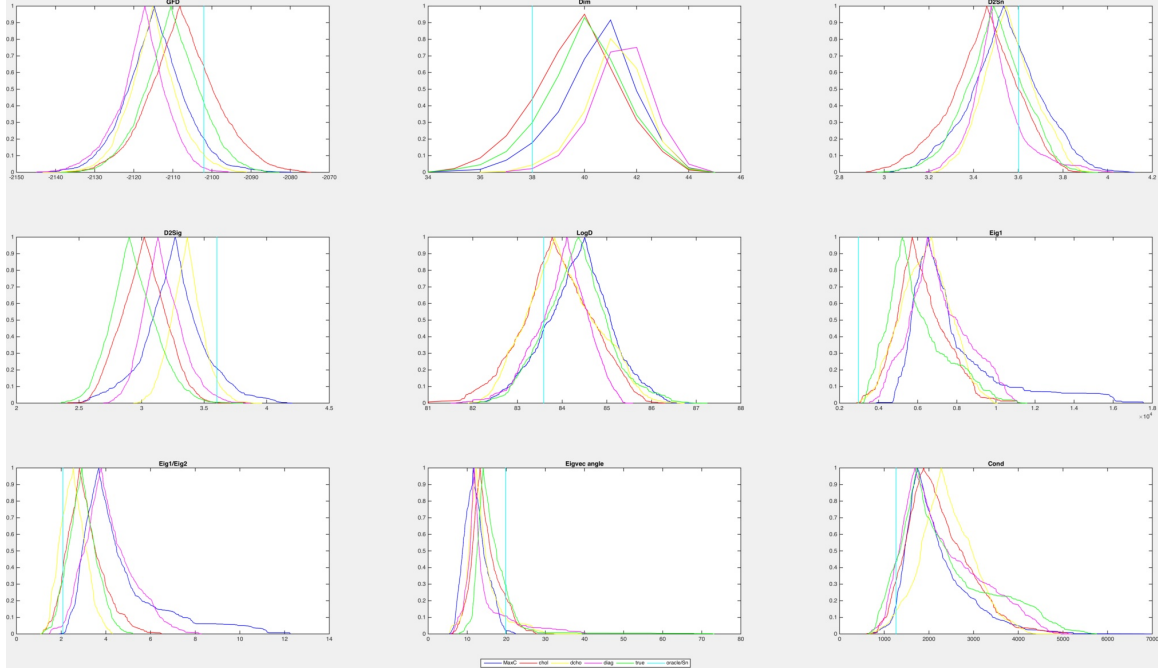


Figure 3.8: Confidence curve plots for RJMCMC with $p = 15$, $n = 30$, $\max C = 3$, $\text{burnin} = 50000$. Although the many of the estimated covariates have few more nonzeros than A , and the GFD's appear to peak at values smaller than the truth, comparing to S_n the estimated covariances behave more similar to Σ .

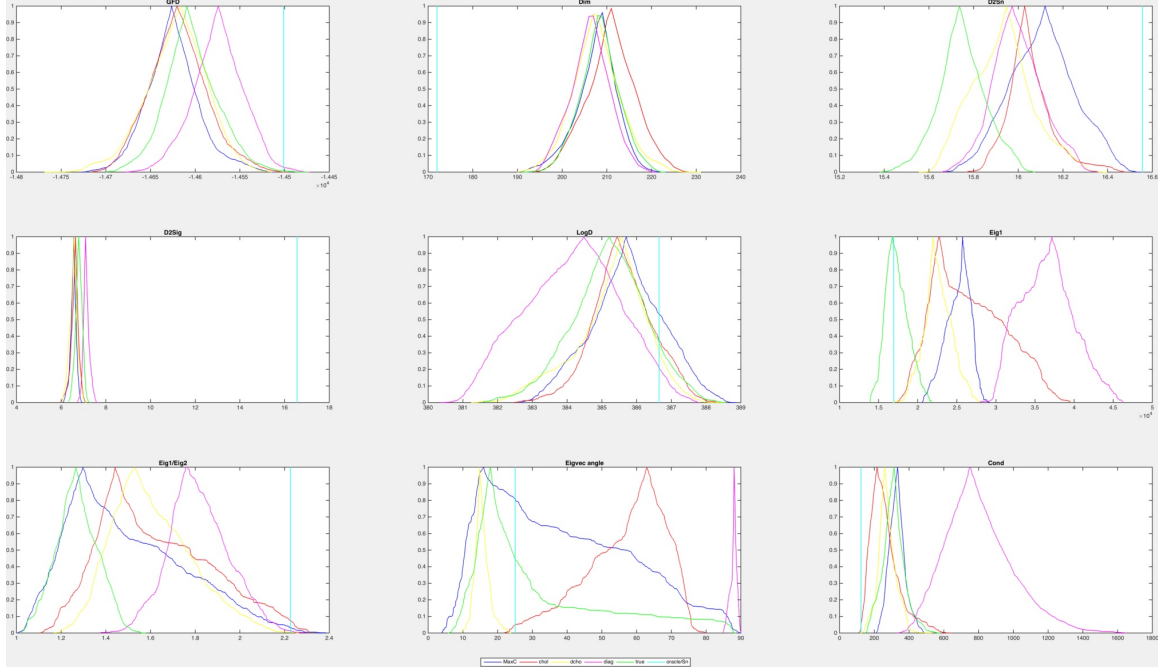


Figure 3.9: Confidence curve plots for RJMCMC with $p = 50$, $n = 50$, $\max C = 5$, $\text{burnin} = 100000$. The estimated covariates are less sparse than A , but the FM-distance to Σ are much shorter than from S_n to Σ .

CHAPTER 4

Phylogenetically dependent gene expressions study

4.1 Introduction

RNA is a critical intermediate between the genetic information encoded in the DNA of a gene and the phenotypes shaped by that gene. Changes in the levels or types of RNA molecules produced by a gene have profound biological effects. For example, changes in the RNA expression patterns of key developmental genes - the Hox, Pax6, and the Wnt family genes - are thought to have been events during the evolution of morphological diversity of animals. Similarly, mis-expression of some gene's RNA can have profound effects, such as promoting tumor formation and cancer.

A gene's RNA does not act alone; it is part of the RNA transcriptome. The transcriptome is comprised of RNA molecules transcribed from the genome: some code for proteins, others have structural, catalytic, or regulatory functions. High-throughput RNA sequencing technologies (RNA-seq) such as the Illumina HiSeq 2000 can now sample the transcriptome at unprecedented depth. This and related technologies have been primarily used for expression profiling, that is, identifying genes that change RNA expression levels in response to a treatment. RNA-seq quantitatively measures gene expression as counts, and typically involves isolating a subject's mRNA, converting to cDNA, and sequencing. Sequencing reads are then computationally mapped to loci of interest (e.g. genes or exons), and the number of reads associated with each locus is stored in a p -loci by n -individuals matrix. Matrices built from individuals representing phenotypically different populations may then be compared in order to correlate differences in gene expression with phenotype. A number of statistical methods exist for analyzing these data (Anders and Huber, 2010; McCarthy et al., 2012; Robinson and Smyth, 2007; ?)

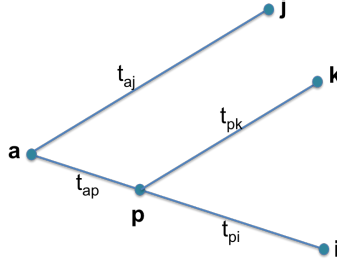


Figure 4.1: A simple three species phylogeny with known branch lengths $t_{aj}, t_{ap}, t_{pk}, t_{pi}$.

A phylogeny describes the relatedness between species. When comparing related species, it is important to take account of their phylogenetic dependency. For two species that are less related, it is not surprising to see the homologs expressed more differently compared to the more related species. For the purpose of detecting differentially expressed genes and change points along the phylogeny, it is crucial to factor in the relatedness of species.

In this chapter, we will provide a brief review of stochastic models used for related gene expression levels. The rest of the chapter is arranged as follows: Section 4.2 introduces the stochastic models have been used for describing the evolutionary process of continuous traits. In Section 4.3 we present some simulation results. Finally, in Section 4.4 we give a discussion.

4.2 Stochastic models for phylogenetically dependent gene expressions

How the expression level of a gene evolve over time is a stochastic event. It is natural to consider known stochastic processes for describing its evolutionary process. In the past decades, the Brownian motion (BM) model, the Ornstein-Uhlenbeck (OU) model, and the Lévy model has been used for capturing the evolutionary processes of a continuous trait, especially gene expressions, along a given phylogeny. A collection of stochastic processes of the same type (OU, BM, or general Lévy) has been used to model the evolutionary history along each of the tree branch. We will first illustrate the three models with the simply phylogeny in Figure 4.1.

In this simple example, we consider three leaf nodes (species i, j, k), an internal node p , and a root node a . The branch length between two nodes are denoted by t_{\cdot} , with the two

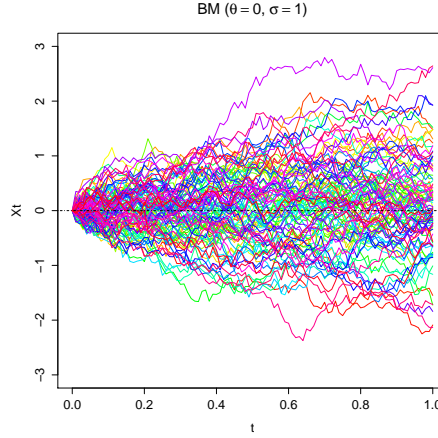


Figure 4.2: Multiple trace plots (n=100) of the same BM process. It does not have a stationary state.

node labels in the subscript. Since the observations are only available for the leaf nodes, an important first task is accessing the marginal joint likelihood of gene expression levels at the the leaves taking account of the phylogeny structure.

4.2.1 Brownian motion (BM)

A BM process can be defined by the following stochastic differential equation:

$$dX_t = \sigma dW_t, X_0 = \theta \quad (4.1)$$

The BM process describes the random motion of particles suspended in a fluid resulting for their collision with the quick atoms or molecules in the gas or liquid. Mathematically, it can be formulated as Eq 4.1. With the same initial state, a BM particle moves randomly to either direction (up & down) in Figure 4.2. The trace plot of 100 particles of the same BM process shows that, as time goes by, these particles are further apart.

The mathematical model of BM has numerous real-world applications, especially in finance when model stock market fluctuations. Its setup for the simple phylogeny (Figure 4.1) is shown in Eq 4.2.

$$\begin{aligned}
x_j|x_a &\sim N(x_a, \sigma_j^2 t_{aj}), & x_p|x_a &\sim N(x_a, \sigma_p^2 t_{ap}), \\
x_k|x_p &\sim N(x_p, \sigma_k^2 t_{pk}), & x_i|x_p &\sim N(x_p, \sigma_i^2 t_{pi}).
\end{aligned}
\tag{4.2}$$

Conditioning on its latest ancestor's expression level x_a , the measurement at species j , x_j , follows a normal distribution, centered at x_a , with variance proportional to the branch length from species a to j . Similarly for the other species. This simple setup leads to a normal joint likelihood of x_i, x_j, x_k . Since we assume BM model, conditioning on the latest common ancestor, the expression levels for those two species are independent. Namely, conditioning on x_a , $x_j \perp\!\!\!\perp x_k, x_j \perp\!\!\!\perp x_i$; conditioning on x_p , $x_k \perp\!\!\!\perp x_i$. Often time, we assume that $\sigma_j = \sigma_p = \sigma_k = \sigma_i$. The computation for the joint likelihood is therefore straight forward and the parameters can be estimated via the maximum likelihood method.

The BM model (Eq 4.2) was in the earlier microarray studies (e.g. (Gu and Gu, 2002; Bedford and Hartl, 2009)). One of the advantages of microarray data is that all the microarray plates are design to have same sets of transcriptomes of interest, and one can assume independence between the plates. In other words, the number of replicates for each species/transcriptome are the same. Independent and identical (iid) samples are assumed to be available. We extended the basic BM model for unbalanced sequencing data by introducing a within species sampling variance.

4.2.2 Ornstein-Uhlenbeck (OU)

Because of its simplicity, the BM model involves relatively easy computation. However, the divergent trace plot in Figure 4.2 contradicts to biology conservation. A different stochastic process, OU, that includes a directional drift, was then brought to the game. An OU process can be defined as the following:

$$dX_t = \alpha(\theta - X_t)dt + \sigma dW_t \tag{4.3}$$

The OU process describes the velocity of a massive Brownian particle under the influence of friction. Figure 4.3 shows the trace plot for 100 particles that follow the same process and started at various initial states. Regardless of the starting point, as time goes by, each

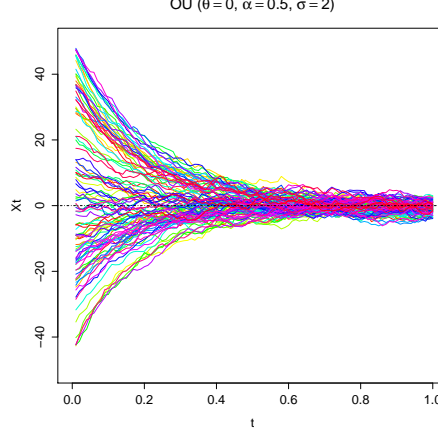


Figure 4.3: Multiple trace plots (n=100) of the same OU process.

particle is pulled closer and closer to its long time mean, at $\theta = 0$ in the example. How soon the curves started far away from θ get to close to it depends on the drift parameter α ; how wide the band for the later part of the graph depends on the standard deviation σ .

Conditioning on its ancestor, the expression at a node follows a normal distribution. Given the information of its latest ancestor, taking x_i as an example,

$$\begin{aligned} E(x_i) &= E(x_p)e^{-\alpha_i t_{pi}} + \theta_i(1 - e^{-\alpha_i t_{pi}}), \\ \text{Var}(x_i) &= \frac{\sigma_i^2}{2\alpha_i}(1 - e^{-2\alpha_i t_{pi}}) + \text{Var}(x_p)e^{-2\alpha_i t_{pi}}. \end{aligned} \quad (4.4)$$

The covariance between two leaf nodes is

$$\text{Cov}(x_i, x_j) = \text{Var}(x_a)e^{-(\alpha_p t_{ap} + \alpha_i t_{pi} + \alpha_j t_{aj})}. \quad (4.5)$$

The joint likelihood of x_i, x_j, x_k again follows a normal distribution, and the computation is relatively easy as under BM model. The OU model includes drifts towards to a particular state. The stationary distribution of x_i is $N\left(\theta_i, \frac{\sigma_i}{2\alpha_i}\right)$. The OU model is therefore suitable for conservation.

For multiple samples per species, there is the mean model, that takes the sample mean at the leaf nodes (e.g. (Brawand et al., 2011; Gu et al., 2013)). To incorporate sampling variation, Rohlf, *et al.* proposed the variance model (?). The variance model includes within species variation, which is similar to our extended BM model.

4.2.3 Lévy processes

Both BM and OU processes belong to a class called Lévy processes, which can be decomposed into three processes as stated in Lévy-Khinchine Representation Theorem (Gardiner, 1985).

Theorem 4.1 (Lévy-Khinchine Representation Theorem). *All Lévy processes have characteristic functions of the form*

$$\phi(k; t) = \exp \left\{ t(\alpha ik - \sigma^2 k^2 / 2 + \int (e^{ikj} - 1 - ikj \mathbb{I}_{|j| < 1}) \nu(dj) \right\}$$

Any Lévy process can be decompose into

1. *A constant directional drift (or trend) with rate a*
2. *A Brownian motion with rate σ^2*
3. *A pure-jump process that draws jumps from the measure $\nu(\cdot)$*

The BM and the OU processes only concern the second and the first two parts of the decomposition, respectively. They do not include a pure-jump process component, hence, it does not allow abrupt changes in the processes. In reality, events like massive gene duplication can happen. It does not occur in a continuous manner and causes a jump in the expression level. If a more comprehensive process, that includes jumps, is considered, the joint likelihood of x_i, x_k, x_i is no longer normal. In fact, many Lévy processes do not have a closed distribution form. Maximum likelihood estimation is not applicable. To estimate the process parameters, a Markov chain Monte Carlo (MCMC) procedure is often needed. Landis, *et al.* (Landis et al., 2013) proposed the framework:

$$\begin{aligned} p(\Theta, \mathbf{J} | D) &\propto L(D | \Theta, \mathbf{J}) p(\mathbf{J} | \Theta) p(\Theta), \\ p(\mathbf{J} | \Theta) &= \prod_i P(J_{t_i}^{(i)} = j^{(i)} | J_0 = 0, \Theta), \\ p(\Theta | D) &= \int p(\Theta, \mathbf{J} | D) d\mathbf{J}. \end{aligned} \tag{4.6}$$

where D denotes the expression data, \mathbf{J} are the sizes of pure jump, Θ includes the jump process parameters, i indicates each branch with length $t_{aj}, t_{ap}, t_{pk}, t_{pi}$.

The jump sizes, \mathbf{J} , can be different along each branch, hence the number of parameters grows quickly when a new species is added to the phylogeny. In addition, MCMC process can take a long time to converge when the parameter space is large. The sampling process tends to be computationally intensive. In (Landis et al., 2013), the authors considered three types of Lévy processes that are combinations of a pure jump and BM.

4.2.4 Parametric bootstrap

Once a stochastic model is selected for describing the gene expression level evolutionary history, one can proceed to check the Gaussianity of the data, identify differentially expressed genes and/or detect break points along the phylogeny. In general, for hypothesis testing, we propose to use the following parametric bootstrap approach instead of a chi-square test:

1. Obtain the MLEs under $H_0 \& H_1$, $\Theta_0 \& \Theta_1$, with observed data using the quasi-Newton method.
2. Compute the LRT statistics LRT_{obs} .
3. Simulate 1000 synthetic data using Θ_0 and compute their LRT statistics LRT_i .
4. Derive thresholds using both the bootstrapped data and χ_1^2 .

Set $LRT_{boot} = \text{quantile}(LRT_i, .95)$, $LRT_{\chi_1^2} = qchisq(.95, df = 1)$. Reject H_0 if $LRT_{obs} >$ thresholds.

In the past, the chi-square threshold, $LRT_{\chi_1^2}$, has been used. However, the likelihood ratio statistics does not follow a chi-square distribution due to the non-independence between the species. The $LRT_{\chi_1^2}$ cutoff tends to be too conservative. The parametric bootstrap threshold, LRT_{boot} , is more appropriate.

4.3 Implementation

To illustrate some of the methods, we applied them to a nine mammalian species dataset (chimp, bonobo, human, gorilla, orangutan, macaque, mouse, platypus, and chicken). The phylogeny was built with the RNA-seq data of a list of conservative homologs. Two out-group species, fish and frog, were added to enhance the quality of the tree (Figure 4.4). The branches shown are proportional to the evolutionary distances between nodes on the phylogeny.

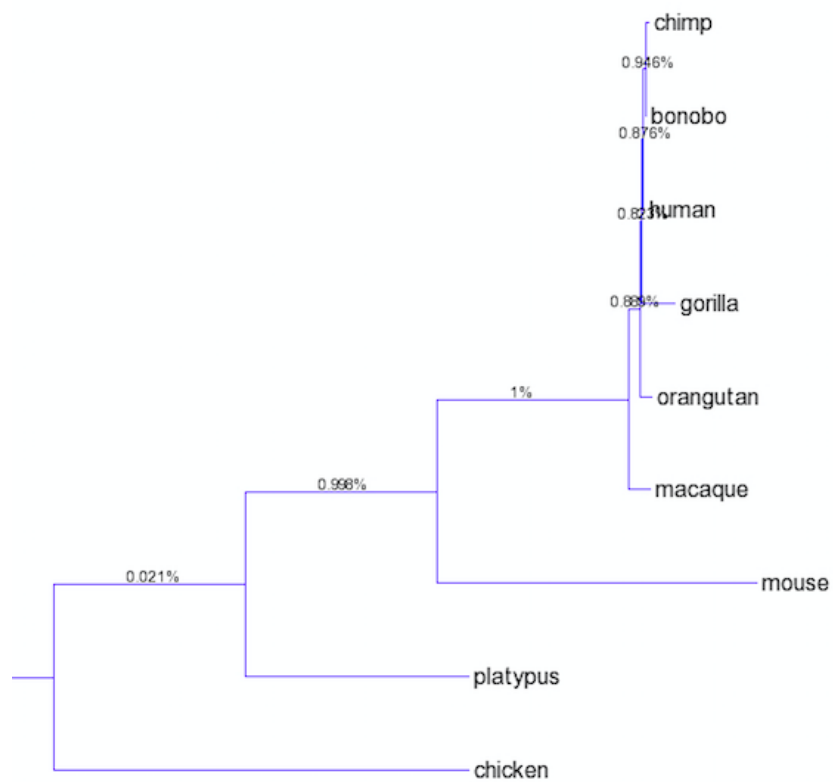


Figure 4.4: A mammalian species tree constructed with RNA-seq data of conservative homologs. Short branches between the primates indicate close relatedness comparing to mouse, platypus, and chicken.

4.3.1 Multi-species multi-tissues Mammalian data

The expression level of 5320 genes were collected for each species from six different tissues. At least two samples were obtained per species. This list of genes does not include the homologs used to build the phylogeny. The complete expression data are visualized via a heat map in Figure 4.5. The cold color (blue) indicates high level of gene expression while the hot color (red) denotes low level. Each row is an individual sample from a tissue. The color bar on the left indicates the tissue types: brain (purple), cerebellum (cyan), heart (orange), kidney (green), liver (burgundy), and testies (navy). The genes have been reordered to enhance the clustering of samples based on Ward distance between samples without considering the phylogeny structure. The color blocks in the tissue color bar shows that the gene expression levels differ more from tissue to tissue. This is not surprising since many genes tissue-specific and they are largely conserved (Merkin et al., 2012). The amount that a gene expressed can differ a lot even within the same individual.

Within a tissue, taking brain tissue as an example (Figure 4.6), the Direct Ward distance clustering shows good separation between the primates and the others. Since the primates are connected with short branches, meaning they are closely related. Comparing to the less related mouse, platypus, and chicken, it is less easy to distinguish among the primates. The dendrogram based on Ward distance is relatively consistent with our constructed phylogeny, where mouse, platypus and chicken are separated from the primates.

Figure 4.7 is a simulated multiple dimensional scaling (MDS) result for the leaf node species. A break point was simulated to occur on the chimp branch (yellow). The top two panels are direct comparisons; the bottom two have been adjusted for the phylogeny. The two columns from left to right are for single gene and multiple genes, respectively. Only with adjustment for the phylogenetic dependency, the break point on the chimp branch was identified. When multiple related genes were used, the separation between chimp and the rest of the group was enhanced. It is important to model the evolutionary processes with stochastic models. In addition, we can borrow power across related genes for detecting break points along the evolutionary history.

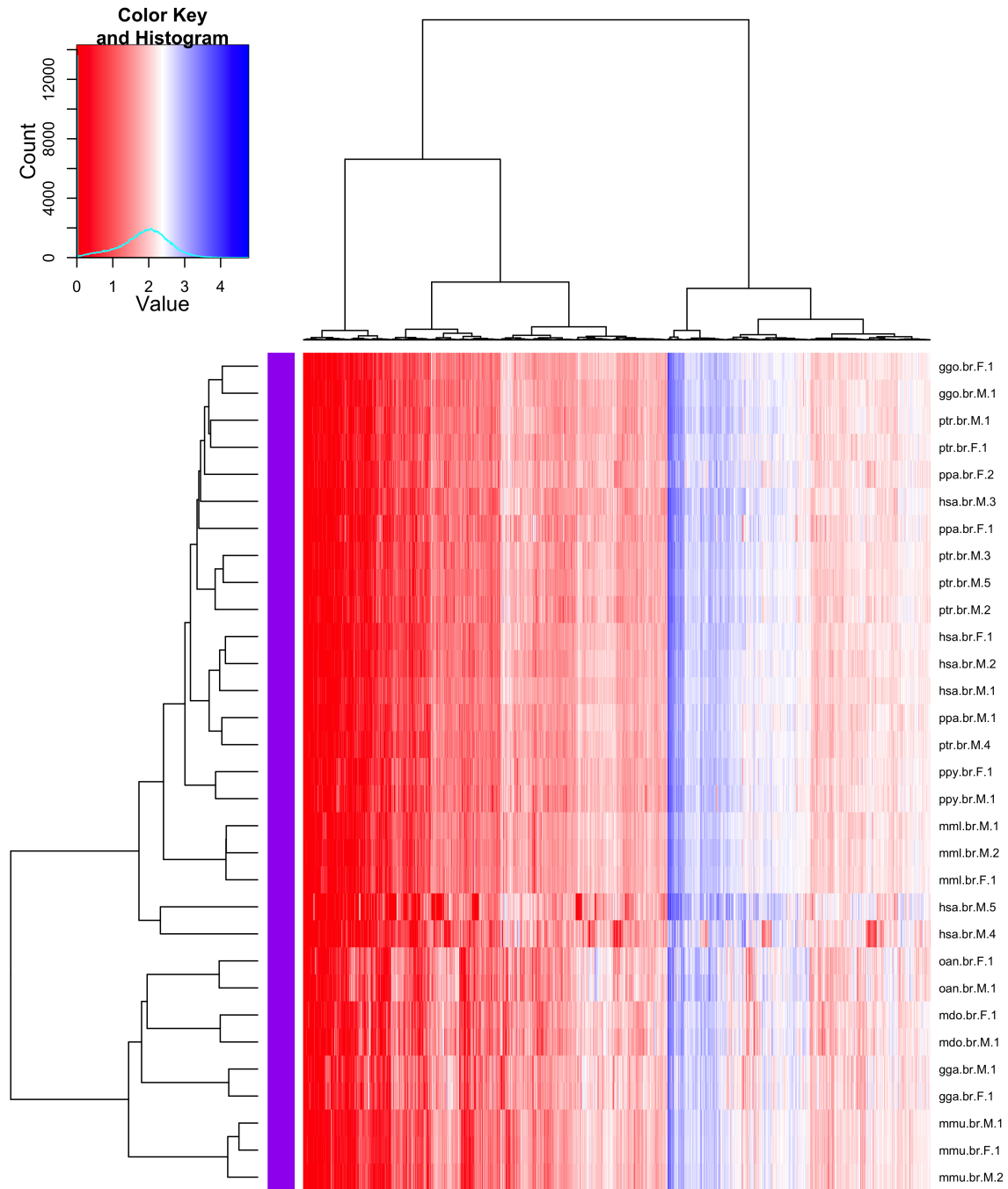


Figure 4.6: The heat map of the expression data for nine mammalian species from brain tissues. The relatedness of species determines affects the gene expression level comparison.

4.3.2 BM vs OU

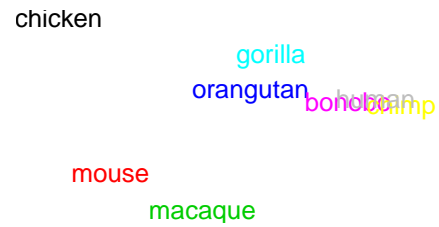
Because of the tissue-specific effect, when applying one of the stochastic models discussed in Section 4.2, the gene expressions from different tissue type should be modeled individually.

Single gene, no adjustment



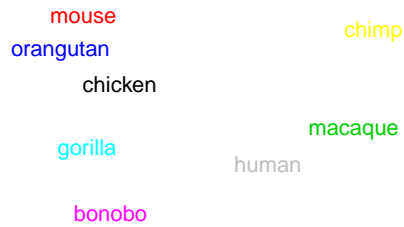
mouse bonobo macaque orangutan human chimp chicken

Multiple genes, no adjustment



chicken gorilla orangutan bonobo human mouse macaque

Single gene, adjusted



mouse orangutan chicken gorilla bonobo human chimp

Multiple genes, adjusted



chicken mouse macaque orangutan gorilla chimp bonobo human

Figure 4.7: MDS plot for simulation with a break point on the chimp branch. Top two panels ignored the phylogeny structure; the bottom two incorporated the phylogenetic relatedness using the Brownian motion model. The panels on the left and right are for single gene and multiple related genes. Top two panels failed to separate chimp from others. When multiple genes were used, the separation in the bottom right is more clear than when single gene was used.

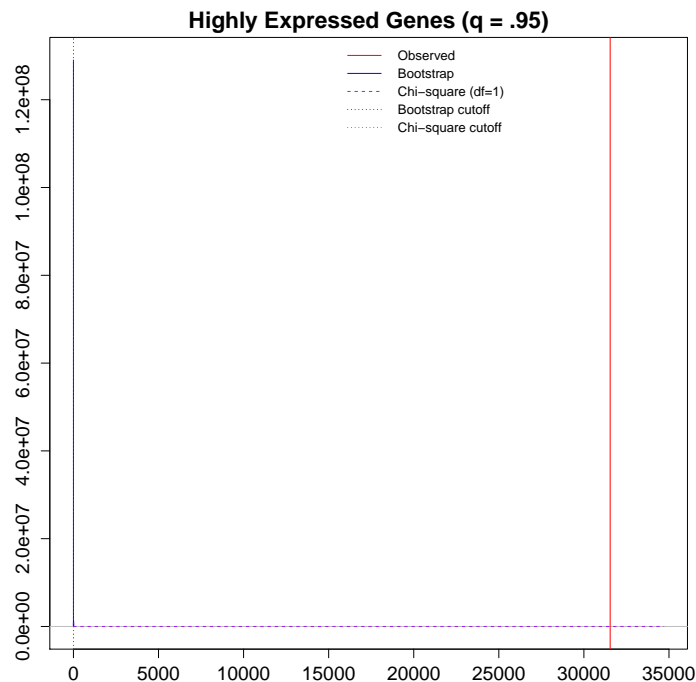


Figure 4.8: Test for BM model based on the top 10 highly expressed genes. There is a strong evidence against that the underlying model is BM.

To overcome the tissue-specific complexity, Merkin, *et al.* suggested to analyze alternative splicing instead (Merkin et al., 2012). Since our focus here is stochastic modeling for continuous traits, we will restrict ourselves to gene expression data for now. Among the 5320 homologs in the dataset, most are conservative, i.e. relatively consistent across species, but at different levels. Instead looking at all 5320 genes at once, we suggest to select a subset of related genes, such as a pathway or highly expressed genes, to apply the stochastic models.

Here we use the top 10 highly expressed genes in the data from brain tissue and test which of BM and OU models is more appropriate (Figure 4.8). The observed likelihood statistic is indicated as a red vertical line, the chi-square distribution and its threshold are in purple, the bootstrapped pseudo likelihood and the corresponding threshold are in blue. The observed value lies on the far right of both thresholds. It presents strong evidence against the underlying model being BM.

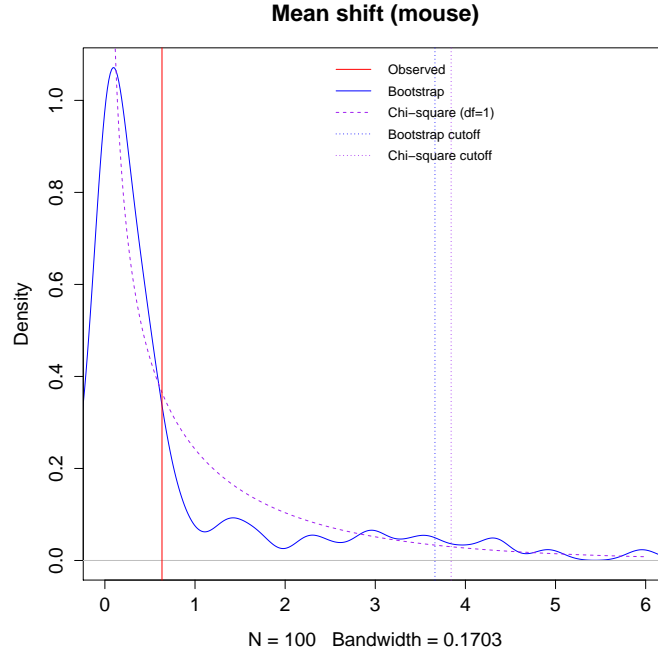


Figure 4.9: Test for single branch mean shift on the mouse branch based on the top 10 highly expressed genes.

4.3.3 Mean shift test

Assuming the OU model, we proceed to check mean shifts on the mouse branch as an example. Similar to before, the observed likelihood ratio statistic is indicated with the red vertical line; the chi-square distribution and its 95th percentile are in purple; the bootstrapped pseudo likelihood and the corresponding 95th percentile are in blue. The chi-square threshold is more liberal in this case. Since the observed statistic lies on the left hand side of the thresholds, we do not reject the null hypothesis that there is no mean shift along the mouse branch.

4.4 Discussion

Despite progress in the development of methods for comparing gene expression levels between two or more transcriptomes using RNA-seq data, a key problem has been ignored: many current methods assume that the samples investigated are genetically independent.

This issue becomes problematic when samples are drawn from a related family group (pedigree), different ethnic groups, or among phylogenetically related species.

The earliest stochastic models used in phylogenetic analysis for continuous traits are BM models. We extended a microarray based BM method to RNA-seq expression data by introducing an additional variance component that captures sampling variation. The BM models do not have a stationary distribution, hence, they are not suitable for modeling conservation. For this reason, the OU models were utilized for phylogenetic analysis in the last decade. With a directional drift, the OU model pulls the particle towards its long term mean. It, therefore, is suitable for conservation. Most of the OU model applications to phylogenies are mean models. The sample means are taken for each species. Few studies have considered the OU variance model, which incorporate sampling variance within species. The variance model is similar to our BM extension model. The OU variance model and our BM extension models capture the sampling variance. However, for the phylogenetic data where very few replicates are available, the gain of variance model might not be significant.

By testing the existence of the drift parameter, one can distinguish between OU and BM models. Both BM and OU processes are special cases of Levy processes, which can be decomposed to a BM process, a constant directional drift, and a pure-jump process. Change of expression level does not have to be a continuous event. For instance, massive gene duplication can cause abrupt changes in the expression level. To allow such events, a jump process is needed. The big drawback of jump processes, in general, is that the joint likelihood for the leaf nodes is no longer Gaussian. Maximum likelihood method is not suitable for estimating the process parameters. Instead, the parameters need to be estimated through Markov chain Monte Carlo (MCMC) methods. With jump processes, each branch allows a different jump size. The number of parameters quickly increases as the phylogeny grows. Likely because of this issue, so far only one study has proposed jump processes (Landis et al., 2013), and the Lévy processes discussed are a combination of BM and a pure-jump. In that study, the directional drift was assumed to be zero. Therefore, those jump models are also not suitable for conservation. Much work remains to improve the MCMC methods and to make general Lévy processes feasible for gene expression modeling.

Currently, the most popular stochastic models in the area of dependent gene expression analysis are the OU mean models. While they are suitable for conservation and are easy to compute, they do not allow for abrupt changes. The direct application of jump processes has proven to be computationally challenging. The study of these models has been very helpful for us to understand the current state of the research area and perhaps point us to step back and try to visualize/model the related expression data differently.

BIBLIOGRAPHY

- Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11:R106.
- Baldwin, E., Bhat, T., Liu, B., Pattabiraman, N., and Erickson, J. (1995). Structural basis of drug resistance for the v82a mutant of hiv-1 proteinase. *Nature Structural & Molecular Biology*, 2:244–249.
- Barbieri, M. M. and O’Hagan, A. (1996). A reversible jump mcmc sampler for bayesian analysis of arma time series. Technical report, Dipartimento di Statistica Probabilità e Statistiche Applicate Università ”La Sapienza” Roma, Italy and Department of Mathematics, University of Nottingham, Nottingham, UK.
- Barnard, G. A. (1995). Pivotal models and fiducial argument. *International Statistical Reviews*, 63:309–323.
- Bedford, T. and Hartl, D. L. (2009). Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences of the United States of America*, 106.4:1133–1138.
- Boutwell, C. L., Rolland, M. M., Herbeck, J. T., Mullins, J. I., and Allen, T. M. (2010). Viral evolution and escape during acute hiv-1 infection. *The Journal of Infectious Diseases*, 202(Suppl 2):S309–S314.
- Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pabo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478:343–478.
- Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). Efficient construction of reversible jump markov chain monte carlo proposal distributions. *Journal of the Royal Statistical Society*, 65.1:3–39.
- Cameron, D. W., Heath-Chiozzi, M., Danner, S., Cohen, C., Kravcik, S., Maurath, C., Sun, E., Henry, D., Rode, R., Potthoff, A., et al. (1998). Randomised placebo-controlled trial of ritonavir in advanced hiv-1 disease. the advanced hiv disease study group. *Lancet*, 351(9102):543–549.
- Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Wadsworth and Brooks.
- Chiang, A. (2001). A simple general method for constructing confidence intervals for functions of variance components. *Technometrics*, 43:356–357.
- Cisewski, J. and Hannig, J. (2012). Generalized fiducial inference for normal linear mixed models. *The Annals of Statistics*.
- Collins, P., Haire, L., Liu, Y., Russell, R., Walker, R., Skehel, J., Martin, S., Hay, A., and SJ, G. (2008). Crystal structure of oseltamivir-resistant influenza virus neuraminidase mutants. *Nature*, 453(7199):1258–61.

- Cuevas, J., Domingo-Calap, P., and Sanjuán, R. (2012). The fitness effects of synonymous mutations in dna and rna viruses. *Mol. Biol. Evol.*, 29(1):17–20. doi: 10.1093/molbev/msr179.
- Damgaard, C., Andersen, E., Knudsen, B., Gorodkin, J., and Kjems, J. (2004). Rnaintevolution in the 5' region of the hiv-1 genome. *J Mol. Biol.*, 336(2):269–79.
- Dawid, A. P. and Stone, M. (1982). The functional model basis of fiducial inference. *The Annals of Statistics*, 10:1054–1074.
- de Campos, C. P. and Benavoli, A. (2011). Inference with multinomial data: why to weaken the priors strength. In *International Joint Conference on Artificial Intelligence*.
- Dellaportas, P., Forster, J. H., and Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12.1:27–36.
- Dempster, A. (1966). New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics*, 37:355–374.
- Dempster, A. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B*, 48:365–377.
- E, L., Hannig, J., and Iyer, H. K. (2008). Fiducial intervals for variance components in an unbalanced two-component normal mixed linear model. *Journal of the American Statistical Association*, 103:854–865.
- E, L., Hannig, J., and Iyer, H. K. (2009). Fiducial generalized confidence interval for median lethal dose (ld50). *Preprint*.
- Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R., and Beerenwinkel, N. (2008). Viral population estimation using pyrosequencing. *PLoS Computational Biology*.
- Fishman, G. (2005). *A First Course in Monte Carlo*. Duxbury.
- Foll, M., Poh, Y., Renzette, N., Ferrer-Admetlla, A., Bank, C., Shim, H., Malaspinas, A., Ewing, G., Liu, P., Wegmann, D., Caffrey, D., Zeldovich, K., Bolon, D., Wang, J., Kowalik, T., Schiffer, C., Finberg, R., and Jensen, J. (2014). Influenza virus drug resistance: A time-sampled population genetics perspective. *PLOS Genetics*.
- Förstner, W. and Moonen, B. (1999). A metric for covariance matrices. *Quo Vadis Geodesia*, pages 113–128.
- Fraser, D. A. S. (1961a). The fiducial method and invariance. *Biometrika*, 48:261–280.
- Fraser, D. A. S. (1961b). On fiducial inference. *The Annals of Mathematical Statistics*, 32:661–676.
- Fraser, D. A. S. (1966). Structural probability and a generalization. *Biometrika*, 53:1–9.
- Fraser, D. A. S. (1968). *The Structure of Inference*. John Wiley & Sons Inc., New Yourk-London-Sydney.

- Gardiner, C. W. (1985). *Stochastic methods*. Springer-Verlag, Berlin–Heidelberg–New York–Tokyo.
- Geman, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4*.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6:721–741.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics. Springer-Verlag, New York.
- Glagovskiy, Y. S. (2006). Construction of fiducial confidence intervals for the mixture of cauchy and normal distributions. Master’s thesis, Colorado State University.
- Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82.4:711–732.
- Gu, J. and Gu, X. (2002). Induced gene expression in human brain after the split from chimpanzee. *Trends Genet.*, 19:63–65.
- Gu, X., Zou, Y., Huang, W., Shen, L., Arendsee, Z., and Su, Z. (2013). Phylogenomic distance methods for analyzing transcriptome evolution based on rna-seq data. *Genome Biology and Evolution*, 5.9:1746–1753.
- Hannig, J. (2009). On asymptotic properties of generalized fiducial inference for discretized data. Technical report, University of North Carolina, at Chapel Hill.
- Hannig, J. (2012). Generalized fiducial inference via discretization. *Statistica Sinica*.
- Hannig, J., E, L., Abdel-Karim, A., and Iyer, H. K. (2006a). Simultaneous fiducial generalized confidence intervals for ratios of means of lognormal distributions. *Austrian Journal of Statistics*, 35:261–269.
- Hannig, J., Iyer, H. K., and Patterson, P. (2006b). Fiducial generalized confidence intervals. *Journal of the American Statistical Association*, 101:254–269.
- Hannig, J., Iyer, H. K., and Wang, J. C. M. (2007). Fiducial approach to uncertainty assessment: account for error due to instrument resolution. *Metrologia*, 44:476–483.
- Hannig, J. and Lee, T. (2009a). Fiducial Inference and Generalizations. *Distribution*, (0707037):1–10.
- Hannig, J. and Lee, T. C. M. (2009b). Generalized fiducial inference for wavelet regression. *Biometrika*, 96:847–860.
- Hannig, J., Wang, J. C. M., and Iyer, H. K. (2003). Uncertainty calculation for the ratio of dependent measurements. *Metrologia*, 4:177–186.
- Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271.

- Huerta, G. and West, M. (1999). Priors and component structures in autoregressive times series models. *Journal of the Royal Statistical Society: Series B*, 61.4:881–889.
- Insua, D. R. and Müller, P. (1998). *Feedforward neural network for nonparametric regression*. Springer, New York.
- Iyer, H. K., Wang, C. M. J., and Mathew, T. (2004). Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, 99:1060–1071.
- Jabara, C., Jones, C., Roach, J., Anderson, J., and Swanstrom, R. (2011). Accurate sampling and deep sequencing of the hiv-1 protease gene using a primer id. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50):20166–20171.
- Jeffereys, H. (1940). Note on the behrens-fisher formula. *Annals of Eugenics*, 10:48–51.
- Knies, J., Dang, K., Vision, T., Hoffman, N., Swanstrom, R., and Burch, C. (2008). Compensatory evolution in rna secondary structures increases substitution rate variation among sites. *Mol. Biol. Evol.*, 25(8):1778–87. doi: 10.1093/molbev/msn130.
- Kuroda, M., Katano, H., Nakajima, N., Tobiume, M., Ainai, A., T, S, Hasegawa, H., Tashiro, M., Sasaki, Y., Arakawa, Y., Hata, S., Watanabe, M., and Sata, T. (2010). Characterization of quasispecies of pandemic 2009 influenza a virus (a/h1n1/2009) by de novo sequencing using a next-generation dna sequencer. *PLoS ONE*, 5(4):e10256.
- Landis, M. J., Schraiber, J. G., and Liang, M. (2013). Phylogenetic analysis using lévy processes: finding jumps in the evolution of continuous traits. *Systematic Biology*, 62.2:193–204.
- Leitner, T., Halapi, E., Scarlatt, G., Rossi, P., Albert, J., Fenyö, E., and Uhlen, M. (1993). Analysis of heterogeneous viral populations by direct dna sequencing. *Biotechniques*, 15(1):120–7.
- Lindley, D. V. (1958). Fiducial distributions and bayes’ theorem. *Journal of the Royal Statistical Society: Series B*, 20:102–107.
- McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor rna-seq expressions with respect to biological variation. *Nucleic Acids Research*, 40.10:4288–97.
- Merkin, J., Russel, C., Chen, P., and Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338.6114:1593–1599.
- Muers, M. (2011). Technology: Getting moore from dna sequencing. *Nature Reviews Genetics*, 12.9:586–587.
- Norström, M., Karsson, A., and Salemi, M. (2012). Towards a new paradigm linking virus molecular evolution and pathogenesis: experimental design and phylodynamic inference. *New Microbiol.*, 35(2):101–11.
- Patterson, P., Hannig, J., and Iyer, H. K. (2004). Fiducial generalized confidence intervals for proportion of conformance. Technical report, Colorado State University.

- Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004). The causes and consequences of hiv evolution. *Nature Reviews Genetics*, 5:52–61.
- Renzette, N., Caffrey, D., Zeldovich, K., Liu, P., Gallagher, G., Aiello, D., Porter, A., Kurt-Jones, E., Bolon, D., Poh, Y., Jensen, J., Schiffer, C., Kowalik, T., Finberg, R., and Wang, J. (2014). Evolution of the influenza a virus genome during development of oseltamivir resistance in vitro. *Journal of Virology*, 88:272–281.
- Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B*, 59.4:731–792.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14.5:465–658.
- Roberts, S. J., Holmes, C., and Denison, D. (2001). Minimum-entropy data partitioning using reversible jump markov chain monte carlo. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 23.8:909–914.
- Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.
- Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23.21:2881–2887.
- Salome, D. (1998). *Statistical inference via fiducial methods*. PhD thesis, University of Groningen.
- Simmonds, P. and Smith, D. (1999). Structural constraints on rna virus evolution. *Journal of Virology*, 73(7):5787–94.
- Sonderegger, D. and Hannig, J. (2012). Bernstein-von mises theorem for generalized fiducial distributions with application to free knot splines. *Preprint*.
- Stevens, W. L. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika*, 37:117–129.
- Troughton, P. and Godsill, S. J. (1998). A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves. *Proceedings of the 1988 IEEE International Conference*, 4.
- Tsui, K. W. and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, 84:602–607.
- Tsui, K. W. and Weerahandi, S. (1991). Corrections: "generalized p-values in significance testing of hypotheses in the presence of nuisance parameters" [journal of the american statistical association 84 (1989), no.406,602-607; mr1010352 (90g:62047)]. *Journal of the American Statistical Association*, 86:256.
- Walley, P. (1996). Inferences from multinomial data: Learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society*, Vol. 58, No. 1:3–57.
- Wandler, D. V. and Hannig, J. (2011). Fiducial inference on maximum mean of a multivariate normal distribution. *Journal of Multivariate Analysis*, 102:87–104.

- Wandler, D. V. and Hannig, J. (2012a). A fiducial approach to multiple comparisons. *Journal of Statistical Planning and Inference*, 142:878–895.
- Wandler, D. V. and Hannig, J. (2012b). Generalized fiducial confidence intervals for extremes. *Extremes*, 15:67–87.
- Wang, J. C. M., Hannig, J., and Iyer, H. K. (2012). Pivotal methods in the propagation of distributions. *Metrologia*, 49:382–389.
- Wang, J. C. M. and Iyer, H. K. (2005). Propagation of uncertainties in measurements using generalized inference. *Metrologia*, 42:145–153.
- Wang, J. C. M. and Iyer, H. K. (2006a). A generalized confidence interval for a measurand in the presence of type-a and type-b uncertainties. *Measurement*, 39:856–863.
- Wang, J. C. M. and Iyer, H. K. (2006b). Uncertainty of analysis of vector measurands using fiducial inference. *Metrologia*, 43:486–494.
- Watts, J., Dang, K., Gorelick, R., Leonard, C., Bess, J. J., Swanstrom, R., Burch, C., and Weeks, K. (2009). Architecture and secondary structure of an entire hiv-1 rna genome. *Nature*, 460(7256):711–6. doi:10.1038/nature08237.
- Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, 88:899–905.
- Weerahandi, S. (1994). Correction: "generalized confidence intervals" [journal of the american statistical association 88 (1993), no. 423, 899-905; mr1242940 (94e:62031)]. *Journal of the American Statistical Association*, 89:726.
- Weerahandi, S. (1995). *Exact statistical methods for data analysis*. Springer-Verlag, New York.
- Wilkinson, G. N. (1977). On resolving the controversy in statistical inference. *Journal of Royal Statistical Society: S*, 39:119–171.
- Wright, C., Morelli, M., Thébaud, G., Knowles, N., Herzyk, P., Paton, D., Haydon, D., and King, D. (2010). Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *J Virol*, 85(5):2266–75.
- Yeo, I.-K. and Johnson, R. A. (2001). A uniform strong law of large number of u-statistics with application to transforming to near symmetry. *Statistics & Probability Letters*, 51:63–69.