# STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL GENETIC DATA

XUAN LI

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

November 2018

# Abstract

This dissertation focuses on three types of high-dimensional genetic data: protein sequences, DNA methylation data, and microRNA expression data. The four major parts are presented in Chapters 2-5, respectively.

In Chapter 2, we develop a new clustering method for protein sequences. First, we reduce the dimensionality based on entropy. Second, the sequences are clustered using the Hamming distance vectors of chosen sites. We apply this new method to an influenza A H3N2 HA data set, which consists of 1960 viral sequences. Our method aggregates these sequences into 23 clusters. Based on the temporal evolution pattern of these clusters, we find that the dominant clusters change from time to time and are often different from the clusters housing vaccine strains.

In Chapter 3, we conduct systematic simulation studies and real data analysis to compare the performance of seven statistical tests for equal-variance hypothesis. Our results show that Brown-Forsythe test and trimmed-mean-based-Levene's test have better performance on DNA methylation data in comparison with other tests.

Detection of differential DNA methylation and differential variability have received a lot of attention in the literature. In Chapter 4, we derive the asymptotic distribution of a joint score test (AW), proposed by Anh and Wang (2013). Furthermore, we propose three improved joint score tests, namely iAW.Lev, iAW.BF, and iAW.TM. Systematic simulation studies show that at least one of the proposed tests performs better than the existing tests for data with outliers or from non-normal distributions. The real data analyses demonstrate that the three proposed tests have higher true validation rates than the existing tests.

Besides DNA methylation, microRNA regulation is another important epigenetic mechanism. In Chapter 5, we propose a novel model-based clustering method to detect differentially variable (DV) miRNAs. We impose biologically meaningful structures on covariance matrices for each cluster of miRNAs. Simulation studies show that the proposed method performs better than other model-based methods when miRNA expression levels are from a multivariate normal distribution. In real data analysis, the proposed method has a higher validation rate than other methods.

# Acknowledgements

First and foremost, I would like to express my utmost sincere appreciation to my supervisors, Professor Cindy Fu and Professor Steven Wang. Without their extensive knowledge, excellent guidance, and constant encouragement, this dissertation would not be possible. I have been extremely lucky to have two supervisors who cared so much about my work and responded to my questions and queries so promptly. I truly respect them for their brilliant insights and enthusiasm on research.

I would like to extend my grateful appreciation to Professor Yuehua Wu and Professor Huaiping Zhu as members of my supervisory committee. My appreciation also goes to all faculty members, staffs, and fellow graduate students in the Department of Mathematics and Statistics at York. I also wish to thank Dr. Weiliang Qiu, Dr. Qiong Li, Dr. Xiaoying Sun, and Dr. Nanwei Wang for their excellent suggestions on my dissertation.

I wish to thank my parents and my brothers from the bottom of my heart for their continuous support, encouragement, for always believing in me and decisions I

make. Their love is my driving force, providing me with inspiration. Also, a special thank you to my boyfriend for his patience, kindness, and wisdom. Finally, I offer my regards and thanks to all of those who supported me in any respect during the completion of the dissertation.

# Table of Contents

# List of Tables

# List of Figures

# 1 Introduction

The innovation of statistical procedures has always been driven by the desire to learn from emerging data. Genetic data can reflect inherited or acquired genetic features of an organism. The analysis of genetic data can advance our understanding of biological mechanisms in biological development, complex genetic disorders, and even the evolution of a species. However, the high dimensionality of genetic data has posed a significant challenge for scientists. With the rapid advance of sequencing technologies (e.g., next-generation sequencing), genetic data has exploded in both dimensionality and complexity. For instance, one complete hemagglutinin protein sequence in influenza contains 566 amino acid sites. Each site can have 20 possible types of amino acids and one gap. Therefore, a complete hemagglutinin protein sequence consists of $21^{566}$ possible states. The number of possible states is far more than the number of available sequences. During the past few years, there have been many publications on the analysis of high-dimensional genetic data (Boulesteix and Strimmer, 2006). Since many traditional statistical methods may

not be applicable in the analysis of high-dimensional data, statistically rigorous and biologically interpretable approaches are required to yield new scientific insights from these enormous, ever-growing genetic data resources. In this dissertation, we are concerned with three types of genetic data: influenza hemagglutinin protein sequences, human DNA methylation data, and human microRNA (miRNA) expression data. We introduce these three types of data and their related problems in the next three sections.

## 1.1 Influenza hemagglutinin protein sequences

Influenza virus is a negative-stranded, segmented, and enveloped RNA virus which incurs acute and infectious respiratory disease globally. Influenza epidemics act as a major cause of human mortality and morbidity. They occur in the winter months of each hemisphere every year, which is well-known as the influenza season (or flu season). In the influenza virus family, there are two main genera: A and B. Among them, influenza A has caused most of the flu epidemics in recent years. Based on their surface proteins, hemagglutinin (HA) and neuraminidase (NA), influenza A viruses can be further classified into 16 known HA and 9 known NA. The HA protein is regarded as the primary antigenic component in the circulating influenza virus. The occurrence of seasonal flu epidemics is highly influenced by the accumulated

small mutations in the HA protein (antigenic drift), which allow the virus to evade recognition of host immune systems and increase the lifetime susceptibility of the host.

Vaccination is the most effective way to prevent or mitigate the severity of seasonal flu. Each year, the formulation of seasonal flu vaccine is reviewed and sometimes updated because circulating influenza strains are continuously changing. However, high mutation rate of the seasonal influenza strains, especially on the HA protein, makes it difficult to select the most proper vaccine strains. Furthermore, some analyses of the HA gene sequence discovered that antigenic drift of the circulating strains from vaccine strains played an essential role in affecting the efficacy of the vaccine (Carrat and Flahault, 2007; Boni, 2008).

Recently, investigators begin to pay attention to flu viral swarms or clusters, which are viewed as major units driven by evolutionary forces, instead of only focusing on phylogenetic reconstruction of virus strains (Plotkin et al., 2002; Łuksza and Lässig, 2014). Nevertheless, in those studies which emphasize the evolutionary history of clades, the statistical approaches are not optimal and can be improved. For instance, Plotkin et al. (2002) made some subjective decisions in their methodology and consequently, the method is not fully automatic. Moreover, Łuksza and Lässig's method has many assumptions and also substantial computation complexity.

Therefore, a fast and effective clustering algorithm is required for flu viral sequences to describe the flu's temporal evolution pattern. The study of this pattern can help us predict future mutation hot spots in influenza protein sequences and design more effective vaccine strains.

## 1.2 Human DNA methylation data

DNA methylation is the most well-characterized epigenetic mechanism that regulates gene expression without changing genetic codes. In humans, methylation frequently occurs by adding a methyl group to the cytosine (C) nucleotide followed by a guanine (G) nucleotide, which is named as CpG site (Wahl et al., 2014). Aberrant methylation patterns and levels have been shown to be associated with many diseases such as cancer (Gopalakrishnan et al., 2008). Since DNA methylation is reversible, it is now considered as a potential therapeutic target in cancer treatment due to its ability to inhibit the expression of oncogenes, which can transform a normal cell into a tumor cell in certain circumstances.

Generally, the DNA methylation level is measured as the ratio of methylated to combined (methylated and unmethylated) levels. The definition is presented as follows. For a given methylation site, let $\tau_i$ denote the original methylation value of subject $i$, where $i = 1, ..., n$. Then $\tau_i$ is defined as the ratio of methylated to

combined intensity values of signals:

$$\tau_i = \frac{M_i}{U_i + M_i + e},$$

where $M_i$ and $U_i$ are the methylated and unmethylated intensity values of subject $i$, and $e$ is a small correction term to regularize probes of low total signal intensity (Teschendorff and Widschwendter, 2012).

Recent advances in next-generation sequencing and microarray technology allow us to measure genome-wide methylation levels at a high resolution (Bock, 2012). A series of Illumina Infinium Methylation platforms provide quantitative array-based methylation measurement at the single-CpG-site level. These platforms include the Illumina Infinium HumanMethylation27 BeadChip (HM27k), the Illumina Infinium HumanMethylation450 BeadChip (HM450k), and the Illumina Infinium MethylationEPIC BeadChip (EPIC). All three Illumina Infinium Methylation technologies are based on bisulfite-converted DNA. They investigate approximately 27k (HM27k), 450k (HM450k), and 850k (EPIC) CpG sites. The genomic regions targeted by the HM27k are proximal promoter region of RefSeq genes that are well-characterized in NCBI Reference Sequence Database and well-described cancer genes. Including 94 percent of the methylation sites of the HM27k, the HM450k covers more regions comprising of CpG islands and related regions, bodies of RefSeq genes and functional transcription regions, and more regulatory regions (Pidsley et al., 2016). Compared

with the HM450k, the EPIC quantifies DNA methylation levels at more distal regulatory regions.

Compared with the HM27k, the HM450k and the EPIC have more interrogated CpG sites, offering more resources for genome-wide association studies (GWAS). However, they also bring more challenges as a result of having two different types of chemical assays in one data, termed Infinium I and Infinium II. Infinium I assay uses two probes (Type I probes) per CpG locus (as HM27k) to generate methylated and unmethylated measurements. Infinium II assay uses a single probe (Type II probes) with two different colors to differentiate methylated (green) and unmethylated signals (red) (Wu et al., 2014; Shiah et al., 2017). The distributions of the methylation values derived from these two assays are significantly different. Type II probes are reported to have a reduced dynamic range and are generally less reproducible (Dedeurwaerder et al., 2011). Along with some common noises of microarray technologies, some biases introduced by Type II probes make the tasks of pre-processing DNA methylation data more challenging. Hence, normalization and quality control are crucial for statistical inference using DNA methylation data. In general, DNA methylation data should be pre-processed and evaluated by the following steps:

(1) Normalization including background correction, within-array normalization, and between-array normalization;

(2) Removal of the CpG sites with low quality, including CpG sites with high detection $p$-value, high missing proportion, and CpG sites residing near SNP, etc;

(3) Principal component analysis (PCA) for detection and adjustment of batch effects;

(4) Cell type estimation and adjustment or other adjustment.

One major goal in the analysis of methylation data is to identify disease-associated CpG sites. Many analyses in the past have focused on the difference of mean methylation levels between the diseased and control groups. Recently, some research suggests that methylation features detected based on variation discrepancy may also play a crucial role in unveiling the underlying mechanisms of complex diseases (Frank, 2010; Feinberg et al., 2010; Hansen et al., 2011; Jaffe et al., 2011; Teschendorff and Widschwendter, 2012). Many DNA methylation analyses show that differentially variable DNA methylation marks are biologically relevant to the disease of interest. However, these investigators make the inference relying on the information from the standard $F$ test or Bartlett's test (Bartlett, 1937). It is well-known that $F$ test or Bartlett's test is highly sensitive to the departure of the normality assumption and the presence of outliers. It has been reported that DNA methylation levels for different sites often follow various distributions and contain outliers (Ahn and Wang, 2013). Therefore,

we need robust equal-variance testing procedures, which can keep the nominal type I error and have reasonable power even if the normality assumption is violated.

Conover et al. (1981) compared 56 equal-variance testing procedures using simulation studies and found that the Brown-Forsythe test (Brown and Forsythe, 1974) was one of the best performers in terms of robustness and testing power. Phipson and Oshlack (2014) compared their proposed two equal-variance tests with $F$ test and Bartlett's test using simulated data generated from a Bayesian hierarchical model and evaluated the impact of outliers on the test statistics. However, systematic comparisons among these equal-variance tests are still needed to evaluate their performance when there are different distributions and outliers, featuring DNA methylation data.

Since discrepancies in both mean methylation levels and methylation variabilities can contribute to the identification of CpG sites relevant to the disease of interest, a more efficient approach is to test equal means and equal variances simultaneously. Some researchers have tried to construct this kind of joint tests. Littell and Folks (1971, 1973) compared four methods of combining independent tests of hypotheses and found that Fisher's method is the most efficient in terms of Bahadur efficiency. Perng and Littell (1976) suggested a joint test for two normal populations by testing equal mean and equal variance simultaneously. The joint test is proved to have

asymptotical optimality based on Bahadur efficiency. Zhang et al. (2012) constructed the exact distribution of the likelihood ratio test statistic of the joint test. Chen et al. (2014) employed a generalized exponential tilt model to derive semiparametric tests, which could evaluate the disparities in both means and variances between diseased and non-diseased groups. However, all these methods have parametric assumptions on data distributions, which may be violated in DNA methylation data. Ahn and Wang (2013) proposed a joint test to assess the discrepancies in means and variances simultaneously. This joint test is derived from a generalized linear model and relaxes some parametric restrictions on the data, which can be applied to DNA methylation data with different distributions. However, Ahn and Wang (2013) did not consider the fact that outliers are prevalent in DNA methylation data and the occurrence of mixture structures increases a lot in the HM450k and EPIC data. Therefore, we propose some robust joint tests to address this problem.

## 1.3 Human microRNA expression data

Besides DNA methylation, another important epigenetic mechanism is the regulation of microRNAs (miRNAs). Ribonucleic acid (RNA) is one of the three major biological macromolecules (DNA, RNA, and protein) that are essential for all known forms of life. RNA can be generally classified into two categories based on whether it

can be translated into protein (coding RNA) or not (non-coding RNA). The functionally important types of non-coding RNAs (ncRNAs) include transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small RNAs (e.g., microRNAs, siRNAs), and the long ncRNAs. The miRNAs are non-coding RNAs of about 22 nucleotides long, which can regulate gene expression through post-transcriptional repression or target mRNA degradation (He and Hannon, 2004; Hammond, 2015). It has been demonstrated that miRNAs play an important role in mammalian development, maintaining tissue homeostasis, cell cycle progression and proliferation, regulation of immune response, and aging of the brain (Hammond, 2015; Silva Rodrigues et al., 2018; Van den Hove et al., 2014). Aberrant miRNA expression patterns are found to be associated with a wider range of human diseases, such as cancer (Lu et al., 2005), metabolic diseases (Fernández-Hernando et al., 2013), Viral pathogenesis (Cullen, 2011).

The miRNA expression profiling becomes increasingly popular because miRNAs can significantly affect many biological processes and are the promising candidates for disease biomarkers. Three major approaches have been widely used for miRNA profiling: quantitative reverse transcription-polymerase chain reaction (qRT-PCR), hybridization-based methods (e.g., microarrays), and high-throughput sequencing (i.e., RNA-seq) (Pritchard et al., 2012). These approaches have different technical advantages, and hence can be used to achieve different research objectives. The

qRT-PCR is a well-established method and can be used to determine absolute quantification. The miRNA microarray is well established and can be easily adapted to existing microarray workflow. RNA sequencing has high accuracy and sensitivity to detect novel miRNAs. Ascribed to more and more comprehensive and in-depth studies of miRNAs, the public miRNA database miRBase (http://www.mirbase.org/) contains more than 38,000 miRNA entries with detailed information on sequences and physical structures (Kozomara and Griffiths-Jones, 2013). This information can help us advance the understanding of the mechanisms of gene regulation and develop novel effective miRNA-based diagnoses and therapies.

With the rapid advance of sequencing platforms, array-based technologies allow investigators to interrogate hundreds of thousands of miRNAs simultaneously in one experiment. Many analyses have been conducted to identify disease-associated miRNAs based on differential means (Bandrés et al., 2006; Resnick et al., 2009; Bandrés et al., 2006). Similar to DNA methylation data, differentially variable miRNAs may also be relevant to the disease of interest or the improvement of therapies (Mar et al., 2011). However, to the best of our knowledge, no studies have incorporated the information of variances because they have all focused on testing equal means in miRNA expression data.

Generally, there are two kinds of methods used in the analyses of gene expression

data: probe-wise approaches and model-based methods. Some common probe-wise approaches for differential variances are $F$ test or Bartlett's test (Bartlett, 1937), Brown-Forsythe test (Brown and Forsythe, 1974), and methods derived from these tests. The probe-wise approaches are flexible and easily implemented, but they ignore the correlations between gene marks and have multiple testing problems. The model-based methods can borrow information across probes and avoid multiple testing problems. Strbenac et al. (2016) used re-sampling methods and user-defined thresholds to identify gene marks based on differential means and differential variances. This method is partially subjective due to user-defined parameters. Bar et al. (2012, 2014) and Bar and Schifano (2018) considered differential means and differential variances in mixture models. However, these models characterize the distributions of the summary statistics (e.g., mean, variance, or difference of means), instead of the observed expression levels. In this dissertation, we propose a three-component multivariate normal mixture model to fit the expression levels of miRNAs in order to identify differentially variable miRNAs between two samples.

# 2 Clustering influenza hemagglutinin protein Sequences

In this chapter, we present a novel clustering method to aggregate the hemagglutinin (HA) protein sequences of flu viruses. This method has two steps: entropy-based dimensionality reduction and clustering. We apply this method to study the evolutionary properties of the HA component of influenza A H3N2 virus - a major cause of seasonal flu. We show that our new method could be used to uncover HA evolution patterns and evaluate recommended vaccine strains.

## 2.1 Methodology

### 2.1.1 Data acquisition

This study used 1960 sequences, of which 1947 sequences were directly downloaded from the Influenza Research Database (IRD), an online repository of influenza

sequences, based on the criteria listed in Table 2.1. The downloaded sequences of the H3 type HA genes were collected from locations around the globe between September 1998 and July 2012. Each of these sequences consists of 1698 nucleotides plus a stop codon (3 nucleotides).

Table 2.1: Search criteria of HA sequences in IRD.

| Option | Criteria |
|---|---|
| "Data to return" | protein |
| "Virus type" | A |
| "Sub type" | H3N2 |
| "Select segments" | HA |
| "Complete sequences" | Complete Segments Only |
| "Date range" | 1998 to 2012 |
| "Host" | Human |
| "Geographic grouping" | All |
| Advanced options | |
| "Month Range" | Sep 1998 to July 2012 |
| "Remove Duplicate Sequences" | Yes |

All other settings were kept the default or blank.

In order to explore the relationship between the collected flu sequences and recommended vaccine strains, we added recommended vaccine strains to the data set. The vaccine sequence information was obtained from the World Health Organization (http://www.who.int/influenza/vaccines/virus/recommendations/en/). Ta-

ble 2.2 lists the information and includes all vaccine strains used from September 1998 to July 2012. When we searched for vaccine strains in our downloaded data set based on strain names, three vaccine strains called "A/Brisbane/10/2007", "A/Perth/16/2009", and "A/Texas/50/2012", were already included in the data set. We acquired the remaining vaccine sequences separately and merged them into the data set manually, resulting in a total of 1960 sequences in our data set.

Table 2.2: Vaccine sequences in the data set.

| Stain Name | Number of sequences | Accession Number |
|---|---|---|
| A/Moscow/10/99 | 2 | AY531035, DQ487341 |
| A/Fujian/411/2002 | 2 | CY088483, CY112933 |
| A/California/7/2004 | 1 | CY114373 |
| A/Wisconsin/67/2005 | 4 | CY033646, CY163936 CY114381, EU103823 |
| A/Brisbane/10/2007 | 3 | CY035022, CY039087 EU199366 |
| A/Perth/16/2009 | 1 | GQ293081 |
| A/Victoria/361/2011 | 1 | KC306165 |
| A/Texas/50/2012 | 2 | KC892248, KC892952 |

The resulting 1960 sequences were then translated into the corresponding amino acids using MEGA software (Tamura et al., 2011). We translated the coding sequences into protein sequences instead of downloading the protein sequences directly from IRD to avoid the handling of the ambiguous amino acid sign "B". The transla-

tion resulted in 566 amino acids for each of the 1960 sequences. Next, we conducted multiple alignments on all 1960 protein sequences simultaneously using MUSCLE software (Edgar, 2004). All sequences were neatly aligned and some of them may have contained a few gaps. The data set can be regarded as 1960 observations (sequences) and 566 categorical variables (amino acid sites). Each site has 21 possible states, 20 types of amino acids and one gap. To better analyze the data set using statistical softwares (R and Matlab), we converted the alphabetical characters (representing amino acids) into numerical values.

Each of these 1960 sequences is related to a calendar year, country, and city of isolation, inferred from the strain name. For the 1947 sequences that were directly downloaded from the IRD, we can also obtain the date of isolation, which allows us to partition the data into different influenza seasons (October 1st through September 30th).

### 2.1.2 Entropy-based dimensionality reduction

A dataset $\mathbb{S}$ with $n$ records and $p$ columns is a sample set of the discrete random vector $A = \{a_1, \ldots, a_p\}$. For each component $a_j$, $1 \leq j \leq p$, $a_j$ takes a value from the state domain $\Psi_j$. For $j \neq j'$, $\Psi_j$ is conceptually different from $\Psi_{j'}$. There are a finite number of distinct categorical values in $domain(\Psi_j)$ and we denote the number

of distinct values as $|\Psi_j|$.

For each component $a_j$, let $p(a_j = k)$, $k \in \Psi_j$. The entropy of $a_j$ can be defined as (Cover and Thomas, 2012):

$$H(a_j) \quad = \quad -\sum_{k\in\Psi_j} p(a_j = k)\log p(a_j = k). \tag{2.1}$$

Since $H(a_j)$ is estimated using the sample set $\mathbb{S}$, we define the estimated entropy as $\hat{H}(a_j) = H(a_j|\mathbb{S})$, i.e.

$$\begin{aligned}\hat{H}(a_j) =& H(a_j|\mathbb{S}) \\ =& -\sum_{k\in\Psi_j} p(a_j = k|\mathbb{S})\log p(a_j = k|\mathbb{S}).\end{aligned} \tag{2.2}$$

In our context, the numbers of records and columns are $n = 1960$ and $p = 566$. Each component of the categorical vector $a_j$, $1 \leq j \leq p$ contains 21 categorical values, using integers 1-20 to represent the 20 types of amino acids and 21 to denote a gap. Hence $\Psi_j = (1, \ldots, 21)$ and $|\Psi_j| = 21$ for $j = 1, ..., 566$.

To avoid subjective decision about the number of clusters, we use Hamming distance vector (HD vector) algorithm (Zhang et al., 2006) to conduct clustering analysis. The most time-consuming part of the HD vector algorithm is to find cluster centers. The computational complexity is $O(\gamma^3)$, where $\gamma$ is defined as the number of unique sequences in the data set, which depends on $M$, the total number of possible positions. Thus, we reduce the dimensionality before clustering analysis. The main

idea behind our dimensionality reduction is to select a smaller number ($p_e$) of sites based on the variabilities of the sites in the data. This is a reasonable approach, as it is well known that parts of HA sequences are well-conserved (Staneková and Varečková, 2010). To identify the most variable (equivalently, least conserved) sites, we use the notion of entropy. Incorporating the known information to (2.2), we compute the entropy for each of the 566 sites as follows:

$$\hat{H}(a_j) \quad = \quad -\sum_{k=1}^{21} p(a_j = k) \log p(a_j = k). \tag{2.3}$$

Note that $\hat{H}(a_j)$ is always positive, and large entropy indicates great variability at a site. Sites with entropy equal to zero were removed, as there is no amino acid variability in those sites and hence, no useful information for clustering. Sites with entropy equal to 0.004377 (i.e. only one observation has a different state value from the other 1959 observations) were also removed. Finally, we sorted the remaining entropy in ascending order and used a Gaussian mixture model to cluster them (Everitt and Hand, 1981). The algorithm results in five classes of entropies. We identified the fifth class with the largest entropy as the selected sites to cluster all the sequences. This class contains 62 sites ($p_e = 62$), which allows us to reduce the dimension to $21^{62}$. That is, our data is now made up of 62 sites (variables) and 1960 sequences (observations).

### 2.1.3 Clustering categorical data

We use Hamming distance to evaluate the distance between any two sequences (Forney, 1966). For two sequences, $A = \{a_1, \ldots, a_{p_e}\}$ and $B = \{b_1, \ldots, b_{p_e}\}$ the Hamming distance is defined as

$$d_s(A, B) \;=\; \sum_{i=1}^{p_e} I(a_i \neq b_i), \tag{2.4}$$

where $I(E)$ is the indicator function that is equal to one if $E$ is true, and zero otherwise. The Hamming distance between any two HA sequences is the number of sites with different amino acids.

According to the Hamming distance vector (HD vector) algorithm (Zhang et al., 2006), we consider a general set-up where $p_e$ nominal categorical attributes are of interest and the $j$th attribute is categorized by $m_j$ ($m_j = |\Psi_j|$) levels. The categorical sample space, $\Omega$, is defined as the collection of all possible $p_e$-dimensional vectors of states. For us, $m_j = 21$ for each $j$, and $j = 1, \ldots, p_e$. Therefore, each sequence can be seen as a vector of length 62 ($p_e = 62$), and each element of the vector is a value taken from one of 21 ($m_j = 21$ for all $j$) possible categories.

Any given data set, which in our case can be represented as $A_1, \ldots, A_n$, gives a distribution of distances on the sample space $\Omega$ from a fixed reference position in $\Omega$. We denote this fixed reference position as $S = \{s_1, \ldots, s_{p_e}\}$. For a sample data set, we use $n$ to denote the sample size ($n = 1960$). Recall the definition of

Hamming distance given in (2.4), and note that it will take values in $0, 1, 2, \ldots, p_e$. The algorithm relies on HD vector, which is defined as a $(p_e + 1)$-element vector $H(S) = \{H_0(S), H_1(S), \ldots, H_{p_e}(S)\}$, where

$$H_q(S) = \sum_{j=1}^{n} 1(d_s(A_j, S) = q), \quad q = 0, \ldots, p_e.$$

Thus, $H_q(S)$ counts the number of all sequences of which the Hamming distance to the given reference position $S$ equal to exactly $q$ (Zhang et al., 2006).

In the categorical sample space $\Omega$, when all the data points have equal probability to occur at each position, the resulting HD vector is defined as $uniform$ HD vector (UHD), and denoted by $U(\Omega) = \{U_0(\Omega), U_1(\Omega), \ldots, U_{p_e}(\Omega)\}$. Given an uniformly distributed data $\{x_1, \ldots, x_n\}$, for a reference position $S \in \Omega$, the corresponding $uniform$ HD vector $U(S)$ has $p_e + 1$ elements. The $q$th element is:

$$U_q(S) = \sum_{i=1}^{n} d_s(x_i, S), \quad q = 0, \ldots, p_e,$$

which is the number of possible sequences with a distance $q$ to the position $S$ (Zhang et al., 2006). Note that the total number of possible positions in the categorical sample space is $M = \prod_{j=1}^{p_e} m_j$, where $m_j$ is the number of states for $j$th attribute.

Based on Theorem 3 of Zhang et al. (2006), the UHD vector has the following form:

$$
\begin{aligned}
U_0(\Omega) =& \frac{n}{M}, \\
U_1(\Omega) =& \frac{n}{M}\{(m_1 - 1) + (m_2 - 1) + \cdots + (m_{p_e} - 1)\}, \\
U_2(\Omega) =& \frac{n}{M}\sum_{i<j}^{p_e}(m_i - 1)(m_j - 1), \\
& \vdots \\
U_{p_e}(\Omega) =& \frac{n}{M}(m_i - 1)(m_j - 1)\ldots(m_{p_e} - 1).
\end{aligned}
$$

Note that the UHD vector does not depend on the position $S$.

The HD vector algorithm sequentially examines the existence of cluster patterns, and extracts the clusters. At each iteration, the algorithm detects only one cluster which is defined by a cluster center and a cluster radius. The cluster center is determined by the position with maximum modified chi-squared statistic based on HD vector and UHD vector. The cluster radius is defined by the first local minimum of the frequency distribution of the HD vector. The cluster will be deleted from the dataset before the next iteration. When there are no more significant clusters in the remaining data, the iteration stops and the algorithm outputs the number of clusters. Thus, the algorithm is fully automatic in finding both the clusters and the number of clusters.

After applying the HD vector algorithm to our data set of sequences, we want to evaluate the method of dimension reduction based on the distances between clusters.

Consider two clusters, $\mathcal{C}_1$ and $\mathcal{C}_2$. Each cluster is made of up of a number of sequences, say, $\mathcal{C}_1 = \{A_1, \ldots, A_{\kappa_1}\}$ and $\mathcal{C}_2 = \{B_1, \ldots, B_{\kappa_2}\}$. The mean Hamming distance is then

$$d_c(\mathcal{C}_1, \mathcal{C}_2) \;=\; \sum_{i,j} \frac{d_s(A_i, B_j)}{\kappa_1 \kappa_2},$$

if $\mathcal{C}_1$ and $\mathcal{C}_2$ are two different clusters. If $\mathcal{C}_1 = \mathcal{C}_2$, then we use

$$d_c(\mathcal{C}_1, \mathcal{C}_1) \;=\; \sum_{i<j} \frac{d_s(A_i, A_j)}{\kappa_1(\kappa_1 - 1)/2} \;=\; \sum_{i \neq j} \frac{d_s(A_i, A_j)}{\kappa_1(\kappa_1 - 1)}$$

This modification is due to the fact that when comparing the same cluster, all distances "along the diagonal" will always be zero.

## 2.2 Results

### 2.2.1 Cluster evaluation

After the dimensionality reduction, we identified 62 sites of high variability across the whole HA sequence. The HA protein is composed of two subunits - HA1 and HA2. The two subunits are linked by disulfide bond and form a protein complex to exert the full function (Knipe and Howley, 2007). Of the 62 sites, 52 lie within the HA1 domain. The remaining 10 sites lie within the HA2 domain.

The 1960 viral sequences were partitioned into 23 clusters. Figure 2.1 shows two dendrograms of the resulting clusters. The top dendrogram is based on the mean

Hamming distance calculated only for the sites of maximal entropy; whereas in the bottom dendrogram, the mean Hamming distance is calculated for all sites. The more variable the corresponding amino acid site, the higher the entropy. The most variable sites tend to be protein mutation hot spots based on the available sequences. Indeed, the dendrograms are largely consistent with regard to tree locations. Clusters 1 - 8 are grouped into a clade, while the remaining clusters, with the exception of 21, are grouped into another clade. Although subtle discrepancies in the specific clade location of some clusters exist, the evolution pattern of clades is generally consistent. We use mean Hamming distance (in amino acids) to measure genetic distance between two clusters of sequences. If the two clusters are close in mean Hamming distance, they are considered to be close in lineage evolution history. We regard clusters of sequences as evolutionary units and infer the genealogy of the clusters by ensemble of trees, those with small mean Hamming distances will be grouped into a clade (i.e. in the same trunk of phylogenic tree).

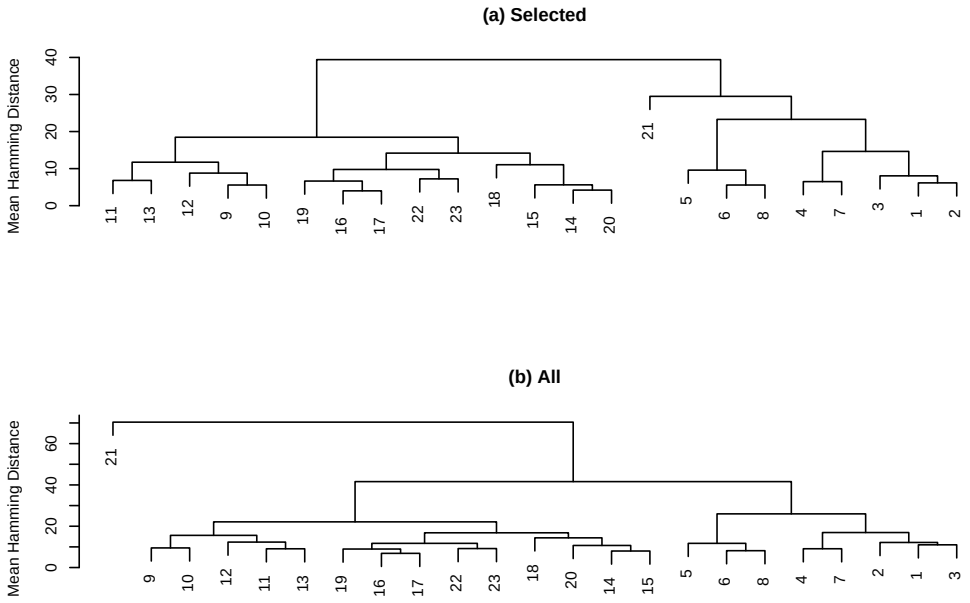


Figure 2.1: Dendrograms of clusters by mean Hamming distance. This plot is drawn using hierarchical cluster analysis with complete linkage. The top dendrogram is based on mean Hamming distance of the 62 sites with highest entropy, and the bottom one is based on mean Hamming distance of all 566 sites.

### 2.2.2 Temporal evolution of flu clusters

Figure 2.2 shows the number of sequences in each cluster sorted by the year of isolation. The clusters that house the vaccine strains are also indicated. The location order of vaccine strains is consistent with the calendar year according to their strain

name (and year closeby) e.g. "A/California/7/2004" and "A/Wisconsin/67/2005", "A/Victoria/361/2011" and "A/Texas/50/2012" are clustered together. This helps verify the validity of the clustering results.

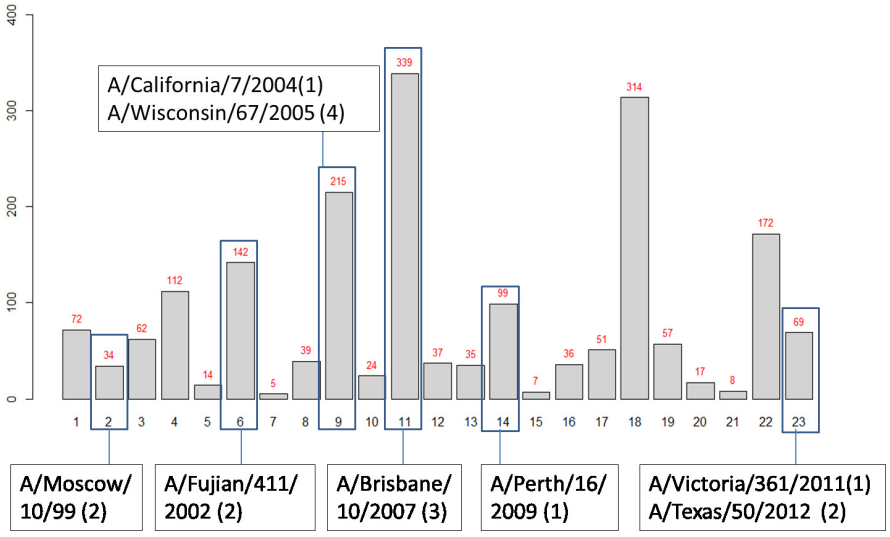


Figure 2.2: Histogram of cluster size and vaccine location. The clusters have been re-ordered in accordance to the sequence of the calendar year.

We can observe from Figure 2.2 that large clusters are generally surrounded by clusters of much smaller sizes. Thus, the dominant clusters can be identified over time. We can also observe that a higher number of small clusters are generated in recent years. This may be due to higher reporting rates, as rapid sequencing technologies have become increasingly available.

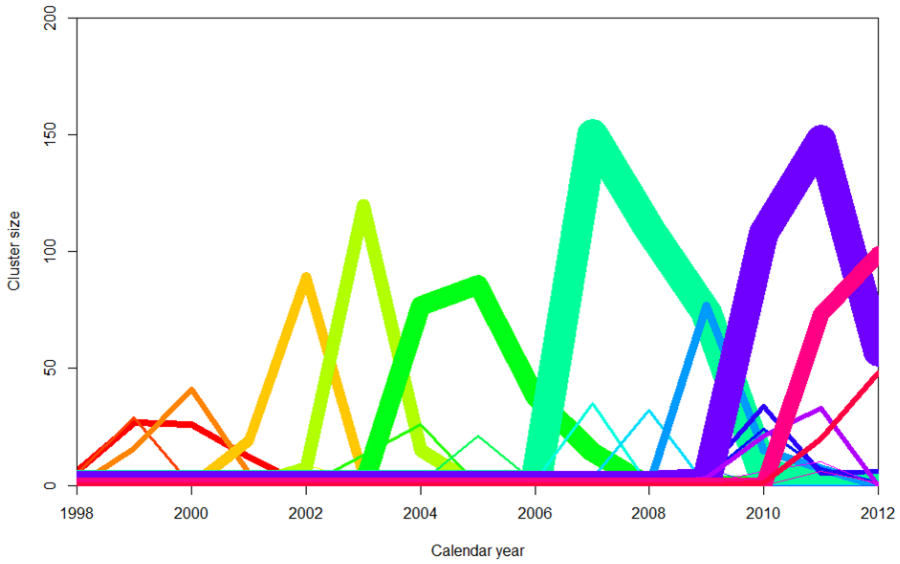


Figure 2.3: The number of HA protein sequences within each cluster versus the calendar year of isolation. Each cluster is indicated by a different colour, and the line width reflects the cluster size.

Each cluster houses strains that exist over one or more influenza seasons. In Figure 2.3, we plot the number of sequences in each cluster as a function of their isolation year. The thickness of the line indicates the size of each cluster. We observe that some clusters have a significantly longer lifespan than others, but no cluster spans more than seven years. It is also observed that clusters first increase and then decrease in size over their lifespan. Dominant clusters of viral sequences replace one another every 2-5 years. However, the occurrence of dominant clusters

in Figure 2.3 is not periodic within the time span. Once HA evolves away from a given region of the sequence space, it does not revisit that region at a later time. This agrees with previous studies of influenza evolution (Plotkin et al., 2002; Knipe and Howley, 2007).

### 2.2.3 Evaluation of recommended vaccines

In Table 2.3, we identified dominant clusters and the clusters containing vaccine strains (vaccine clusters) from 2000 to 2012. Ideally, the vaccine cluster can be the same as the dominant cluster. Considering the time lag that exists between the disease outbreak and time of isolation, the vaccine clusters should be as close to the dominant clusters as possible.

A common observation over all of the years shown indicates that the vaccine clusters and the dominant clusters are different. For example, from 2000-2004 the same vaccine strain "A/Moscow/10/99" was used, however, the dominant cluster changes each year in this time period, moving from a mean distance of 6.15 amino acids (aa) of the vaccine sequence to 8.37aa, and then 18.68aa in 2002-2004.

Table 2.3: Vaccines and Clusters by Year

| Season | Vaccine | Cluster | Dominant cluster | Vaccine cluster |
|---|---|---|---|---|
| 2000-2001 | A/Moscow/10/99 | **1** 3 4 5 | 1 | 2 |
| 2001-2002 | A/Moscow/10/99 | 1 **4** 5 6 | 4 | 2 |
| 2002-2003 | A/Moscow/10/99 | 4 5 **6** 7 8 | 6 | 2 |
| 2003-2004 | A/Moscow/10/99 | 4 **6** 8 9 | 6 | 2 |
| 2004-2005 | A/Fujian/411/2002 | (6) 8 **9** 10 | 9 | 6 |
| 2005-2006 | A/California/7/2004 | 1 (**9**) 10 | 9 | 9 |
| 2006-2007 | A/Wisconsin/67/2005 | (9) 10 **11** 12 13 | 11 | 9 |
| 2007-2008 | A/Wisconsin/67/2005 | (9) **11** 13 | 11 | 9 |
| 2008-2009 | A/Brisbane/10/2007 | (**11**) 14 15 16 17 19 20 21 | 11 | 11 |
| 2009-2010 | A/Brisbane/10/2007 | (11) 14 16 17 **18** 19 20 | 18 | 11 |
| 2010-2011 | A/Perth/16/2009 | 11 (14) 16 17 **18** 19 20 21 22 23 | 18 | 14 |
| 2011-2012 | A/Perth/16/2009 | (14) 16 17 18 21 **22** 23 | 22 | 14 |

From Table 2.3, we can also see that cluster extinction often coincides with the existence or introduction of a well-matched vaccine strain. In particular, the extinction of clusters 6 and 9 coincides with the introduction of vaccines housed in the same cluster. A similar observation can be made in cluster 11. Ultimate extinction of a cluster, however, is the result of a combination of various factors, including vaccine strain and competition between strains. An exploration of strain fitness is a course for future work.

## 2.3 Discussion

In this chapter, we have presented a new method for clustering protein sequences and have applied the method to clustering HA sequences of seasonal influenza A H3N2. The inclusion of vaccine sequences in the analysis allows us to present important relationships between the vaccine strains and the dominating flu clusters.

The traditional method for clustering genetic sequences is to use hierarchical agglomerative clustering methods to construct the phylogenetic tree of sequences based on the pairwise distance of the whole sequence (Plotkin et al., 2002). The number of resulting clusters are determined by subjective decisions. Our proposed method is parameter free and doesn't depend on subjective decisions about the number of clusters.

We demonstrate that it is not necessary to perform clustering methods on the entire HA genome. The 62 sites of the largest entropy can provide similar measures of Hamming distance. These 62 sites lie within the HA1 and HA2 regions of the HA genome. Our clustering methodology separates the HA data set into 23 clusters. Based on the analyses of these clusters, we find that dominant clusters replace one another every 2-5 years; the dominant clusters are often different from the clusters housing the vaccine strain, and the extinction of a dominant cluster often coincides with the existence or introduction of a well-matched vaccine.

Our results are highly consistent with previous studies of HA evolution (Plotkin et al., 2002; Nelson and Holmes, 2007; Łuksza and Lässig, 2014). Plotkin et al. (2002) found that the persistence of clusters could be used to predict next season's influenza sequences. In their analysis, however, some sites of high variability in HA2 are neglected. Their results are not as detailed as ours. Through choosing those most variable sites of the whole sequence, all the potential evolutionary hot spots can be taken into account.

Our method can be applied to other components of the influenza virus genome. The comparisons among studies of other genetic parts of the influenza virus can be conducted to provide more information for epidemics prediction and vaccine design.

We use the term "dominant" to denote clusters with the greatest number of

sequences in a given season, but this definition only identifies the cluster containing the largest number of unique sequences. Therefore, the definition does not account for (a) the actual number of sequences reported in a season, or (b) their relation to the frequency of the strain within the population. Although the first issue is relatively straightforward to fix, the second is more problematic. The influenza sequences available via IRD are based on voluntary contributions and are therefore not the result of random sampling. It is thus possible that systematic biases exist in the data set, including yearly and regional variations (Łuksza and Lässig, 2014). Translating the observed sequences on IRD into an appropriate representation of population-level frequencies is an important statistical problem which requires careful consideration in our future work.

Lastly, we point out that our analysis is based on the Hamming distance (2.4), which means that we regard the sequences close in Hamming distance (in amino acids) as close in lineage evolution history. This approach is purely mathematical in that it does not include any potential information on the level of importance of specific amino acid differences, or their locations. Incorporating such additional information will improve on the quality of our analysis, and will be included in future analysis.

# 3 Tests for homogeneity of variances applied to DNA methylation data

As mentioned in Chapter 1, other than differentially methylated marks, differentially variable methylation marks are also relevant to some diseases. However, many inferences are presented based on the $F$ test or Bartlett's test. Both tests are sensitive to the departure of the normality assumption and the presence of outliers. More than 50 tests have been proposed in the statistical literature to improve the $F$ test/Bartlett's test. Conover et al. (1981) compared 56 equal-variance testing procedures using simulation studies, with the Brown and Forsythe's test being one of the top performers. The Brown and Forsythe's test has larger statistical power than other tests when samples are from non-normal distributions, while it maintains the nominal Type I error rate. To our knowledge, the Brown and Forsythe's test has not yet been applied to DNA methylation data. Phipson and Oshlack (2014) compared their proposed two equal-variance tests with $F$ test and Bartlett's test using

simulated data and evaluated the impact of outliers on the performance of the tests. However, systematic comparisons among these equal-variance tests are still needed in order to evaluate their performance for different distributions and the presence of outliers. In this chapter, we aim to help researchers choose the right test for equal variances in their DNA methylation data analysis. We compare Phipson and Oshlack's equal-variance tests and five commonly used equal-variance tests in the literature ($F$ test, Bartlett's test, Levene's test, trimmed-mean-based Levene's test, and Brown-Forsythe test) via systematic simulation studies and real data analysis.

## 3.1 Methodology

The scientific question we would like to address is to test if the variances of two populations (e.g., diseased and non-diseased subjects) are the same based on their corresponding samples. Let $X_i$ and $Y_i$ denote the methylation value and the disease status of subject $i$, where $i = 1, ..., n$, with $n = n_0 + n_1$, $n_0$ is the number of the non-diseased subjects (controls, $Y_i = 0$) and $n_1$ is the number of the diseased subjects (cases, $Y_i = 1$). We would like to test the null hypothesis $H_0 : \sigma_0^2 = \sigma_1^2$ versus the alternative hypothesis $H_a : \sigma_0^2 \neq \sigma_1^2$, where $\sigma_0^2$ and $\sigma_1^2$ are the variances of the non-diseased subjects and diseased subjects, respectively.

Next, we would like to compare the performance of the seven equal-variance tests:

$F$ test, Bartlett's test, Levene's test, trimmed-mean-based Levene's test, Brown and Forsythe's test, Phipson and Oshlack's equal variance test based on absolute difference, and Phipson and Oshlack's equal variance test based on squared difference. We denoted the seven tests by $F$, Bartlett, Levene, L.trim, BF, PO.AD, and PO.SQ, respectively.

The $F$ test is to test homogeneity of variance for a two-sample problem based on the ratio of variances. To test homogeneity of variance in multiple-sample situation, one popular test is Bartlett's test (Bartlett, 1937; Shoemaker, 2003). The Levene, L.trim, BF, PO.AD, and PO.SQ tests employ the ideas of equal-mean tests (e.g., t-test or one-way ANOVA) and replace the original data $x_{ki}$ in the test statistics by the transformed data $z_{ki} = |x_{ki} - c|$ or $z_{ki} = (x_{ki} - c)^2$, where the subscription $k$ indicates the group, $i$ indicates the subject within the group, and $c$ is a measure of central tendency, such as within-group mean or overall mean.

Specifically, Levene, L.trim, and BF tests replace $x_{ki}$ by $z_{ki}$ in one-way ANOVA's $F$ test statistic; PO.AD and PO.SQ tests replace $x_{ki}$ by $z_{ki}$ in the moderated t-test statistic (Smyth, 2004). The definitions of these seven equal-variance tests are given as follows.

Phipson and Oshlack (2014) proposed two equal-variance tests based on the fol-

lowing two linear regression models:

$$\begin{aligned} z_i^* =& \beta_0 + \beta_1 y_i + \epsilon_i, \\ z_i^{**} =& \gamma_0 + \gamma_1 y_i + \xi_i, \\ i =& 1, \ldots, n_0 + n_1, \end{aligned} \tag{3.1}$$

where

$$\begin{aligned} z_i^* =& c_i^* |x_i - g(x_i)|, \\ z_i^{**} =& c_i^* [x_i - g(x_i)]^2, \end{aligned}$$

and $g(x_i) = \frac{1}{n_0} \sum_{i=1}^{n} x_i I(y_i = 0)$ for controls or $g(x_i) = \frac{1}{n_1} \sum_{i=1}^{n} x_i I(y_i = 1)$ for cases. The value of $c_i^*$ is

$$c_i^* = \begin{cases} \sqrt{\frac{n_0}{n_0 - 1}} & \text{if } y_i = 0 \text{ (i.e., controls)}, \\ \sqrt{\frac{n_1}{n_1 - 1}} & \text{if } y_i = 1 \text{ (i.e., cases)}, \end{cases}$$

In matrix terminology, the two linear models can be written as:

$$\begin{aligned} E(\boldsymbol{z}^*) =& \boldsymbol{Y}_d \boldsymbol{\beta}, \\ E(\boldsymbol{z}^{**}) =& \boldsymbol{Y}_d \boldsymbol{\gamma}, \end{aligned} \tag{3.2}$$

where

$$\boldsymbol{Y}_d = \begin{pmatrix} 1, y_1 \\ \vdots, \vdots \\ 1, y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix}.$$

And the regression coefficients can be estimated as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{Y}_d^T \boldsymbol{Y}_d)^{-1} \boldsymbol{Y}_d^T \boldsymbol{z}^*, \\ \hat{\boldsymbol{\gamma}} &= (\boldsymbol{Y}_d^T \boldsymbol{Y}_d)^{-1} \boldsymbol{Y}_d^T \boldsymbol{z}^{**}, \end{aligned} \tag{3.3}$$

Phipson and Oshlack (2014) mentioned that testing for equal-variance between cases and controls is equivalent to testing if the slope $\beta_1$ (or $\gamma_1$) is equal to zero. Phipson and Oshlack (2014) applied moderated t-test (Smyth, 2004) to borrow information across CpG sites to improve the test of the null hypothesis $H_0 : \beta_1 = 0$ (or $\gamma_1 = 0$) for a given CpG site. The moderated t-test statistics is defined as:

$$\tilde{t} = \frac{\hat{\beta}_1}{\tilde{s}\sqrt{\nu}}, \tag{3.4}$$

where $\nu$ is the diagonal element from the positive definite matrix $(\boldsymbol{Y}_d^T \boldsymbol{Y}_d)^{-1}$ and $\tilde{s}$ is the standard deviation of the squeezed variance calculated according to Smyth's (2004) procedures.

The $F$ test statistic is asymptotically F distributed under the null hypothesis:

$$F = \frac{S_1^2}{S_0^2} \xrightarrow{d} F_{n_1-1,n_0-1},$$

where

$$S_k^2 = \frac{1}{(n_k - 1)} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2,$$
$$\bar{x}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ki},$$

$k = 0, 1$, $n_0$ and $n_1$ are the sample sizes for controls and cases, respectively, and $N = n_0 + n_1$.

For two-sample comparison ($K = 2$), Bartlett's test statistic is

$$X^2 = \frac{(n_1 + n_0 - 2)\log(S_p^2) - \sum_{k=1}^{2}(n_k - 1)\log(S_k^2)}{1 + \frac{1}{3}\left(\sum_{k=1}^{2}\left(\frac{1}{(n_k-1)}\right) - \frac{1}{(n_1+n_0-2)}\right)} \xrightarrow{d} \chi_1^2,$$

where

$$S_p^2 = \frac{1}{(N-2)} \sum_{k=1}^{2} (n_k - 1) S_k^2.$$

The numerator of the Bartlett's test for two-sample comparison of variances is

$$\begin{aligned}
numer &= (n_1 + n_0 - 2) \log\left(S_p^2\right) - \left[(n_1 - 1) \log\left(S_1^2\right) + (n_0 - 1) \log\left(S_0^2\right)\right] \\
&= \log\left\{ \frac{\left[(S_p)^2\right]^{n_1+n_0-2}}{\left[(S_1)^2\right]^{n_1-1} \left[(S_0)^2\right]^{n_0-1}} \right\} \\
&= \log\left\{ \left[\frac{(S_p)^2}{(S_1)^2}\right]^{n_1-1} \left[\frac{(S_p)^2}{(S_0)^2}\right]^{n_0-1} \right\} \\
&= \log\left\{ \left[\frac{1}{(n_1+n_0-2)} \left[(n_1-1) + (n_0-1)\frac{S_0^2}{S_1^2}\right]\right]^{n_1-1} \left[\frac{1}{(n_1+n_0-2)} \left[(n_0-1) + (n_1-1)\frac{S_1^2}{S_0^2}\right]\right]^{n_0-1} \right\} \\
&= \log\left\{ \left[\frac{1}{(n_1+n_0-2)} \left[(n_1-1) + (n_0-1)\frac{1}{F}\right]\right]^{n_1-1} \left[\frac{1}{(n_1+n_0-2)} \left[(n_0-1) + (n_1-1)F\right]\right]^{n_0-1} \right\},
\end{aligned}$$

where $\frac{S_1^2}{S_0^2}$ is the $F$ test statistic.

Hence, Bartlett's test for two-sample comparison of variances is similar to $F$ test, but they have small differences in performance when applied to data with a small sample size due to different derived asymptotic distributions.

Levene's test statistic for two-sample comparison of variances is defined as

$$W = \frac{(n-2)[n_1(\bar{w}_1 - \bar{w})^2 + n_0(\bar{w}_0 - \bar{w})^2]}{\sum_{i=1}^{n_1}(w_{1i} - \bar{w}_1)^2 + \sum_{j=1}^{n_0}(w_{0j} - \bar{w}_0)^2},$$

where

$$
\begin{aligned}
w_{1i} =& |x_{1i} - \bar{x}_1|, \quad \bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}, \\
w_{0j} =& |x_{0j} - \bar{x}_0|, \quad \bar{x}_0 = \frac{1}{n_0}\sum_{j=1}^{n_0} x_{0i}, \\
\bar{w}_1 =& \frac{1}{n_1}\sum_{i=1}^{n_1} w_{1i}, \\
\bar{w}_0 =& \frac{1}{n_0}\sum_{j=1}^{n_0} w_{1j}, \\
\bar{w} =& \frac{1}{n}\left[\sum_{i=1}^{n_1} w_{1i} + \sum_{j=1}^{n_0} w_{0j}\right].
\end{aligned}
$$

Trimmed-mean-based Levene's test for two-sample comparison of variances has the same format as Levene's test. The only difference is in the definition of $w_{1i}$ and $w_{0j}$:

$$
\begin{aligned}
w_{1i} =& |x_{1i} - \check{x}_1|, \\
w_{0j} =& |x_{0j} - \check{x}_0|,
\end{aligned}
$$

where $\check{x}_1$ and $\check{x}_0$ are within-group 10% trimmed means for cases and controls, respectively.

Brown-Forsythe test statistic uses the same format as Levene's test for two-sample comparison of variances. The only difference is in the definition of $w_{1i}$ and $w_{0j}$:

$$
\begin{aligned}
w_{1i} =& |x_{1i} - \tilde{x}_1|, \\
w_{0j} =& |x_{0j} - \tilde{x}_1|,
\end{aligned}
$$

where $\tilde{x}_1$ and $\tilde{x}_0$ are medians for cases and controls, respectively.

## 3.2 Simulation studies

### 3.2.1 Simulation setting

We conducted two simulation studies. Each study contains several scenarios and only evaluates the balanced samples. For each scenario, we generated 100 simulated data sets. For each simulated data set, we generated DNA methylation levels for 1000 CpG sites. For each CpG site, we tested if the DNA methylation levels are differentially variable between non-diseased and diseased subjects using each of the seven equal-variance tests. A test is claimed as significant if its p-value is $< 0.05$. Two-sided tests were used by the R statistical software (R Core Team, 2008) for the simulation studies.

In Simulation Study I, we evaluated the performance of the seven tests by the following aspects: (1) the violation of the normality assumption, (2) the presence of heterogeneity of means, (3) the existence of outliers, (4) various sample sizes. We employed three parametric distributions to generate the methylation data: normal distribution, t distribution, and chi-squared distribution. To evaluate the impact of different group means on these tests, we considered two scenarios: equal group means (eqM) and different group means (diffM). To evaluate the influence of outliers, we randomly picked a diseased subject and replaced its DNA methylation value by the

maximum DNA methylation level of all CpG sites. We also used three sample sizes: 20 (small), 50 (median), and 200 (large), to evaluate the effect of sample size on the performance.

The distribution settings for the scenarios in Simulation Study I are summarized in Table 3.1. Simulation study I had 3 (distributions) $\times$ 2 (scenarios of group means) $\times$ 2 (with or without outlier) $\times$ 3 (sample sizes) = 36 different comparisons.

Table 3.1: The distribution settings for the scenarios in Simulation Study I.

| Mean & Variance | Normal | | t distribution | | chi-squared distribution | |
|---|---|---|---|---|---|---|
| (mean, var) | Non-D | D | Non-D | D | Non-D | D |
| eqM & eqV | $N(0,1)$ | $N(0,1)$ | $t_1 0$ (0, 1.25) | $t_1 0$ (0, 1.25) | $\chi^2_2$ (2,4) | $\chi^2_2$ (2,4) |
| eqM & diffV | $N(0,1)$ | $N(0,2)$ | $t_1 0$ (0, 1.25) | $t_{10/3}$ (0, 2.5) | $\chi^2_2$ (2,4) | $\chi^2_{0.5,1.5}$ (2,7) |
| diffM & eqV | $N(0,1)$ | $N(1.5,1)$ | $t_1 0$ (0, 1.25) | $t_{15,1.489}$ (1.57, 1.25) | $\chi^2_2$ (2,4) | $\chi^2_{1,0.5}$ (1.5,4) |
| diffM & diffV | $N(0,1)$ | $N(1.5,2)$ | $t_1 0$ (0, 1.25) | $t_{6,2.393}$ (2.75, 2.5) | $\chi^2_2$ (2,4) | $\chi^2_4$ (4,8) |

eqM: equal-mean; eqV: equal-variance; diffM: different-mean; diffV:different-variance; D: diseased; Non-D: non-diseased; $N(a,b)$: normal distribution with mean $a$ and variance $b$; $t_c$: t-distribution with degrees of freedom $c$; $t_{d,e}$: non-central t-distribution with degrees of freedom $d$ and non-centrality parameter $e$; $\chi^2_f$: chi-squared distribution with degrees of freedom $f$; $\chi^2_{g,h}$: non-central chi-squared distribution with degrees of freedom $g$ and non-centrality parameter $h$.

In Simulation Study II, we considered two Bayesian hierarchical models. First, we generated the DNA methylation values from a normal distribution with the variance sampled from an inverse chi-squared distribution (denoted as c.N). Second, we generated the DNA methylation values from a chi-squared distribution, the degrees

of freedom of which were generated from the scaled inverse chi-squared distribution scale-inv-$\chi^2(d_0, s_0^2)$ (denoted as c.chisq). To evaluate the type I error rate (equal-variance scenario), we set the degrees of freedom $d_0 = 20$ and the scaling factor $s_0^2 = 0.64$ for both non-diseased and diseased subjects. To evaluate the power of the tests (different-variance scenario), we set the scaling factor as $s_0^2 = 0.64$ for non-diseased subjects and $s_0^2 = 1.5$ for diseased subjects. The degrees of freedom are set to be $d_0 = 20$ for both non-diseased and diseased subjects. To evaluate the effect of outliers and sample size, we used the same procedures in Simulation Study I. Thus, Simulation Study II had 2 (distributions) $\times$ 2 (with or without outlier) $\times$ 3 (sample sizes) = 12 different comparisons.

For each simulated data set, we assessed the performance of an equal-variance test by the estimated type I error rate and power. The estimated type I error rate is the proportion of significant tests detected by the equal-variance test among the 1000 CpG sites in a simulated data set generated from the null hypothesis (i.e., CpG sites are non-differentially variable). Estimated power is the proportion of significant tests detected by the equal-variance test among the 1000 CpG sites in a simulated data set generated from the alternative hypothesis (i.e., CpG sites are differentially variable).

### 3.2.2 Simulation results

Figure 3.1 and Table 3.2 summarize the observed results from all the scenarios: (1) $F$ test, Bartlett's test, Levene's test, and PO.AD test have type I error rates higher than the nominal value (0.05) in most scenarios, while L.trim, BF, and PO.SQ maintain the nominal type I error rates for almost all scenarios; (2) BF test and PO.SQ test perform better than the other tests in terms of having high power while maintaining the nominal type I error rate; (3) $F$ test and Bartlett's test have very similar performance and perform best under normality assumption, while both of them have type I error rates higher than the nominal value (0.05) when the normality assumption is violated; (4) PO.AD test tends to have a type I error rate higher than the nominal value of 0.05 for a majority of simulation scenarios, while PO.SQ test can maintain the nominal type I error rate of 0.05 for almost all simulation scenarios; (5) For almost all of the scenarios where PO.AD test maintains the nominal type I error rate, PO.AD test has the largest power, while the PO.SQ test is less powerful than other tests in the similar situation of PO.AD; and (6) the power has improved a lot by increasing the sample size from 20/50 subjects per group to 200 subjects per group. In addition, we observe that the ranks of the seven tests do not change much as the sample size increases. The ranks of power by different sample sizes are presented in Appendix A.

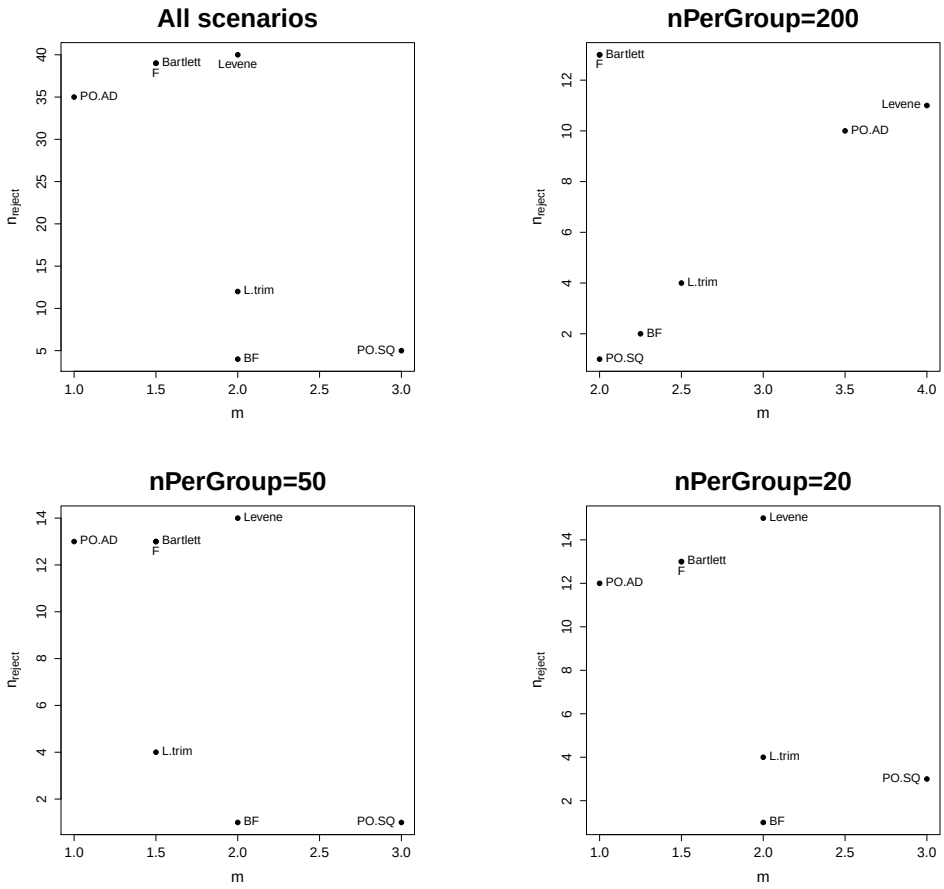


Figure 3.1: Plots of $n_{reject}$ versus m, where $n_{reject}$ is the number of scenarios where the test has inflated type I error rate and $m$ is the median of the ranks of power. For ranks with ties, average ranks are used. The upper-right, bottom-left, and bottom-right panels are based on scenarios with sample size 200, 50, and 20 subjects per group, respectively. The upper-left panel is based on all 48 scenarios.

Table 3.2: The summary of the simulated 48 comparisons for the seven tests.

| n | Distribution | F | Bartlett | Levene | L.trim | BF | PO.AD | PO.SQ |
|---|---|---|---|---|---|---|---|---|
| 200 | N(incl. c.N) | 2(3) | 2(3) | 5(3) | 3.25(0) | 3.25(0) | 4.75(2) | 2(1) |
| 200 | t | -(4) | -(4) | 2(2) | 3(1) | 4(1) | 2(2) | 2(0) |
| 200 | chisq (incl. c.chisq) | -(6) | -(6) | -(6) | 1(3) | 1.5(1) | -(6) | 2.5(0) |
| 50 | N(incl. c.N) | 1.5(3) | 1.5(3) | 2(5) | 4(1) | 4.5(0) | 1(4) | 3(1) |
| 50 | t | -(4) | -(4) | 2(3) | 1(0) | 2(0) | 1(3) | 3(0) |
| 50 | chisq (incl. c.chisq) | -(6) | -(6) | -(6) | 1(3) | 2(1) | -(6) | 2.5(0) |
| 20 | N(incl. c.N) | 1.5(3) | 1.5(3) | -(6) | 4(1) | 4(0) | 1(4) | 3(1) |
| 20 | t | -(4) | -(4) | 2(3) | 1.5(0) | 2.5(0) | 1(2) | 3.5(0) |
| 20 | chisq (incl. c.chisq) | -(6) | -(6) | -(6) | 1(3) | 2(1) | -(6) | 3(2) |
| | Total | 1.5(39) | 1.5(39) | 2(40) | 2(12) | 2(4) | 1(35) | 3(5) |

$^{N}$ : Normal distribution;

$^{c.N}$ : Bayesian hierarchical model with normal distribution;

$^{t}$ : t distribution;

$^{chisq}$ : Chi-squared distribution;

$^{c.chisq}$ : Bayesian hierarchical model with chi-squared distribution;

$^{m(n_{reject})}$ : $m$ denotes the median of the ranks of the power, $n_{reject}$ denotes the number of the scenarios where the test has inflated type I error rate;

"-" : no power can be considered because the test has inflated type I error rates for all the scenarios in the situation.

## 3.3 Real data analysis

### 3.3.1 Data description

To evaluate the performance of the seven equal-variance tests for real data sets, we used two data sets (GSE37020 and GSE20080) downloaded from the public repository: Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo). Both data sets contain DNA methylation (DNAm) profiles of 27,578 CpG sites measured from liquid based cytology (LBC) cervical smear samples by IlluminaHumanMethylation27 platform.

GSE37020 contains a total of 48 samples, 24 of which have normal histology and the remaining 24 are cervical intraepithelial neoplasia of grade 2 or higher (CIN2+). All of them are human papillomavirus (HPV) positive. Normal and CIN2+ samples are age-matched. GSE20080 also contains 48 samples. A total of 30 samples (11 HPV positive samples and 19 HPV negative samples) have normal cytology. The other 18 samples (all HPV positive) are with CIN2+. Moreover, normal and CIN2+ samples were age-matched. After the procedures of quality control and data preprocessing, the remaining 22,859 are matched CpG sites in both data sets. We used these 22,859 CpG sites in our real data analysis. The procedures and results of quality control (QC) and data preprocessing are presented in Appendix B.

We used GSE37020 as the discovery set to detect CpG sites differentially variable between normal cytology samples and CIN2+ samples. To control for multiple comparisons, we applied the Benjamini and Hochberg's method to adjust p-values so that the false discovery rate (FDR) is controlled at the level of 0.05. Specifically, a CpG site was claimed significant if its FDR-adjusted p-value was $< 0.05$. We then validated these differentially variable CpG sites by using the GSE20080 data set. If an equal-variance test for a given CpG site had FDR-adjusted p-value $< 0.05$ in the analysis of GSE37020 and had un-adjusted p-value $< 0.05$ in the analysis of GSE20080, we then claimed that the significance of the test in GSE37020 was validated in GSE20080.

### 3.3.2 Results

For the real data set GSE37020, the numbers of significant CpG sites (i.e., CpG sites with FDR-adjusted p-value $< 0.05$) obtained by the seven equal-variance tests are 2318 (F), 2315 (Bartlett), 235 (Levene), 15 (L.trim), 7 (BF), 130 (PO.AD), and 0 (PO.SQ), respectively. The numbers of significant CpG sites detected by $F$ test and Bartlett test are much larger than those detected by other tests. No significant CpG sites were detected by the PO.SQ test.

The numbers/proportions of significant CpG sites validated by GSE20080 are

1154/49.8% (F), 1164/50.3% (Bartlett), 183/77.9% (Levene), 9/60.0% (L.trim), 3/42.9% (BF), and 91/70% (PO.AD), respectively (see Table 3.3). The six tests (except PO.SQ) have large proportions of significant CpG sites validated by the testing set GSE20080. Overall, the robust equal-variance tests have a larger proportion of validated significant CpG sites than F or Bartlett test.

Since the $F$ test and Bartletts test are sensitive to outliers, we check the number/proportion of significant CpG sites containing outliers detected based on GSE37020. The numbers/proportions are 1503/64.8% (F), 1501/64.8% (Bartlett), 70/29.8% (Levene), 2/13.3% (L.trim), 2/28.6% (BF), and 64/49.2% (PO.AD), respectively (see the second column of Table 3.3). For the $F$ test and Bartletts test, more than 60% significant CpG sites contain outliers. For robust tests (e.g., Levene, L.trim, and BF), the proportions are relatively small ($< 30\%$).

Table 3.3: The performance of seven equal-variance tests based on data sets GSE37020 and GSE20080.

| Test | GSE37020 nSig(p.adj $< 0.05$) | GSE20080 nValid(pval $< 0.05$) | pValid |
|---|---|---|---|
| F | 2318 | 1154 | 49.8% |
| Bartlett | 2315 | 1164 | 50.3% |
| Levene | 235 | 183 | 77.9% |
| L.trim | 15 | 9 | 60.0% |
| BF | 7 | 3 | 42.9% |
| PO.AD | 130 | 91 | 70% |
| PO.SQ | 0 | 0 | - |

$^{\text{nSig}}$ : the number of significant CpG sites detected in GSE37020 based on FDR-adjusted p-value $< 0.05$;

$^{\text{nValid}}$ : the number of validated CpG sites in GSE20080 based on unadjusted p-value $< 0.05$;

$^{\text{pTV}}$ : $= \frac{nValid}{nSig}$, the proportion of significant CpG sites detected in GSE37020 and validated in GSE20080;

We then checked if the significant CpG sites containing outliers in GSE37020 would still contain outliers in GSE20080. The number/proportion of such CpG sites are 495/32.9% (F), 497/33.1% (Bartlett), 34/48.6% (Levene), 0/0% (L.trim), 0/0% (BF), and 31/48.4% (PO.AD), respectively (see third column of Table 3.3).

Table 3.4: The number (proportion) of significant CpG sites that contain outliers in GSE37020 and GSE20080.

| Test | GSE37020 nOut.Sig (pOut.Sig) | GSE20080 nOut.Valid (pOut.Valid) |
|---|---|---|
| F | 1503 (64.8%) | 495 (32.9%) |
| Bartlett | 1501 (64.8%) | 497 (33.1%) |
| Levene | 70 (29.8%) | 34 (48.6%) |
| L.trim | 2 (13.3%) | 0 (0%) |
| BF | 2 (28.6%) | 0 (0%) |
| PO.AD | 64 (49.2%) | 31 (48.4%) |
| PO.SQ | 0 (-) | 0 (-) |

$^{\text{nOut.Sig (pOut.Sig)}}$ : the number (proportion) of significant CpG sites containing outliers detected in GSE37020;

$^{\text{nOut.Valid (pOut.Valid)}}$ : the number (proportion) of the significant CpG sites with outliers that also contain outliers in GSE20080.

We next checked the parallel boxplots of DNA methylation level versus case-control status for the top CpG site (i.e., having the smallest p-value for testing equal variance) obtained by each of the seven tests based on GSE37020. The top CpG sites detected by the seven equal-variance tests are cg26363196 (F, Bartlett, PO.AD), cg00027083 (Levene and L.trim), and cg06675478 (BF), respectively. All these top CpG sites were validated in GSE20080. Figure 3.2 shows the boxplots of these three unique top CpG sites. We found that all these three top CpG sites

contain at least one outlier in either GSE37020 or GSE20080.

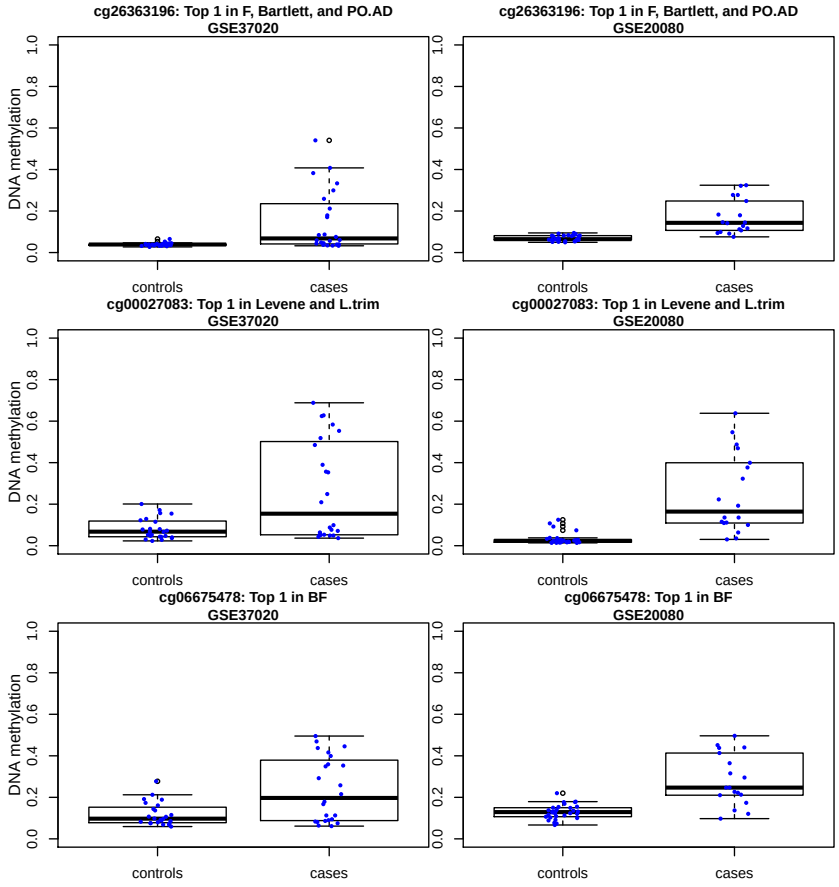


Figure 3.2: Parallel boxplots of DNA methylation level versus case-control status for the obtained three unique top CpG sites.

## 3.4 Discussion

Recently, Phipson and Oshlack (2014) proposed two new tests for homogeneity of variances for DNA methylation data analysis. However, the performance of their methods has not been compared with existing tests that are robust against the violation of the normality assumption, such as Levene's test, trimmed-mean-based Levene's test, and Brown Forsythe's test.

In this chapter, we systematically compare the performance of the two new equal-variance tests with the $F$ test, Bartlett's test, Levene's test, trimmed-mean-based Levene's test, and Brown Forsythe's test via two sets of simulation studies and one real-data analysis. Based on the simulation results, BF, L.trim, and PO.SQ tests for equality of variance have relatively high power while keeping the nominal type I error rate for most of the simulation scenarios. Levene's test has type I error rates higher than the nominal value 0.05 for a majority of the scenarios, even for the scenarios where data are generated from a normal distribution. All the seven equal-variance tests have low power when data are generated from chi-squared distributions. Compared to real DNA methylation data, our simulation studies do not cover all scenarios encountered in real DNA methylation data analysis. However, our simulation studies provide useful information about the performance of the seven equal-variance tests.

Our simulation studies and real data analysis confirm the fact that F/Bartlett's

test are very sensitive to the violation of the normality assumption and the presence of outliers. However, outliers might be biologically important as pointed out by Teschendorff and Widschwendter. The real data analysis in this chapter also shows that a number (30% - 50%) of significant CpG sites with outliers detected in GSE37020 also contain outliers in GSE20080. Our real data analysis also agrees with Teschendorff and Widschwendter's observation that changes in DNA methylation for differentially variable CpG sites are heterogeneous and stochastic as shown in the parallel boxplots for CpG cg26363196 in Figure 3.2. We notice in the real data analyses that some outliers might be artifacts. For example, more than 60% of the 1503 significant CpG sites containing outliers do not contain outliers in GSE20080.

# 4 Robust joint score tests for DNA methylation data analysis

Since both mean and variance are biologically meaningful in DNA methylation analysis, it is more efficient to test for equal means and equal variances simultaneously. The joint likelihood ratio test (jointLRT) and the two-sample Kolmogorov-Smirnov (KS) test are two traditional methods for this task. Ahn and Wang (2013) proposed a joint score test (AW), which is a quadratic form of a vector of two tests. One of them is to test for equal means, and the other is to test for equal variances. However, they did not provide the derivation of the asymptotic distribution for this test nor the comparison of AW with jointLRT or KS that are the benchmark tests in the statistical literature.

In this chapter, we derive the asymptotic distribution of the AW joint test statistic and make comprehensive comparisons between AW, jointLRT, and KS tests. Although a normal distribution is usually assumed for methylation data, the viola-

tion of the normality assumption and the presence of outlying points can often be observed in the analysis of real data. Bi-modal distributions are also encountered frequently in practice. To improve on the robustness of the AW joint test, we propose three tests based on absolute deviation from mean (iAW.Lev), median (iAW.BF), and trimmed mean (iAW.TM), respectively.

## 4.1 Methodology

### 4.1.1 Asymptotic distribution of AW-type joint test statistics

Let $X_i$ and $Y_i$ denote the methylation value and the disease status of subject $i$, where $i = 1, \ldots, n$, with $n = n_0 + n_1$, $n_0$ being the number of non-diseased subjects (controls, $Y_i = 0$) and $n_1$ being the number of diseased subjects (cases, $Y_i = 1$). To detect methylation loci that are relevant to the disease based on means and variances, the corresponding hypothesis is considered as $H_0 : \mu_0 = \mu_1$ and $\sigma_0^2 = \sigma_1^2$ versus $H_1 : \mu_0 \neq \mu_1$ or $\sigma_0^2 \neq \sigma_1^2$, in which $\mu_0$ and $\mu_1$ are means of methylation levels for controls and cases, respectively, and $\sigma_0^2$ and $\sigma_1^2$ are the corresponding variances.

Instead of directly testing the above hypothesis, Ahn and Wang (2013) proposed to test $H_0' : \beta_1 = \beta_2 = 0$ versus $H_a' : \beta_1 \neq 0$ or $\beta_2 \neq 0$, where $\beta_1$ and $\beta_2$ are the

regression coefficients of the following logistic regression:

$$logit[Pr(Y_i = 1|x_i, z_i)] = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i, \tag{4.1}$$

and $z_i$ is the within-group squared deviation for subject $i$, which is defined as

$$z_i = \begin{cases} (x_i - \bar{x}_1)^2, \text{ if } Y_i = 1 \\ \\ (x_i - \bar{x}_0)^2, \text{ if } Y_i = 0, \end{cases} \tag{4.2}$$

and $\bar{x}_1 = \sum_{i=1}^{n} \frac{x_i I[Y_i=1]}{n_1}$ and $\bar{x}_0 = \sum_{i=1}^{n} \frac{x_i I[Y_i=0]}{n_0}$ are the sample means for cases and controls.

The log-likelihood function of the logistic regression (4.1) is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_i + \beta_2 z_i) - \log[1 + exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)],$$

where $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2)^T$. The score statistics are partial derivatives of the log-likelihood function with respect to the parameters of interest, evaluated at the values postulated by the null hypothesis $H_0' : \beta_1 = \beta_2 = 0$.

We have

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_0} &= \sum_{i=1}^{n} (Y_i - \pi_i), \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_1} &= \sum_{i=1}^{n} x_i (Y_i - \pi_i), \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_2} &= \sum_{i=1}^{n} z_i (Y_i - \pi_i), \end{aligned}$$

where

$$\pi_i = Pr(Y_i = 1|x_i, z_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)}.$$

Under $H_0' : \beta_1 = \beta_2 = 0$,

$$\pi_i = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} = \pi_0.$$

Let $\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_0} = 0$ under $H_0'$. The maximum likelihood estimate of $\pi_0$ is:

$$\hat{\pi}_0 = \bar{Y} = \sum_{i=1}^{n} \frac{Y_i}{n}.$$

Hence, the score statistics are

$$U_1 = \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_1}\Big|_{\hat{\pi}_0 = \bar{Y}, \beta_1 = \beta_2 = 0} = \sum_{i=1}^{n} x_i (Y_i - \bar{Y}),$$
$$U_2 = \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_2}\Big|_{\hat{\pi}_0 = \bar{Y}, \beta_1 = \beta_2 = 0} = \sum_{i=1}^{n} z_i (Y_i - \bar{Y}).$$

Under $H_0'$, let $\mathbf{U} = (U_1, U_2)^T$ and $\boldsymbol{\Sigma}_0 = Cov(\mathbf{U})$ denote the vector of the score statistics and the covariance matrix of $\mathbf{U}$, respectively. Based on Fahrmeir (1987) and Dobson (1990, Page 51),

$$\mathbf{U} \xrightarrow{d} N(0, \boldsymbol{\Sigma}_0),$$

and

$$\boldsymbol{\Sigma}_0^{-1/2}\mathbf{U} \xrightarrow{d} N(0, \mathbf{I}_2),$$

where $\mathbf{I}_2$ is a $2 \times 2$ identity matrix. Therefore,

$$(\boldsymbol{\Sigma}_0^{-1/2}\mathbf{U})^T(\boldsymbol{\Sigma}_0^{-1/2}\mathbf{U}) = \mathbf{U}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{U} \xrightarrow{d} \chi_2^2. \tag{4.3}$$

The AW test statistic $T = \mathbf{U}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{U}$ is a quadratic form of the vector of the score statistics, where $\hat{\mathbf{\Sigma}}$ is the estimate of the covariance matrix $\mathbf{\Sigma}_0$ under $H_0'$.

**Proposition 4.1.1** *Under $H_0' : \beta_1 = \beta_2 = 0$, the estimated covariance matrix $\hat{\mathbf{\Sigma}}$ has the following analytical form:*

$$\widehat{\mathbf{\Sigma}} = n\bar{y}(1-\bar{y}) \begin{pmatrix} \hat{\sigma}_x^2 & \hat{\sigma}_{xz} \\ \hat{\sigma}_{xz} & \hat{\sigma}_z^2 \end{pmatrix}, \tag{4.4}$$

*where $\hat{\sigma}_x^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n}$ and $\hat{\sigma}_z^2 = \sum_{i=1}^n \frac{(z_i - \bar{z})^2}{n}$ are the sample variances for $x_i$ and $z_i$, and $\hat{\sigma}_{xz} = \sum_{i=1}^n \frac{(x_i - \bar{x})(z_i - \bar{z})}{n}$ is the sample covariance between $x_i$ and $z_i$, and $\bar{y}$ is the realization of $\bar{Y}$. The asymptotic distribution of $T = \mathbf{U}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{U}$ is a chi-squared distribution with two degrees of freedom.*

**Proof.** Note that in logistic regression, $Y_i$ are random variables, while $x_i$ and $z_i$ are fixed (i.e., non-random). We can get

$$\begin{aligned} \mathrm{E}(U_1) &= \sum_{i=1}^n x_i \mathrm{E}(Y_i - \bar{Y}) = 0, \\ \mathrm{E}(U_2) &= \sum_{i=1}^n z_i \mathrm{E}(Y_i - \bar{Y}) = 0. \end{aligned}$$

Hence, we have

$$\begin{aligned} \mathbf{\Sigma}_0 =& \mathrm{Cov}(\mathbf{U}) \\ =& \mathrm{E}\left(\mathbf{U}\mathbf{U}^T\right) - [\mathrm{E}(\mathbf{U})][\mathrm{E}(\mathbf{U})]^T \\ =& \mathrm{E}\left(\mathbf{U}\mathbf{U}^T\right). \end{aligned} \tag{4.5}$$

Define $\mathbf{J} = (1)_{n\times n}$, $\mathbf{I}$ is a $n \times n$ identity matrix, $\mathbf{X} = \begin{pmatrix} x_1 & z_1 \\ \vdots & \vdots \\ x_n & z_n \end{pmatrix}$, $\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$,

then $\bar{\mathbf{X}} = \frac{1}{n}\mathbf{J}\mathbf{X}$, $\mathbf{U} = \begin{pmatrix} \sum_{i=1}^{n} x_i(Y_i - \bar{Y}) \\ \sum_{i=1}^{n} z_i(Y_i - \bar{Y}) \end{pmatrix} = \mathbf{X}^T(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y}$.

$$\begin{aligned} \mathrm{E}\left(\mathbf{U}\mathbf{U}^T\right) =& \mathrm{E}\left[\mathbf{X}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}\mathbf{Y}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{X}\right] \\ =& \mathrm{E}\left[(\mathbf{X} - \bar{\mathbf{X}})^T\mathbf{Y}\mathbf{Y}^T(\mathbf{X} - \bar{\mathbf{X}})\right] \\ =& (\mathbf{X} - \bar{\mathbf{X}})^T\mathrm{E}\left(\mathbf{Y}\mathbf{Y}^T\right)(\mathbf{X} - \bar{\mathbf{X}}). \end{aligned}$$

We have

$$\begin{aligned} \mathbf{Y}\mathbf{Y}^T &= \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}\begin{pmatrix} Y_1, \cdots, Y_n \end{pmatrix} \\ &= \begin{pmatrix} Y_1^2 & Y_1Y_2 & \dots & Y_1Y_n \\ \vdots & \vdots & \ddots & \vdots \\ Y_nY_1 & Y_nY_2 & \dots & Y_n^2 \end{pmatrix}. \end{aligned}$$

Note that $Y_i^2 = Y_i$ and $\mathrm{E}\left(Y_i\right) = \pi_i$. We have

$$\begin{aligned} \mathrm{E}\left(\mathbf{Y}\mathbf{Y}^T\right) =& \mathrm{E}\begin{pmatrix} Y_1^2 & Y_1Y_2 & \dots & Y_1Y_n \\ \vdots & \vdots & \ddots & \vdots \\ Y_nY_1 & Y_nY_2 & \dots & Y_n^2 \end{pmatrix} \\ =& \mathrm{E}\begin{pmatrix} Y_1 & Y_1Y_2 & \dots & Y_1Y_n \\ \vdots & \vdots & \ddots & \vdots \\ Y_nY_1 & Y_nY_2 & \dots & Y_n \end{pmatrix} \\ =& \begin{pmatrix} \pi_1 & \pi_1\pi_2 & \dots & \pi_1\pi_n \\ \vdots & \vdots & \ddots & \vdots \\ \pi_n\pi_1 & \pi_n\pi_2 & \dots & \pi_n \end{pmatrix}. \end{aligned}$$

Under $H_0' : \beta_1 = \beta_2 = 0$, $\pi_i = \pi_0$. Therefore, we have

$$\begin{aligned}\mathrm{E}\left(\mathbf{Y}\mathbf{Y}^T\right) =& \begin{pmatrix} \pi_0 & \pi_0^2 & \dots & \pi_0^2 \\ \vdots & \vdots & \ddots & \vdots \\ \pi_0^2 & \pi_0^2 & \dots & \pi_0 \end{pmatrix} \\ =& \pi_0^2\mathbf{J} + \pi_0(1-\pi_0)\mathbf{I}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{E}\left(\mathbf{U}\mathbf{U}^T\right) =& (\mathbf{X}-\bar{\mathbf{X}})^T(\pi_0^2\mathbf{J} + \pi_0(1-\pi_0)\mathbf{I})(\mathbf{X}-\bar{\mathbf{X}}) \\ =& (\mathbf{X}-\bar{\mathbf{X}})^T\pi_0^2\mathbf{J}(\mathbf{X}-\bar{\mathbf{X}}) + (\mathbf{X}-\bar{\mathbf{X}})^T\pi_0(1-\pi_0)\mathbf{I}(\mathbf{X}-\bar{\mathbf{X}}) \\ =& \pi_0^2(\mathbf{X}-\bar{\mathbf{X}})^T\mathbf{J}(\mathbf{X}-\bar{\mathbf{X}}) + \pi_0(1-\pi_0)(\mathbf{X}-\bar{\mathbf{X}})^T\mathbf{I}(\mathbf{X}-\bar{\mathbf{X}}).\end{aligned}$$

Since $(\mathbf{X}-\bar{\mathbf{X}})^T\mathbf{J}(\mathbf{X}-\bar{\mathbf{X}}) = (0)_{2\times 2}$, we have

$$\begin{aligned}\mathrm{E}\left(\mathbf{U}\mathbf{U}^T\right) =& \pi_0(1-\pi_0)(\mathbf{X}-\bar{\mathbf{X}})^T\mathbf{I}(\mathbf{X}-\bar{\mathbf{X}}) \\ =& \pi_0(1-\pi_0)(\mathbf{X}-\bar{\mathbf{X}})^T(\mathbf{X}-\bar{\mathbf{X}}) \\ =& \pi_0(1-\pi_0)\begin{pmatrix} \sum_{i=1}^n (x_i-\bar{x})^2 & \sum_{i=1}^n (x_i-\bar{x})(z_i-\bar{z}) \\ \sum_{i=1}^n (x_i-\bar{x})(z_i-\bar{z}) & \sum_{i=1}^n (z_i-\bar{z})^2 \end{pmatrix}.\end{aligned}$$

Under $H_0' : \beta_1 = \beta_2$, we can estimate $\pi_0$ by the realization of $\bar{Y}$, $\bar{y} = \frac{n_1}{n}$. Thereby, we have

$$\begin{aligned}\hat{\mathbf{\Sigma}} =& \bar{y}(1-\bar{y})\begin{pmatrix} \sum_{i=1}^n (x_i-\bar{x})^2 & \sum_{i=1}^n (x_i-\bar{x})(z_i-\bar{z}) \\ \sum_{i=1}^n (x_i-\bar{x})(z_i-\bar{z}) & \sum_{j=1}^n (z_j-\bar{z})^2 \end{pmatrix} \\ =& n\bar{y}(1-\bar{y})\begin{pmatrix} \hat{\sigma}_x^2 & \hat{\sigma}_{xz} \\ \hat{\sigma}_{xz} & \hat{\sigma}_z^2 \end{pmatrix},\end{aligned}$$

where $\hat{\sigma}_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$, and $\hat{\sigma}_z^2 = \sum_{i=1}^{n} \frac{(z_i - \bar{z})^2}{n}$ are the sample variances for $x_i$ and $z_i$, and $\hat{\sigma}_{xz} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(z_i - \bar{z})}{n}$ is the sample covariance between $x_i$ and $z_i$. Based on formula (4.3), by Slutsky's theorem,

$$T = \mathbf{U}^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{U} \xrightarrow{d} \chi_2^2. \tag{4.6}$$

□

### 4.1.2 Three improved joint score tests

To improve on the robustness of the deviation in (4.2), we propose three improved joint score tests. In the first improved joint score test (denoted as iAW.Lev), we replace the within-group squared deviation by within-group absolute deviation Levene (1960):

$$z_i^L = \begin{cases} |x_i - \bar{x}_1|, \text{ if } Y_i = 1; \\ \\ |x_i - \bar{x}_0|, \text{ if } Y_i = 0. \end{cases} \tag{4.7}$$

For the logistic regression $logit(Pr(Y_i = 1)|x_i, z_i^L) = \beta_0^L + \beta_1^L x_i + \beta_1^L z_i^L$, under the null hypothesis $H_0^*$: $\beta_1^L = \beta_2^L = 0$, the joint score test statistic $T^{Lev}$ is asymptotically chi-squared distributed with two degrees of freedom:

$$T^{Lev} = (\mathbf{U}^{Lev})^T (\hat{\mathbf{\Sigma}}^{Lev})^{-1} \mathbf{U}^{Lev} \xrightarrow{d} \chi_2^2,$$

where $\mathbf{U}^{Lev} = \left(U_1, U_2^L\right)^T$, $U_2^L = \sum_{i=1}^n z_i^L \left(y_i - \bar{y}\right)$,

$$\hat{\mathbf{\Sigma}}^{Lev} = n\bar{y}\left(1-\bar{y}\right)\begin{pmatrix} \hat{\sigma}_x^2 & \hat{\sigma}_{xz^L} \\ \hat{\sigma}_{xz^L} & \hat{\sigma}_{z^L}^2 \end{pmatrix},$$

$\hat{\sigma}_{z^L}^2$ is the sample variance for $z_i^L$, and $\hat{\sigma}_{xz^L}$ is the sample covariance between $x_i$ and $z_i^L$. Note that the proposed improved joint test is different from Levene's test (Levene, 1960) in that Levene's test regards $z_i^L$ as random and uses ANOVA, while the proposed improved joint test regards $z_i^L$ as fixed (i.e., non-random) and uses a logistic regression framework.

In the second improved joint score test, we replace the sample means in the $T^{Lev}$ by sample medians Brown and Forsythe (1974):

$$z_i^{BF} = \begin{cases} |x_i - \tilde{x}_1|, \text{ if } Y_i = 1; \\ \\ |x_i - \tilde{x}_0|, \text{ if } Y_i = 0, \end{cases} \tag{4.8}$$

where $\tilde{x}_1$ and $\tilde{x}_0$ are the sample medians for cases and controls, respectively. Under the null hypothesis $H_0^{BF} : \beta_0^{BF} = \beta_1^{BF} = 0$, the joint score test statistic $T^{BF}$ follows asymptotically the chi-squared distribution with two degrees of freedom:

$$T^{BF} = (\mathbf{U}^{BF})^T (\hat{\mathbf{\Sigma}}^{BF})^{-1} \mathbf{U}^{BF} \xrightarrow{d} \chi_2^2,$$

where $\mathbf{U}^{BF} = \left(U_1, U_2^{BF}\right)^T$, $U_2^{BF} = \sum_{i=1}^n z_i^{BF} \left(y_i - \bar{y}\right)$,

$$\hat{\mathbf{\Sigma}}^{BF} = n\bar{y}\left(1-\bar{y}\right)\begin{pmatrix} \hat{\sigma}_x^2 & \hat{\sigma}_{xz^{BF}} \\ \hat{\sigma}_{xz^{BF}} & \hat{\sigma}_{z^{BF}}^2 \end{pmatrix},$$

$\hat{\sigma}^2_{z^{BF}}$ is the sample variance for $z_i^{BF}$, and $\hat{\sigma}_{xz^{BF}}$ is the sample covariance between $x_i$ and $z_i^{BF}$.

In the third improved joint score test, we replace the sample means in the $T^{Lev}$ by trimmed sample means Brown and Forsythe (1974):

$$z_i^{TM} = \begin{cases} |x_i - \check{x}_1|, \text{ if } Y_i = 1; \\ \\ |x_i - \check{x}_0|, \text{ if } Y_i = 0, \end{cases} \tag{4.9}$$

where $\check{x}_1$ and $\check{x}_0$ are the 25% trimmed sample means for cases and controls, respectively. The 25% trimmed mean for a sample is the sample mean after 25% from both the lowest and highest values are trimmed.

For the logistic regression model $logit(Pr(Y_i = 1)|x_i, z_i^{TM}) = \beta_0^{TM} + \beta_1^{TM} x_i + \beta_1^{TM} z_i^{TM}$, under the null hypothesis $H_0^{TM}$: $\beta_1^{TM} = \beta_2^{TM} = 0$, the joint score test statistic $T^{TM}$ is asymptotically chi-squared distributed with two degrees of freedom:

$$T^{TM} = (\mathbf{U}^{TM})^T (\hat{\mathbf{\Sigma}}^{TM})^{-1} \mathbf{U}^{TM} \xrightarrow{d} \chi^2_2,$$

where $\mathbf{U}^{TM} = \left(U_1, U_2^{TM}\right)^T$, $U_2^{TM} = \sum_{i=1}^n z_i^{TM} (y_i - \bar{y})$,

$$\hat{\mathbf{\Sigma}}^{TM} = n\bar{y}(1-\bar{y}) \begin{pmatrix} \hat{\sigma}^2_x & \hat{\sigma}_{xz^{TM}} \\ \hat{\sigma}_{xz^{TM}} & \hat{\sigma}^2_{z^{TM}} \end{pmatrix},$$

$\hat{\sigma}^2_{z^{TM}}$ is the sample variance for $z_i^{TM}$, and $\hat{\sigma}_{xz^{TM}}$ is the sample covariance between $x_i$ and $z_i^{TM}$.

## 4.2 Simulation studies

### 4.2.1 Simulation setting

We conducted comprehensive simulations to compare the performance of the three improved tests with the three existing methods: the joint likelihood ratio test based on the normal distribution (jointLRT) (Pearson and Neyman, 1930; Wilks, 1938), the Kolmogorov-Smirnov test (KS) (William, 1971), and AW test. Zhang et al. (2012) attained the mathematical expression and the exact distribution of jointLRT test statistics under normal distribution. Due to computational complexity, we used the asymptotic distribution of jointLRT in our simulation studies.

The simulation studies examined the following four aspects and their impacts on these six tests: (1) various sample sizes, (2) the presence of heterogeneity of means and variances, (3) the violation of the normality assumption, and (4) outliers. We considered various sample sizes: $(n_0, n_1)$=(100, 100), $(n_0, n_1)$=(50, 50), and $(n_0, n_1)$=(20, 20). Four parametric models were employed to generate the methylation data: the normal distribution, the Beta distribution, the chi-square distribution, and the mixture of two normal distributions. To evaluate the impact of outliers, we replaced the DNA methylation level of one randomly picked disease subject by $max\{x_{1,max},\ (Q_3 + 3(Q_3 - Q_1))\}$, where $x_{1,max}$ denotes the maximum DNA methy-

lation level of the diseased samples, and $Q_1$ and $Q_3$ are the first and third quartiles, respectively.

We computed the empirical Type I error rates and the power of the six tests under different scenarios: (1) Type I error scenario (eqM & eqV): distributions of non-diseased and diseased samples are the same; (2) Power scenario I (diffM & eqV): distributions of non-diseased and diseased samples are different in means only; (3) Power scenario II (eqM & diffV): distributions of non-diseased and diseased samples are different in variances only; and (4) Power scenario III (diffM & diffV): distributions of non-diseased and diseased samples are different in both means and variances. We conducted $10,000$ repetitions to estimate Type I error rates under scenario (1). For the simulated power, the results were based on 5,000 repetitions. We used the critical values of the observed test statistics under the null to determine the simulated power in order to make the power comparison fair.

### 4.2.2 Simulation results

Overall, the three improved joint score tests performed better than the other three methods when methylation levels contained outliers and had different variances between diseased and non-diseased samples. Furthermore, iAW.BF is the most robust in terms of power among all the scenarios. The KS test had conservative empirical

Type I error rates and the lowest power in many scenarios.

Table 4.1: The type I error ($\times 100$) and power ($\times 100$) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation $\beta$-values were generated from the normal distribution without or with an outlier. The numbers of non-diseased and diseased samples are (100, 100).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 4.9 | 3.5 | 5.0 | 5.0 | 5.2 | 5.4 |
| | No | 1 | 0.9 | 0.5 | 0.8 | 1.0 | 1.0 | 1.2 |
| | No | 0.5 | 0.4 | 0.3 | 0.4 | 0.5 | 0.5 | 0.6 |
| diffM&eqV | No | 5 | 97.3 | 94.8 | 97.2 | 97.3 | 97.1 | 96.8 |
| | No | 1 | 91.0 | 85.7 | 91.0 | 89.8 | 90.2 | 88.7 |
| | No | 0.5 | 87.5 | 76.7 | 87.2 | 85.0 | 85.3 | 83.9 |
| eqM&diffV | No | 5 | 89.6 | 18.1 | 87.3 | 84.2 | 82.5 | 82.1 |
| | No | 1 | 75.6 | 6.3 | 69.0 | 63.2 | 61.2 | 60.6 |
| | No | 0.5 | 69.0 | 2.7 | 59.0 | 54.0 | 51.7 | 51.8 |
| diffM&diffV | No | 5 | 84.4 | 57.0 | 83.3 | 80.3 | 79.8 | 78.9 |
| | No | 1 | 66.1 | 36.2 | 63.8 | 58.4 | 56.9 | 56.4 |
| | No | 0.5 | 58.9 | 24.9 | 55.3 | 49.0 | 47.7 | 47.8 |
| eqM&eqV | Yes | 5 | 12.1 | 3.9 | 3.5 | 5.0 | 5.1 | 5.1 |
| | Yes | 1 | 3.5 | 0.6 | 0.5 | 0.8 | 0.8 | 0.9 |
| | Yes | 0.5 | 2.0 | 0.4 | 0.2 | 0.5 | 0.4 | 0.5 |
| diffM&eqV | Yes | 5 | 96.1 | 95.2 | 98.4 | 98.2 | 98.2 | 98.4 |
| | Yes | 1 | 84.8 | 82.9 | 94.9 | 93.3 | 94.2 | 93.2 |
| | Yes | 0.5 | 79.5 | 78.0 | 91.6 | 89.3 | 90.1 | 88.0 |
| eqM&diffV | Yes | 5 | 46.8 | 16.8 | 55.0 | 69.1 | 67.7 | 67.8 |
| | Yes | 1 | 23.2 | 3.9 | 32.8 | 45.3 | 46.2 | 43.5 |
| | Yes | 0.5 | 17.7 | 2.2 | 24.3 | 35.7 | 35.7 | 32.7 |
| diffM&diffV | Yes | 5 | 57.0 | 59.5 | 77.7 | 78.8 | 79.7 | 78.4 |
| | Yes | 1 | 29.1 | 32.4 | 59.9 | 57.9 | 60.3 | 57.3 |
| | Yes | 0.5 | 22.0 | 26.3 | 50.1 | 47.9 | 50.0 | 46.4 |

When methylation levels were generated from normal distributions without outliers, all tests had their empirical Type I error rates close to the nominal levels,

except for KS (Table 4.1). For Power Scenarios I, II, and III, three improved joint score tests had similar performance, but slightly lower power for jointLRT and AW. When methylation values were from normal distributions with an outlier, the three improved joint score tests can maintain empirical Type I error rates well at all nominal levels. Whereas the empirical Type I error rates of jointLRT were inflated at all nominal levels, AW and KS had very conservative empirical Type I error rates at all levels (Table 4.1). For Power Scenarios I, II and III, the three improved tests had similar or greater power than AW. For Power Scenarios II and III (i.e., different variances), KS had poor estimated power despite the presence or absence of an outlier. Similar findings of KS were also observed in other parametric distributions (Tables 4.2 and 4.4).

Similar findings were also observed for the Beta distribution setting (Table 4.2). When the Beta distributions of two groups were different in variances (Power Scenarios II and III) and contained outliers, the three improved tests had significantly greater power than AW.

Table 4.2: The type I error (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation $\beta$-values were generated from the beta distribution without or with an outlier. The numbers of non-diseased and diseased samples are (100, 100).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 5.4 | 3.6 | 5.0 | 5.1 | 5.3 | 5.2 |
| | No | 1 | 1.1 | 0.6 | 0.9 | 1.0 | 1.1 | 0.9 |
| | No | 0.5 | 0.6 | 3.3 | 0.4 | 0.5 | 0.5 | 0.4 |
| diffM&eqV | No | 5 | 97.3 | 94.9 | 97.9 | 97.8 | 97.7 | 97.8 |
| | No | 1 | 90.8 | 82.9 | 92.5 | 91.0 | 91.7 | 92.3 |
| | No | 0.5 | 85.3 | 78.5 | 89.5 | 86.4 | 88.5 | 88.9 |
| eqM&diffV | No | 5 | 88.5 | 18.4 | 87.1 | 82.7 | 82.1 | 82.6 |
| | No | 1 | 71.0 | 4.0 | 68.1 | 61.1 | 60.9 | 63.4 |
| | No | 0.5 | 61.8 | 2.5 | 58.5 | 51.1 | 53.5 | 56.0 |
| diffM&diffV | No | 5 | 84.2 | 66.0 | 88.2 | 85.6 | 85.9 | 86.1 |
| | No | 1 | 62.8 | 37.6 | 71.4 | 65.1 | 66.1 | 69.0 |
| | No | 0.5 | 52.5 | 31.4 | 63.6 | 56.2 | 59.3 | 62.3 |
| eqM&eqV | Yes | 5 | 11.1 | 3.4 | 3.8 | 5.1 | 5.1 | 4.7 |
| | Yes | 1 | 2.9 | 0.6 | 0.5 | 1.1 | 0.9 | 0.9 |
| | Yes | 0.5 | 1.7 | 0.4 | 0.2 | 0.5 | 0.5 | 0.4 |
| diffM&eqV | Yes | 5 | 97.1 | 95.5 | 98.7 | 98.4 | 98.8 | 98.8 |
| | Yes | 1 | 88.3 | 84.7 | 95.4 | 93.3 | 94.8 | 94.3 |
| | Yes | 0.5 | 83.1 | 80.4 | 92.2 | 89.1 | 91.9 | 91.4 |
| eqM&diffV | Yes | 5 | 31.7 | 15.8 | 26.3 | 60.1 | 59.0 | 61.1 |
| | Yes | 1 | 12.1 | 3.7 | 9.4 | 32.5 | 31.7 | 33.8 |
| | Yes | 0.5 | 8.0 | 2.3 | 4.3 | 23.1 | 24.2 | 25.5 |
| diffM&diffV | Yes | 5 | 28.5 | 59.7 | 37.7 | 53.2 | 52.7 | 54.4 |
| | Yes | 1 | 8.9 | 33.5 | 19.7 | 24.8 | 26.2 | 27.0 |
| | Yes | 0.5 | 5.8 | 27.2 | 12.6 | 16.4 | 19.1 | 19.3 |

When methylation values were generated from a two-component normal mixture distribution without outliers, both iAW.BF and AW had appropriate empirical Type I error rates. However, iAW.Lev and iAW.TM had significantly inflated empirical

Type I error rates. Additionally, jointLRT and KS had conservative empirical Type I error rates. Under all Power Scenarios, iAW.BF had greater power than AW and jointLRT. When methylation values were from two-component normal mixture distributions with an outlier, iAW.BF had appropriate simulated Type I error rates at each level. Although iAW.Lev and iAW.TM had increased empirical Type I error rates, they are much smaller than those rates of jointLRT. The KS and AW tests had conservative empirical Type I error rates. All of the three improved tests had significantly greater power than AW under Power scenarios II (i.e., different variances only) and III (i.e., different means and different variances).

When methylation values were generated from a chi-squared distribution without outliers, iAW.BF, iAW.TM and AW kept empirical Type I error rates well, though iAW.Lev presented increased empirical Type I error rates. While jointLRT had inflated empirical Type I error rates, and KS had more conservative empirical Type I error rates. For Power scenarios II and III (i.e., different variances), iAW.BF and iAW.TM had significantly greater power than AW, and iAW.Lev had similar power to AW for three power scenarios. When methylation values were generated from a chi-squared distribution with an outlier, the performance of all tests were similar, with the exception of AW, where it had conservative empirical Type I error rates.

Table 4.3: The type I error (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation $\beta$-values were generated from the mixture of two normal distributions without or with an outlier. The numbers of non-diseased and diseased samples are (100, 100).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 2.0 | 3.6 | 4.8 | 9.5 | 5.2 | 11.4 |
| | No | 1 | 0.5 | 0.7 | 0.8 | 3.0 | 1.1 | 4.3 |
| | No | 0.5 | 0.2 | 0.4 | 0.4 | 1.8 | 0.6 | 2.8 |
| diffM&eqV | No | 5 | 98.8 | 100 | 97.5 | 99.0 | 99.3 | 98.4 |
| | No | 1 | 93.9 | 99.8 | 91.9 | 94.3 | 97.8 | 91.0 |
| | No | 0.5 | 90.6 | 99.7 | 87.8 | 90.2 | 96.4 | 84.6 |
| eqM&diffV | No | 5 | 35.0 | 98.1 | 54.5 | 88.9 | 59.2 | 75.8 |
| | No | 1 | 10.7 | 81.1 | 35.6 | 72.5 | 36.1 | 52.7 |
| | No | 0.5 | 6.2 | 72.5 | 28.7 | 63.9 | 28.9 | 42.5 |
| diffM&diffV | No | 5 | 47.3 | 99.6 | 56.1 | 89.8 | 80.8 | 82.5 |
| | No | 1 | 19.6 | 93.7 | 36.9 | 72.4 | 60.7 | 56.5 |
| | No | 0.5 | 14.0 | 91.0 | 29.7 | 62.6 | 52.6 | 44.8 |
| eqM&eqV | Yes | 5 | 24.3 | 3.4 | 2.3 | 5.5 | 5.2 | 6.7 |
| | Yes | 1 | 6.1 | 0.6 | 0.4 | 1.2 | 0.8 | 1.8 |
| | Yes | 0.5 | 3.4 | 0.4 | 0.2 | 0.7 | 0.4 | 1.0 |
| diffM&eqV | Yes | 5 | 90.8 | 100 | 98.9 | 99.5 | 99.9 | 99.5 |
| | Yes | 1 | 76.4 | 100 | 95.1 | 97.3 | 98.2 | 96.3 |
| | Yes | 0.5 | 69.0 | 100 | 91.6 | 95.3 | 97.3 | 93.9 |
| eqM&diffV | Yes | 5 | 0.4 | 97.5 | 13.0 | 81.9 | 48.1 | 71.5 |
| | Yes | 1 | 0 | 86.2 | 5.2 | 60.4 | 29.1 | 45.9 |
| | Yes | 0.5 | 0 | 71.5 | 3.3 | 51.0 | 23.0 | 37.3 |
| diffM&diffV | Yes | 5 | 10.0 | 99.5 | 44.8 | 88.7 | 80.0 | 85.0 |
| | Yes | 1 | 3.0 | 96.1 | 26.3 | 70.3 | 61.9 | 62.5 |
| | Yes | 0.5 | 1.8 | 91.3 | 18.7 | 60.7 | 54.2 | 52.5 |

From the results of the four tables, we found that iAW.BF could control empirical Type I error rates well and have similar or greater power than AW under all scenarios including the existence of outliers, skewed distributions, and mixtures of two normal distributions. For the scenarios of mixtures of two normal distributions, iAW.Lev and

iAW.TM can maintain empirical Type I error rates at proper levels and had similar or greater power than AW. In comparison, AW can maintain appropriate empirical Type I error rates for any parametric distributions as designed without outliers. When the methylation values were generated from a distribution with an outlier, AW tended to have conservative empirical Type I error rates and smaller estimated power. The jointLRT, on the other hand, only performed best for methylation values generated from normal distributions without outliers. KS can keep conservative empirical Type I error rates under all scenarios, and it had poor estimated power in many scenarios.

Simulation studies were also conducted when the sample size was moderate (50, 50) and small (20, 20). The results are provided in Appendix C. We observed that empirical Type I error rates increased and power decreased when the sample size decreased from 100 to 50 subjects per group. Furthermore, the three improved joint score tests still performed significantly better than AW under a moderate or small sample size.

Table 4.4: The type I error (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation $\beta$-values were generated from the chi-square distribution without or with an outlier. The numbers of non-diseased and diseased samples are (100, 100).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 14.4 | 3.7 | 4.8 | 5.9 | 5.2 | 4.7 |
| | No | 1 | 5.8 | 0.7 | 0.8 | 1.3 | 0.9 | 1.0 |
| | No | 0.5 | 4.1 | 0.4 | 0.4 | 0.7 | 0.5 | 0.5 |
| diffM&eqV | No | 5 | 90.5 | 99.7 | 99.8 | 99.7 | 99.9 | 99.9 |
| | No | 1 | 61.1 | 96.8 | 99.0 | 97.0 | 99.5 | 99.4 |
| | No | 0.5 | 46.2 | 95.3 | 98.1 | 94.9 | 99.0 | 98.9 |
| eqM&diffV | No | 5 | 20.6 | 10.1 | 27.4 | 32.0 | 35.9 | 35.4 |
| | No | 1 | 7.8 | 2.0 | 11.4 | 12.0 | 17.0 | 15.6 |
| | No | 0.5 | 4.9 | 1.3 | 6.8 | 8.3 | 11.0 | 10.3 |
| diffM&diffV | No | 5 | 20.7 | 44.4 | 61.3 | 58.1 | 72.8 | 68.7 |
| | No | 1 | 5.3 | 18.5 | 39.4 | 29.3 | 49.3 | 43.3 |
| | No | 0.5 | 2.9 | 14.1 | 30.5 | 21.2 | 39.3 | 33.9 |
| eqM&eqV | Yes | 5 | 20.4 | 4.0 | 4.5 | 6.9 | 5.2 | 4.9 |
| | Yes | 1 | 10.3 | 0.6 | 0.6 | 1.6 | 1.0 | 0.9 |
| | Yes | 0.5 | 7.8 | 0.4 | 0.3 | 0.9 | 0.5 | 0.4 |
| diffM&eqV | Yes | 5 | 71.4 | 99.6 | 99.8 | 99.5 | 99.9 | 99.9 |
| | Yes | 1 | 24.6 | 96.3 | 99.3 | 95.7 | 99.4 | 99.3 |
| | Yes | 0.5 | 13.4 | 94.6 | 98.7 | 92.5 | 98.8 | 98.6 |
| eqM&diffV | Yes | 5 | 29.6 | 9.4 | 34.9 | 38.6 | 43.0 | 41.4 |
| | Yes | 1 | 10.8 | 1.8 | 13.2 | 16.6 | 20.5 | 20.1 |
| | Yes | 0.5 | 6.6 | 1.2 | 7.7 | 10.7 | 14.2 | 14.3 |
| diffM&diffV | Yes | 5 | 23.4 | 41.3 | 68.7 | 61.0 | 74.9 | 71.3 |
| | Yes | 1 | 6.5 | 16.2 | 45.5 | 32.5 | 51.7 | 48.2 |
| | Yes | 0.5 | 3.2 | 12.0 | 35.6 | 23.3 | 42.0 | 38.4 |

## 4.3 Real Data Analysis

### 4.3.1 Data description

We applied all six statistical tests to three publicly available DNA methylation data sets: GSE37020 (Teschendorff and Widschwendter, 2012), GSE20080 (Teschendorff et al., 2010), and GSE107080 (Zhang et al., 2017)) from Gene Expression Omnibus (GEO)(*www.ncbi.nlm.nih.gov/geo*).

The two HM27k data sets GSE37020 and GSE20080 have been described in Section 3.3.1. The procedures of quality control and preprocessing about the two data sets are presented in Appendix B. We used clean GSE37020 as the discovery set and clean GSE20080 as the validation set to detect CpG sites differentially methylated (DM) or differentially variable (DV) between CIN2+ samples and normal samples. For a given CpG site in a given data set, we applied each of the six joint tests to test for equalities of both means and variances. For a given joint test, we claimed a CpG site in the analysis of GSE37020 as significant methylation candidate (different in means or variances) if the false discovery rate (FDR) (Benjamini and Hochberg, 1995) adjusted p-value for the CpG site is less than 0.05. The function *p.adjust* in the statistical software $R$ was used to calculate the FDR-adjusted p-value. For a significant site in the analysis of GSE37020, if the corresponding un-adjusted p-value

in the analysis of GSE20080 is less than 0.05 and the difference in the directions of means and variances are consistent between the two data sets, then we claim that the significance in the analysis of GSE37020 is truly validated in the analysis of GSE20080. We use the differences of medians and mean absolute deviations between cases and controls to evaluate the directions.

GSE107080 contained DNA methylation profiles of about 850K sites measured from whole blood samples using Illumina Infinium MethylationEPIC (EPIC) platform. GSE107080 included 100 individuals with illicit drug injection and hepatitis C type virus (IDU+/HCV+) and 305 individuals without illicit drug injection and hepatitis C type virus (IDU-/HCV-). All the individuals were recruited from a well-established longitudinal cohort, Veteran Aging Cohort Study. The procedures of quality control and preprocessing for GSE107080 are presented in Appendix B.

For dataset GSE107080, the samples were randomly split into two sets of approximate equal size (due to odd numbers of cases and controls) as the training set and the validation set. The training set contained 148 controls (IDU-/HCV-) and 48 cases (IDU+/HCV+), and the validation set contained 147 controls and 47 cases. We used a similar method as the above to determine if the significance of a CpG site is truly validated.

### 4.3.2 Results

For GSE37020, the numbers of significant CpG sites (i.e., CpG sites with FDR-adjusted p-value $< 0.05$) obtained by the six joint tests are 4556 (jointLRT), 1288 (KS), 1850 (AW), 2041 (iAW.Lev), 1843 (iAW.BF), and 1838 (iAW.TM). The truly validated CpG sites are 1705 (jointLRT), 47 (KS), 220 (AW), 666 (iAW.Lev), 296 (iAW.BF), and 342 (iAW.TM).

Table 4.5: The performance of six joint tests based on HM27k data sets GSE37020 and GSE20080.

| Test | nSig | nValidation | nTV | pTV(%) | nFV | pFV(%) |
|---|---|---|---|---|---|---|
| JointLRT | 4556 | 2213 | 1705 | 77.0 | 508 | 23.0 |
| KS | 1288 | 60 | 47 | 78.3 | 13 | 21.7 |
| AW | 1850 | 262 | 220 | 84.0 | 42 | 16.0 |
| iAW.Lev | 2041 | 747 | 666 | 89.2 | 81 | 10.8 |
| iAW.BF | 1843 | 339 | 296 | 87.3 | 43 | 12.7 |
| iAW.TM | 1838 | 387 | 342 | 88.4 | 45 | 11.6 |

nSig : the number of significant CpG sites detected in GSE37020 (adjusted p-value $< 0.05$);
nValidation : the number of validated CpG sites in GSE20080 (unadjusted p-value $< 0.05$);
nTV : the number of truly validated CpG sites with the same difference directions in means and variances between two samples;
pTV : $= \frac{nTV}{nValidation}$, the proportion of significant CpG sites detected in GSE37020 and truly validated in GSE20080;
nFV : the number of falsely validated CpG sites in GSE20080 with inconsistent difference direction in means or variances between two samples;
pFV : $= \frac{nFV}{nValidation}$, the proportion of significant CpG sites detected in GSE37020 but falsely validated in GSE20080.

Table 4.5 presents the numbers/proportions of truly and falsely validated significant CpG sites. The three improved joint score tests have higher true validation ratios than joint LRT, KS, and AW. Among all the tests, iAW.Lev had the highest true validation rate (89.2%) and the lowest false validation rate (10.8%), followed by iAW.TM and iAW.BF.

Figure 4.1 shows the parallel boxplots of DNA methylation levels versus case-control status for the top CpG site (i.e., having the smallest p-value among those truly validated CpG sites for testing the homogeneity of means and variances simultaneously) obtained by each of the six joint tests. All these top CpG sites were validated in GSE20080. It has been found that the high incidence of cervical lesions is associated with the genes ST6GALNAC3, CRB1 and RGS7, where cg26363196 (jointLRT), cg00321478 (AW) and cg21303386 (iAW.Lev) might reside (Farkas et al., 2013; Kudela et al., 2016). Furthermore, the gene PRRG2, where cg2196766 (KS) might reside, is involved in the signal transduction pathway, which might be a novel biomarker for CIN2+ diagnosis (Yazicioglu et al., 2013). The gene FPRL2, where cg06784466 (iAW.BF, iAW.TM) might reside, is related to innate immunity and host defense mechanisms (Devosse et al., 2009).

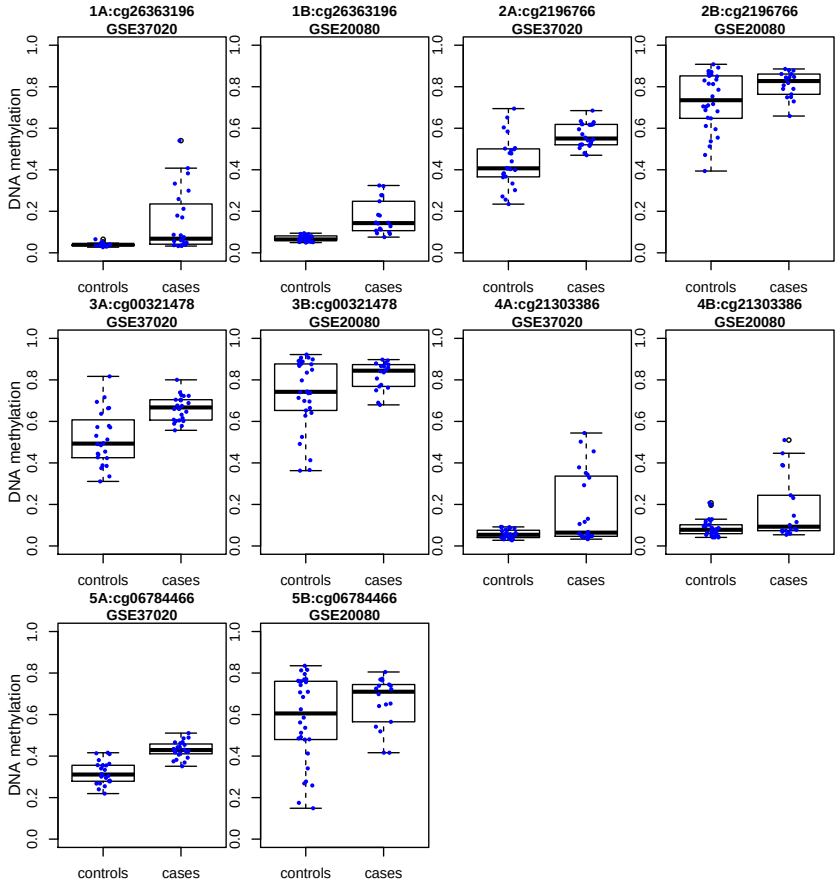


Figure 4.1: Paired parallel boxplots of DNA methylation levels (y-axis) versus case-control status (x-axis) for the 5 unique top CpG sites acquired by the six joint tests based on HM27k data sets. The dots indicate subjects.1A and 1B are for cg26363196 (jointLRT). 2A and 2B are for cg2196766 (KS). 3A and 3B are for cg00321478 (AW). 4A and 4B are for cg21303386 (iAW.Lev). 5A and 5B are for cg06784466 (iAW.BF, iAW.TM). 1A, 2A, 3A, 4A, 5A are based on GSE37020. 1B, 2B, 3B, 4B, 5B are based on GSE20080.

For GSE107080, the number of significant CpG sites (i.e., CpG sites with FDR-adjusted p-value $< 0.05$) obtained by the six joint tests in the training set are $51,994$ (jointLRT), 10 (KS), 12 (AW), 709 (iAW.Lev), 22 (iAW.BF), and 22 (iAW.TM). The corresponding numbers of validated CpG sites in the validation set (i.e., CpG sites with unadjusted p-value $< 0.05$) are $19,806$ (jointLRT), 3 (KS), 5 (AW), 201 (iAW.Lev), 7 (iAW.BF), and 9 (iAW.TM). After checking the direction of differences, the truly validated CpG sites are 5652 (jointLRT), 1 (KS), 2 (AW), 89 (iAW.Lev), 4 (iAW.BF), and 5 (iAW.TM).

Table 4.6 presents the numbers/proportions of truly and falsely validated significant CpG sites based on GSE107080. The three improved tests have higher true validation rates than joint LRT, KS and AW tests. Amongst them, iAW.BF and iAW.TM have a more than ten percent higher proportion of true validation than AW.

Table 4.6: The performance of six joint tests based on EPIC data GSE107080.

| Test | nSig | nValidation | nTV | pTV(%) | nFV | pFV(%) |
|---|---|---|---|---|---|---|
| JointLRT | 51994 | 19806 | 5652 | 28.5 | 14154 | 71.5 |
| KS | 10 | 3 | 1 | 33.3 | 2 | 66.7 |
| AW | 12 | 5 | 2 | 40.0 | 3 | 60.0 |
| iAW.Lev | 709 | 201 | 89 | 44.3 | 112 | 55.7 |
| iAW.BF | 22 | 7 | 4 | 57.1 | 3 | 42.9 |
| iAW.TM | 22 | 9 | 5 | 55.6 | 4 | 44.4 |

$^{\text{nSig}}$ : the number of significant CpG sites detected in the training set of GSE107080 (adjusted p-value $< 0.05$);

$^{\text{nValidation}}$ : the number of validated CpG sites in the validation set of GSE107080 (unadjusted p-value $< 0.05$);

$^{\text{nTV}}$ : the number of truly validated CpG sites with the same difference directions in means and variances between two samples;

$^{\text{pTV}}$ : $= \frac{nTV}{nValidation}$, the proportion of significant CpG sites detected in the training set and truly validated in the validation set;

$^{\text{nFV}}$ : the number of falsely validated CpG sites in the validation set with inconsistent difference direction in means or variances between two samples;

$^{\text{pFV}}$ : $= \frac{nFV}{nValidation}$, the proportion of significant CpG sites detected in the training set but falsely validated in the validation set.

## 4.4 Discussion

The three improved joint score tests are derived from the generalized linear model framework as AW, thus they maintain the strengths of AW in terms of efficiency. Furthermore, the three improved tests use an absolute deviation instead of a squared deviation used by AW to enhance the robustness. For skewed methylation distribu-

tions or distributions with outliers, squared deviation used by AW can be enormously affected by extreme values, which can lead to erroneous results, thus AW tends to have conservative empirical Type I error rates and smaller power in some scenarios. Our proposed methods rectify this problem and thus can maintain good power even if the distribution is skewed or contains outliers. Furthermore, when compared to the squared deviation, the absolute deviation retains the same magnitude of the original measurement scales and are consequently more interpretable. The iAW.Lev tends to have inflated empirical Type I error rates under skewed and mixture distributions. In comparison, iAW.BF and iAW.TM employ a median or trimmed mean as the central tendency to calculate absolute deviation. Both of them are robust and can minimize the impact of outliers and skewed distributions in evaluating the overall dispersion of the sample data.

The performance of the jointLRT is highly dependent on the validity of normality assumptions. However, the empirical distribution of methylation data are often skewed or contain outlying observations. The KS test is inclined to have conservative empirical Type I error rates and lowest power under many scenarios. Therefore, it might not be suitable for DNA methylation analysis as expected.

We would like to address one limitation in our simulation studies. Since the analytical form of the underlying probability distribution of methylation data is

rarely known, we have applied various settings in an attempt to mimic the reality. We also try to evaluate our methods in four different aspects. However, our simulation study might not cover all the possible cases that one might encounter in reality. Nevertheless, the results from real data analyses provide strong evidence to support the thesis that our proposed tests are in general more robust in comparison with the AW test.

# 5 Model-based clustering for detecting differentially variable microRNAs

Besides DNA methylation, microRNA (miRNA) regulation is another important epigenetic mechanism. Aberrant miRNA expression pattern has been linked to many diseases (Lu et al., 2005; Cullen, 2011; Fernández-Hernando et al., 2013). In addition to differential mean expression, differentially variable expression levels can also significantly affect gene regulation. Moreover, the correlations between different miRNAs can help us advance our understanding of the mechanism of genetic disorders and find the truly important candidate of miRNAs for diagnoses and therapy.

In this chapter, we focus on better identifying miRNA probes based on their differential variances by employing the framework of Qiu et al. (2008) and modifying the marginal model to detect differentially variable (DV) miRNAs. We impose special structures on covariance matrices for each cluster of miRNAs based on prior information about the relationship between variances in cases and controls.

## 5.1 Framework

We assume that the miRNA expression data have been normalized to remove the confounding effects and transformed so that the distribution of the miRNA expression levels is close to a normal distribution. Based on Qiu et al. (2008), for a given miRNA, we denote $X_i$ as the processed expression level for the $i$th subject, $i = 1, \ldots, m$, where $m = m_c + m_n$, $m_c$ is the number of diseased samples (cases) and $m_n$ is the number of non-diseased samples (controls). Let $\boldsymbol{X} = (X_1, \ldots, X_{m_c}, X_{m_c+1}, \ldots, X_{m_c+m_n})^T$ denote the processed expression profiles over $m$ samples and follow a three-component mixture of multivariate normal distributions with marginal density:

$$\begin{aligned} &\pi_1 f_1(\boldsymbol{x}; \boldsymbol{\theta}_1) + \pi_2 f_2(\boldsymbol{x}; \boldsymbol{\theta}_2) + \pi_3 f_3(\boldsymbol{x}; \boldsymbol{\theta}_3), \\ &\pi_1 + \pi_2 + \pi_3 = 1, \pi_k > 0, k = 1, 2, 3, \end{aligned} \tag{5.1}$$

where $\pi_1, \pi_2, \pi_3$ are mixing proportions. The $m \times 1$ vector $\boldsymbol{x}$ is a realization of the random vector $\boldsymbol{X}$; $\boldsymbol{\theta}_k$, is the parameter set for the $k$-th component distribution $f_k$, $k = 1, 2, 3$; and $f_1$, $f_2$, and $f_3$ are the density functions for multivariate normal distributions with the mean vectors

$$\boldsymbol{\mu}_1 = \begin{pmatrix} \mu_{1c}\mathbf{1}_{m_c} \\ \mu_{1n}\mathbf{1}_{m_n} \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} \mu_{2c}\mathbf{1}_{m_c} \\ \mu_{2n}\mathbf{1}_{m_n} \end{pmatrix}, \quad \boldsymbol{\mu}_3 = \begin{pmatrix} \mu_{3c}\mathbf{1}_{m_c} \\ \mu_{3n}\mathbf{1}_{m_n} \end{pmatrix}, \tag{5.2}$$

and covariance matrices

$$\begin{aligned}\boldsymbol{\Sigma}_1 &= \begin{pmatrix} \sigma_{1c}^2 \boldsymbol{R}_{1c} & \mathbf{0} \\ \mathbf{0} & \sigma_{1n}^2 \boldsymbol{R}_{1n} \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma_2^2 \boldsymbol{R}_{2c} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \boldsymbol{R}_{2n} \end{pmatrix}, \\ \boldsymbol{\Sigma}_3 &= \begin{pmatrix} \sigma_{3c}^2 \boldsymbol{R}_{3c} & \mathbf{0} \\ \mathbf{0} & \sigma_{3n}^2 \boldsymbol{R}_{3n} \end{pmatrix},\end{aligned} \tag{5.3}$$

respectively, where the correlation matrices are

$$\boldsymbol{R}_t = (1-\rho_t)\left[\boldsymbol{I}_{n_t} + \frac{\rho_t}{(1-\rho_t)}\mathbf{1}_{n_t}\mathbf{1}_{n_t}^T\right], \tag{5.4}$$

$t = 1c, 1n, 2c, 2n, 3c,$ or $3n$. Note that $n_t = m_c$ if $t = 1c, 2c,$ or $3c$; $n_t = m_n$ if $t = 1n, 2n,$ or $3n$. Without loss of generality, we assume the first $m_c$ elements are for the diseased samples (cases) and the remaining $m_n$ elements are for the non-diseased samples (controls). Let $\boldsymbol{\theta}_1 = (\mu_{1c}, \sigma_{1c}^2, \rho_{1c}, \mu_{1n}, \sigma_{1n}^2, \rho_{1n})^T$, $\boldsymbol{\theta}_2 = (\mu_{2c}, \sigma_2^2, \rho_{2c}, \mu_{2n}, \rho_{2n})^T$, $\boldsymbol{\theta}_3 = (\mu_{3c}, \sigma_{3c}^2, \rho_{3c}, \mu_{3n}, \sigma_{3n}^2, \rho_{3n})^T$.

We assume that the available miRNAs belong to one and only one of the following clusters: (1) $\sigma_{1c}^2 > \sigma_{1n}^2$, miRNAs having higher variances in cases than in controls (denoted as the OV cluster), (2) $\sigma_{2c}^2 = \sigma_{2n}^2 = \sigma_2^2$, miRNAs having equal variances between cases and controls (denoted as the EV cluster), (3) $\sigma_{3c}^2 < \sigma_{3n}^2$, miRNAs having smaller variances in cases than in controls (denoted as the UV cluster). We allow the means and correlations to be different between cases and controls in the EV cluster.

Based on the characteristics of miRNA expression profiles, Model 5.1 makes some assumptions to capture the structural information of miRNA expression data: (a)

miRNA expression profiles in the same cluster have the same marginal distribution; (b) miRNA expression profiles are marginally independent; (c) marginal means and variances of expression levels for a given miRNA from the same type of samples (cases or controls) are the same; (d) marginal correlations between any pair of expression levels for a given miRNA from the same type of samples are the same; (e) marginal correlations between any pair of expression levels for a given miRNA from different types of samples are zero.

Next, we derive a selection method for miRNA candidate based on Model 5.1. The $j$-th miRNA is assigned to one of the three clusters based on its posterior probability

$$\begin{aligned} Pr(\text{gene } j \in \text{ cluster } k|\boldsymbol{x}_j) =& \frac{\pi_k f_k(\boldsymbol{x}_j;\boldsymbol{\theta}_k)}{\pi_1 f_1(\boldsymbol{x}_j;\boldsymbol{\theta}_1)+\pi_2 f_2(\boldsymbol{x}_j;\boldsymbol{\theta}_2)+\pi_3 f_3(\boldsymbol{x}_j;\boldsymbol{\theta}_3)}, \\ k =& 1,2,3, \end{aligned} \tag{5.5}$$

where $\boldsymbol{x}_j$ is the processed profile for the $j$-th miRNA, $j = 1, \ldots, p$. It means that gene $j$ with profile $\boldsymbol{x}_j$ will be classified to cluster $C_{k_0}$ if the posterior probability that the $j$-th miRNA belongs to cluster $C_{k_0}$ given $\boldsymbol{x}_j$ is the largest, i.e.,

$$k_0 = \arg\max_k Pr(\text{gene } j \in C_k|\boldsymbol{x}_j). \tag{5.6}$$

## 5.2 Parameter estimation

Based on Titterington et al. (1985), the complete data can be represented as

$$\{\boldsymbol{y}_j, j = 1, \ldots, p\} = \{(\boldsymbol{x}_j^T, \boldsymbol{z}_j^T)^T, j = 1, \ldots, p\},$$

where $p$ is the number of miRNAs, $\boldsymbol{x}_j$ is an $m \times 1$ vector, $m = m_c + m_n$, $m_c$ is the number of diseased samples, $m_n$ is the number of non-diseased samples, $\boldsymbol{z}_j = (z_{1j}, z_{2j}, 1 - z_{1j} - z_{2j})$ and

$$z_{kj} = \begin{cases} 1 & \text{if } \boldsymbol{x}_j \text{ is in category } k \\ 0 & \text{otherwise} \end{cases}, \quad j = 1, \ldots, p, \quad k = 1, 2.$$

Then the likelihood corresponding to $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p)$ can be written as

$$\begin{aligned} g(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p | \boldsymbol{\Psi}) &= \prod_{j=1}^{p} f(\boldsymbol{x}_j, \boldsymbol{z}_j) \\ &= \prod_{j=1}^{p} f(\boldsymbol{x}_j | \boldsymbol{z}_j) f(\boldsymbol{z}_j) \\ &= \prod_{j=1}^{p} \left\{ \left[ f_1(\boldsymbol{x}_j)^{z_{1j}} f_2(\boldsymbol{x}_j)^{z_{2j}} f_3(\boldsymbol{x}_j)^{(1-z_{1j}-z_{2j})} \right] \left[ \pi_1^{z_{1j}} \pi_2^{z_{2j}} (1 - \pi_1 - \pi_2)^{(1-z_{1j}-z_{2j})} \right] \right\}, \end{aligned}$$

where

$$\boldsymbol{Z}_i \sim \text{Multinomial}\left(1, \pi_1, \pi_2, 1 - \pi_1 - \pi_2\right), \tag{5.7}$$

and

$$\begin{aligned} f(\boldsymbol{z}_i) &= \begin{cases} \pi_1 & \text{if } z_{1j} = 1 \text{ and } z_{2j} = 0, \\ \pi_2 & \text{if } z_{1j} = 0 \text{ and } z_{2j} = 1, \\ 1 - \pi_1 - \pi_2 & \text{if } z_{1j} = z_{2j} = 0, \end{cases} \qquad (5.8) \\ &= \pi_1^{z_{1j}} \pi_2^{z_{2j}} (1 - \pi_1 - \pi_2)^{(1-z_{1j}-z_{2j})}, \end{aligned}$$

and

$$\begin{aligned} f(\boldsymbol{x}_j | \boldsymbol{z}_j) &= \begin{cases} f_1(\boldsymbol{x}_j) & \text{if } z_{1j} = 1 \text{ and } z_{2j} = 0, \\ f_2(\boldsymbol{x}_j) & \text{if } z_{1j} = 0 \text{ and } z_{2j} = 1, \\ f_3(\boldsymbol{x}_j) & \text{if } z_{1j} = z_{2j} = 0, \end{cases} \qquad (5.9) \\ &= f_1(\boldsymbol{x}_j)^{z_{1j}} f_2(\boldsymbol{x}_j)^{z_{2j}} f_3(\boldsymbol{x}_j)^{(1-z_{1j}-z_{2j})}. \end{aligned}$$

The log complete likelihood function is

$$\ell_0(\boldsymbol{\Psi}) = \sum_{j=1}^{p} \boldsymbol{z}_j^T \boldsymbol{V}(\boldsymbol{\pi}) + \sum_{j=1}^{p} \boldsymbol{z}_j^T \boldsymbol{U}_j(\boldsymbol{\theta}),$$

where

$$\boldsymbol{z}_j = \begin{pmatrix} z_{1j} \\ z_{2j} \\ 1 - z_{1j} - z_{2j} \end{pmatrix}, \quad \boldsymbol{V}(\boldsymbol{\pi}) = \begin{pmatrix} \log(\pi_1) \\ \log(\pi_2) \\ \log(1 - \pi_1 - \pi_2) \end{pmatrix}, \quad \boldsymbol{U}_j(\boldsymbol{\theta}) = \begin{pmatrix} \log\left(f_1(\boldsymbol{x}_j; \boldsymbol{\theta}_1)\right) \\ \log\left(f_2(\boldsymbol{x}_j; \boldsymbol{\theta}_2)\right) \\ \log\left(f_3(\boldsymbol{x}_j; \boldsymbol{\theta}_3)\right) \end{pmatrix},$$

and

$$\boldsymbol{\theta}_1 = \left(\mu_{1c}, \sigma_{1c}^2, \rho_{1c}, \mu_{1n}, \sigma_{1n}^2, \rho_{1n}\right)^T,$$

$$\boldsymbol{\theta}_2 = \left(\mu_{2c}, \sigma_{2}^2, \rho_{2c}, \mu_{2n}, \rho_{2n}\right)^T,$$

$$\boldsymbol{\theta}_3 = \left(\mu_{3c}, \sigma_{3c}^2, \rho_{3c}, \mu_{3n}, \sigma_{3n}^2, \rho_{3n}\right)^T,$$

and

$$\boldsymbol{\Psi} = \left(\pi_1, \pi_2, \mu_{1c}, \sigma_{1c}^2, \rho_{1c}, \mu_{1n}, \sigma_{1n}^2, \rho_{1n}, \mu_{2c}, \sigma_{2}^2, \rho_{2c}, \mu_{2n}, \rho_{2n}, \mu_{3c}, \sigma_{3c}^2, \rho_{3c}, \mu_{3n}, \sigma_{3n}^2, \rho_{3n}\right)^T.$$

From some initial approximation, the EM algorithm generates $\boldsymbol{\Psi}^{(0)}$, a sequence $\{\boldsymbol{\Psi}^{(t)}\}$ of estimates. Each iteration consists of the following two steps:

**E-step:** Evaluate $Q\left(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(t)}\right) = \mathrm{E}\left[\log(g(\boldsymbol{y}|\boldsymbol{\Psi})|\boldsymbol{x}, \boldsymbol{\Psi}^{(t)}\right]$.

**M-step:** Find $\boldsymbol{\Psi} = \boldsymbol{\Psi}^{(t+1)}$ to maximize $Q\left(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(t)}\right)$.

The procedures of parameter estimation are presented in Appendix D.

## 5.3 Simulation studies

### 5.3.1 Simulation setting

We conducted four sets of simulation studies. In the first set (denoted as SimI), we generated miRNA data from the proposed marginal mixture model of multivariate normal, where estimated model parameters for GSE67139 (i.e., the discovery set) are used as the true values of the model parameters ($\pi_1 = 0.31, \pi_2 = 0.58, \pi_3 = 0.11, \mu_{1c} = -0.14, \sigma^2_{1c} = 1.49, \rho_{1c} = 0.08, \mu_{1n} = 0.14, \sigma^2_{1n} = 0.45, \rho_{1n} = 0.32, \mu_{2c} = 0.03, \sigma^2_{2c} = 1.01, \rho_{2c} = 0.04, \mu_{2n} = -0.03, \sigma^2_{2n} = 1.01, \rho_{2n} = 0.11, \mu_{3c} = 0.13, \sigma^2_{3c} = 0.28, \rho_{3c} = 0.04, \mu_{3n} = -0.13, \sigma^2_{3n} = 1.69, \rho_{1n} = -0.01$). We generated 100 data sets, each of which has 1,000 miRNAs for 50 cases and 50 controls. The numbers of miRNAs in each cluster are 310 (OV), 580 (EV) and 110 (UV).

In the second set (denoted as SimII), we generated miRNA data from a mixture of three-component multivariate t distribution with the same mean vectors and covariance matrices as those in SimI and with three degrees of freedom. SimII is used to evaluate the performance of the proposed method when the normality assumption for any one of the three clusters (OV, EV, and UV) is violated.

In the third set (denoted as SimIII) of the simulation studies, we generated miRNA data from the same model as that in SimI, except that the marginal corre-

lations within subject-groups were set to zero ($\rho_{kc} = 0$ and $\rho_{kn}$=0). SimIII is used to evaluate the performance of the proposed method when there are no marginal correlations.

In the fourth set (denoted as SimIV) of the simulation studies, we generated miRNA data from the same model as that in SimII, except that the marginal correlations within subject-groups were set to zero ($\rho_{kc}$=0 and $\rho_{kn} = 0$). SimIV is used to evaluate the performance of the proposed method when there are no marginal correlations and when the normality assumption for any one of the three clusters (OV, EV, and UV) is violated.

We compared the proposed method (denoted as gs) with the six methods which are based on Bar et al.'s (2014) N3 model, and Bar and Schifano's (2018) L2N model. Both N3 and L2N models have been implemented in R package DVX (Bar and Schifano, 2018). For both N3 and L2N, DVX outputs raw p-values, q-values, and posterior probabilities $p_{gk}$ that the probe g belongs to cluster $k$ given its expression profile and estimated model parameters, $k = 1, 2, 3$. Hence, for both N3 and L2N, we used three methods to assign probes to two clusters: DV probes and non-DV probes. The first method is based on the $q$-value. If a miRNA has a $q$-value $< 0.05$, we assign that to be differentially variable; and non-differentially variable otherwise. The second method is based on the false-discovery-rate (FDR) adjusted p-value. If

a miRNA has a FDR-adjusted $p$-value $< 0.05$, we assign that to be differentially variable; and non-differentially variable otherwise. The third method is based on the posterior probabilities. We assign a miRNA to cluster k* if the posterior probability $p_{gk^*}$ is the largest among the three posterior probabilities, $p_{g1}$, $p_{g2}$, and $p_{g3}$. We denote the three miRNA-assignment methods as N3.q (L2N.q), N3.f (L2N.f), and N3 (L2N), respectively.

For simulated datasets, we calculated the magnitude of agreement between the true cluster memberships of miRNAs and the detected cluster memberships by each of the seven methods using the Jaccard index (Jaccard, 1912; Qiu et al., 2008). The maximum value of the Jaccard index is one, indicating perfect agreement. The minimum value of the Jaccard index is zero, indicating that the agreement is by chance. The definition of Jaccard index is presented as follows.

Let $A$ and $B$ are two sets, each with $p$ binary attributes, $A = \{A_j \in A | A_j = 0 \text{ or } 1, j = 1, \ldots, p\}$, $B = \{B_j \in B | B_j = 0 \text{ or } 1, j = 1, \ldots, p\}$. The total numbers of each combination of attributes for both $A$ and $B$ are specified as follows:

$$M_{11} = \sum_{j=1}^{p} I(A_j = 1 \text{ and } B_j = 1), \quad M_{01} = \sum_{j=1}^{p} I(A_j = 0 \text{ and } B_j = 1),$$
$$M_{10} = \sum_{j=1}^{p} I(A_j = 1 \text{ and } B_j = 0), \quad M_{00} = \sum_{j=1}^{p} I(A_j = 0 \text{ and } B_j = 0),$$

where $I(x)$ is an indicator function, with value as 0 or 1. The Jaccard index is given

as

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}. \tag{5.10}$$

We also evaluate the performance using the false positive rate (FPR) (i.e., the proportion of detected DV probes among the true non-DV probes) and the false negative rate (FNR) (i.e., the proportion of detected non-DV probes among the true DV probes). The smaller the FPR (FNR), the better the performance.

### 5.3.2 Simulation results

When data were generated from a mixture of multivariate normal distributions (SimI and SimIII), the values of the Jaccard index obtained by the gs method were close to one (the perfect agreement). Under the assumption of a mixture of multivariate normal distributions, the gs method performed better than the other six methods in terms of larger Jaccard index, smaller FPR and smaller FNR (Figure 5.1 and 5.2). The values of FPR and FNR obtained by the gs method were significantly smaller than those by the other six methods. The results of two-sided Wilcoxon signed-rank tests of Jaccard index, FPR and FNR are shown in Appendix E.

When data were generated from a mixture of multivariate t distributions (SimII and SimIV), the gs method had a smaller Jaccard index value than the other six methods (Figure 5.1 and 5.2). It means that when the simulated miRNA was from

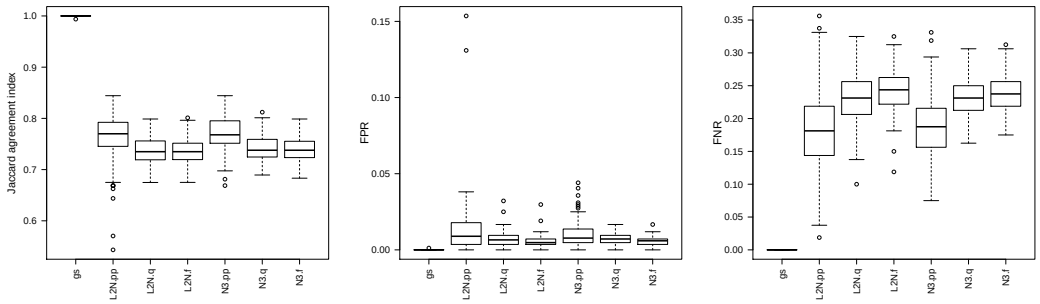


Figure 5.1: The boxplots of estimated Jaccard indices, FPR, and FNR based on the 100 simulated datasets in SimI. The closer to one the Jaccard index, the better the performance of the method. The closer to zero the FPR (FNR), the better the performance of the method.

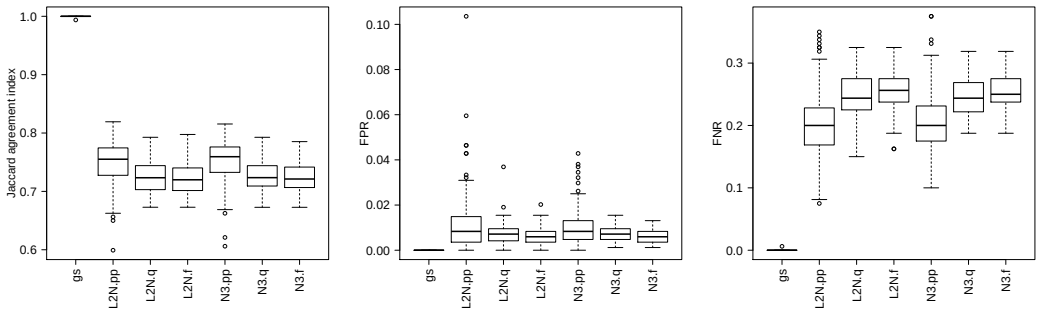


Figure 5.2: The boxplots of estimated Jaccard indices, FPR, and FNR based on the 100 simulated datasets in SimIII. The closer to one the Jaccard index, the better the performance of the method. The closer to zero the FPR (FNR), the better the performance of the method.

the multivariate non-normal distribution, the gs method has a higher probability of classifying them into the wrong sets (DV and non-DV). The gs method had a smaller FNR but larger FPR values than the other six methods (Figure 5.1 and 5.2). This result shows that when the underlying assumption that the marginal miRNA expression levels are from multivariate normal distributions is violated, the gs method may uncover more false miRNA candidates .

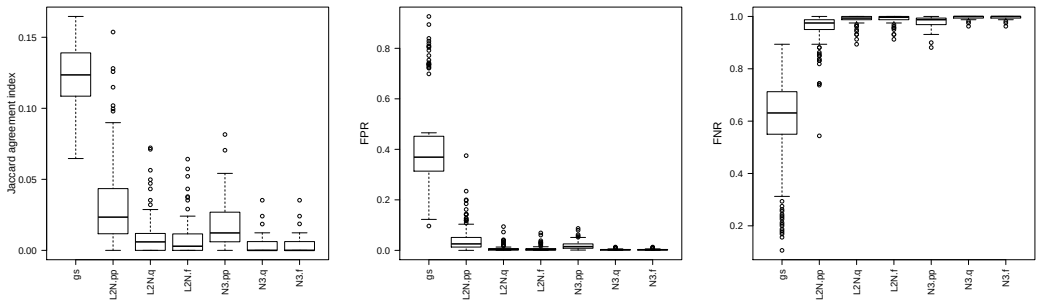


Figure 5.3: The boxplots of estimated Jaccard indices, FPR, and FNR based on the 100 simulated datasets in SimII. The closer to one the Jaccard index, the better the performance of the method. The closer to zero the FPR (FNR), the better the performance of the method.

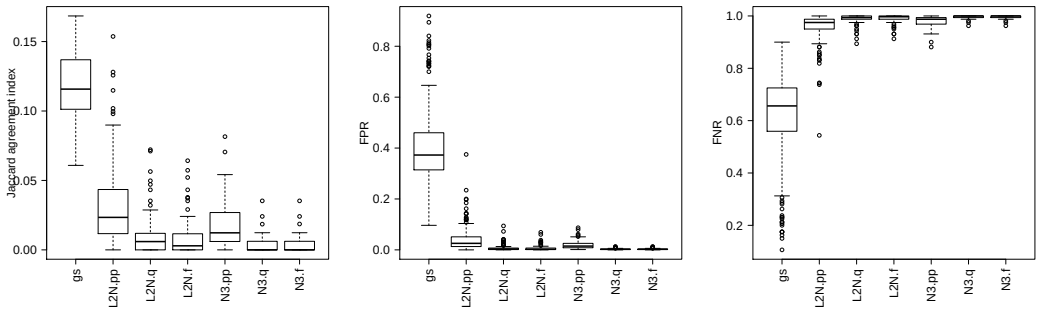


Figure 5.4: The boxplots of estimated Jaccard indices, FPR, and FNR based on the 100 simulated datasets in SimIV. The closer to one the Jaccard index, the better the performance of the method. The closer to zero the FPR (FNR), the better the performance of the method.

## 5.4 Real data analysis

### 5.4.1 Data description

We downloaded two miRNA data sets from online database GEO: GSE67138 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67138) and GSE67139 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67139). Both data sets are from the same project that aims to detect miRNAs that are differentially expressed between human hepatocellular carcinoma (HCC) tumor tissues with and without vascular invasion. GSE67138 is the first batch containing 57 samples (34 invasive tumor tissues and 23 non-invasive tumor tissues), while GSE67139 is the

second batch containing 120 samples (60 invasive tumor tissues and 60 non-invasive tumor tissues). The expression levels of miRNAs in both GEO datasets were measured by using Affymetrix Multispecies miRNA-1 Array (GPL8786). Both datasets contain 847 miRNAs. Since both data sets have been normalized, we used the two data sets directly in the further analyses. Since GSE67139 has a larger sample size than GSE61738, we regarded GSE67139 as the discovery set and GSE67138 as the validation set.

Following Qiu et al.'s (2008) data preprocessing steps, we first performed the Box-Cox transformation and miRNA-profile scaling on the two miRNA expression data sets. We then applied the seven methods (the gs method and the six existing model-based methods) to the discovery set (GSE67139) to detect miRNAs differentially variable between invasive tumors and non-invasive tumors. If a miRNA is assigned to the OV or UV clusters based on the seven model-based clustering methods, we claim that this miRNA has significantly different variances between invasive tumors and non-invasive tumors. We then applied the same procedure to the validation set (GSE67138). We claim that a miRNA is a validated DV miRNA (1) if the miRNA is DV in both discovery and validation sets, and (2) if the sign of the difference $(s_c^2 - s_n^2)$ is the same in both datasets, where $s_c^2$ and $s_n^2$ are sample variances for cases and controls, respectively. We next calculated the proportion of the validated DV

miRNAs (i.e., validation rate) $pValid = \frac{nValid}{nSig}$, where $nSig$ is the number of DV miRNAs in the discovery set (GSE67139) and $nValid$ is the number of DV miRNAs significant and sharing the same difference direction of variances in both data sets. To estimate the variation of the validation rate $pValid$, we obtained the 100 bootstrap validation rates based on 100 bootstrap discovery sets and validation sets. We then tested to see if the median bootstrap validation rate of the gs method is the same as that of each of the other six methods by using two-sided Wilcoxon signed-rank tests.

### 5.4.2 Results

The numbers of the DV miRNAs in the discovery set (GSE67139) and the numbers and proportions of the validated DV miRNAs were shown in Table 5.1. The gs method detected 358 DV probes based on the discovery set (GSE67139), 67 of which were validated in the validation set (GSE67138). Amongst the 67 validated DV miRNAs, 66 miRNAs were in Cluster OV and only one miRNA was in Cluster UV. The proportion of the validated DV miRNAs is 0.19 for the gs method, which is higher than those of the N3 and L2N methods. Moreover, the gs method had the highest median bootstrap validation rate among all seven methods (Figure 5). For all the seven methods, the number of validated OV miRNAs (nValid.OV) was much higher than the number of validated UV miRNAs (nValid.UV). This observation

is consistent with that observed by other researchers using DNA methylation data (Teschendorff and Widschwendter, 2012).

Table 5.1: The performance of model-based clustering methods on miRNA expression data sets GSE67139 and GSE67138.

| Method | nSig | n.OV | n.UV | nValid | nValid.OV | nValid.UV | pValid |
|---|---|---|---|---|---|---|---|
| gs | 358 | 262 | 96 | 67 | 66 | 1 | 0.19 |
| L2N.pp | 225 | 121 | 104 | 30 | 29 | 1 | 0.13 |
| L2N.q | 173 | 69 | 104 | 17 | 16 | 1 | 0.10 |
| L2N.f | 157 | 60 | 97 | 16 | 15 | 1 | 0.10 |
| N3.pp | 247 | 141 | 106 | 34 | 33 | 1 | 0.14 |
| N3.q | 202 | 96 | 106 | 25 | 24 | 1 | 0.12 |
| N3.f | 178 | 74 | 104 | 18 | 17 | 1 | 0.10 |

$^{\text{nSig}}$ : the number of DV miRNAs detected in the discovery set (GSE67139);
$^{\text{n.OV}}$ : the number of OV miRNAs detected in GSE67139;
$^{\text{n.UV}}$ : the number of UV miRNAs detected in GSE67139;
$^{\text{nValid}}$ : the number of validated DV miRNAs in the validation set (GSE67138);
$^{\text{nValid.OV}}$ : the number of validated OV miRNAs in GSE67138;
$^{\text{nValid.UV}}$ : the number of validated UV miRNAs in GSE67138;
$^{\text{pValid}}$ : $= \frac{nValid}{nSig}$, the proportion of significant DV miRNAs detected in GSE67139 and truly validated in GSE67138.

The gs method detected 67 validated differentially variated miRNAs (66 OV and 1 UV), seven of which are only differential in variances. The seven DV-only miRNAs are hsa-miR-1826, hsa-miR-191, hsa-miR-194-star, hsa-miR-222, hsa-miR-502-3p, hsa-miR-93, and hsa-miR-99b. With the exception of hsa-miR-1826, all DV-only miRNAs have been associated with HCC. Elyakim et al. (2010) showed that miR-191 is a candidate oncogene target for hepatocellular carcinoma therapy.

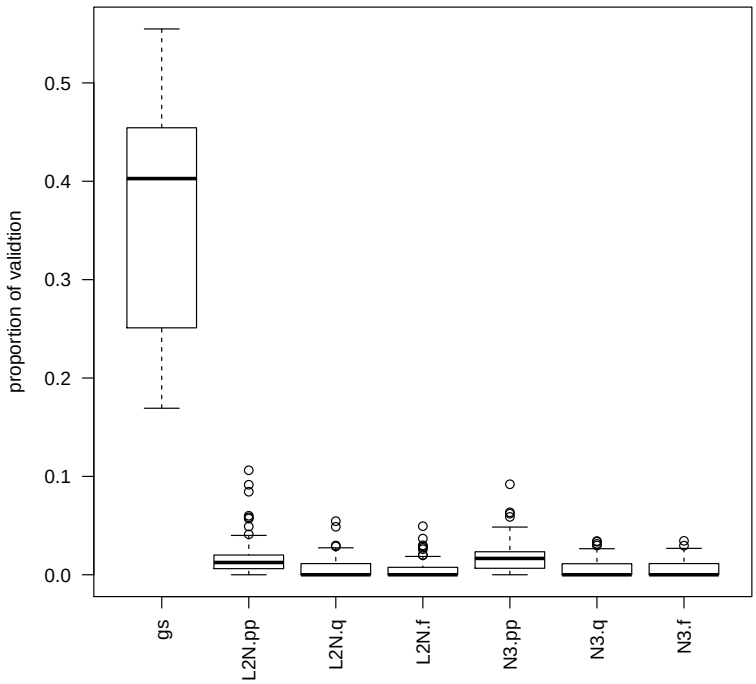


Figure 5.5: The boxplots of validation rates based on 100 bootstrap samples.

Law and Wong (2011) reported the association of miR-194 with metastatic behavior of HCC. Murakami et al. (2006) reported that miR-222 is increased in poorly versus moderately versus well-differentiated hepatomas. Jin et al. (2016) reported that miR-502-3p suppressed cell proliferation, migration, and invasion in HCC by targeting SET. Li et al. (2009) confirmed that the miR-106b-25 cluster, which miR-93 belongs to, is over-expressed in HCC. Morishita et al. (2016) found that miR-99b is up-regulated in HBV-infected HCC cells.

Based on the miRSystem analysis, there are 1,639 genes targeted by the identified seven DV-only miRNAs. These 1,639 genes are in six Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways: CALCIUM SIGNALING PATHWAY, SALIVARY SECRETION, AMYOTROPHIC LATERAL SCLEROSIS (ALS), MAPK SIGNALING PATHWAY, PPAR SIGNALING PATHWAY, and ALZHEIMER'S DISEASE. All of these six pathways have been linked to HCC in the literature. For example, Huang et al. (2017) reported that increased mitochondrial fission induced cytosolic calcium signaling in HCC cells. Chen et al. (2017) reported that in a mice study, DNA methylation marks that are differentially methylated between livers with HCC and livers without HCC are enriched in the SALIVARY SECRETION pathway. Seol et al. (2016)'s results suggest that Riluzole, an amyotrophic lateral sclerosis (ALS) drug, has an anti-cancer effect on HCC. Feng et al. (2018) reported that cantharidic acid inhibits HCC cell proliferation by inducing cell apoptosis through the p38 MAPK signaling pathway. Nwosu et al. (2017) reported that down-regulated genes (HCC vs. non-HCC) were enriched in PPAR SIGNALING PATHWAY based on each of the eight HCC datasets downloaded from the Gene Expression Omnibus (GEO). Jin et al. (2015) reported that Kynurenine 3-monooxygenase (KMO), an enzyme playing a critical role in Huntingtons and Alzheimers diseases, exhibits tumor-promoting effects toward HCC. Hence, DV-only miRNAs are biologically relevant to HCC.

## 5.5 Discussion

In this chapter, we propose a novel model-based clustering method (the gs method) to detect miRNAs that have different variances between cases and controls. We impose special structures on covariance matrices for each cluster of miRNAs based on prior information about the relationship between variances of cases and controls. The proposed method is different from probe-wise equal-variance tests in that it does not involve hypothesis testing. The real data analysis shows that the gs method has a larger median bootstrap validation rate than the six existing model-based equal-variance detecting methods. The simulation studies show that the gs method outperforms the six existing model-based detection methods if the miRNA expression data follow a mixture of multivariate normal distributions.

Several model-based clustering algorithms have been proposed to detect differential variable genetic probes in the literature, such as the N3 methods (Bar et al., 2014) and the L2N methods (Bar and Schifano, 2018). The N3 methods and L2N methods do not seem to work as well as the gs method when the underlying distribution of miRNA expression levels is multivariate normal. The poor performance of N3 and L2N is probably due to the fact that the gs method directly models the observed expression levels to avoid losing information, while the N3 and L2N methods model the summary statistics (e.g., mean, variance, or difference of means). Moreover,

the N3 and L2N methods apply a couple of approximations to derive the marginal densities, while approximations might cause deviations from true marginal densities.

In summary, the proposed gs method assumes expression levels from the mixture of multivariate normal distributions. The proposed gs method performs better than the other model-based methods if the underlying assumption is satisfied.

# 6 Conclusion and future work

In this chapter, we summarize the results of this dissertation and discuss some directions for future research.

First, we propose a new method to cluster high-dimensional protein sequences and apply it to influenza A H3N2 hemagglutinin sequences. After the entropy-based dimensionality reduction, the 1960 protein sequences active from 1998 to 2012 are aggregated into 23 clusters using the Hamming Distance vector algorithm. Based on these clusters, we investigate the relationship between past vaccines and the dominant cluster in each influenza season. We find that the dominant flu clusters replace one another every 2-5 years. The dominant clusters are not periodic within the time span. Once the HA cluster evolves away from a given region of the sequence space, it does not reoccur within that region. Furthermore, the dominant clusters often diverge from the clusters housing vaccine strains.

One possible research is to improve on the Hamming Distance vector algorithm into the algorithm of clustering categorical data and continuous data simultaneously.

Such an improved algorithm will enable us to consider more information about amino acids and provide more details about the resulting evolutionary pattern. Furthermore, we can introduce an amino acid substitution model (Dang et al., 2010) and epidemiological model (Du et al., 2017) to integrate the genetic information and transmission information of the influenza virus. This integration can help us forecast the flu epidemic dynamics and design more effective future vaccines.

For DNA methylation studies, we propose three robust joint score tests. We make extensive comparisons among the three proposed tests with jointLRT, KS test, and the AW test. Systematic simulation studies show that at least one of the three proposed tests perform better (i.e., having larger power, while keeping nominal Type I error rates) than the existing tests for data with outliers or from non-normal distributions. The analyses of the three real data sets demonstrate that the three proposed tests have higher true validation rates than those from jointLRT, KS, and AW. The three proposed joint score tests are robust against the violation of normality assumption and the presence of outlying observations in comparison with three existing tests. Among the three proposed tests, iAW.BF seems to be the most robust and effective one for all simulated scenarios and also in real data analyses. Moreover, we conduct systematic simulation studies and real data analysis to compare the performance of seven statistical tests for equal-variance hypothesis on DNA methy-

lation data, including two tests that were recently proposed in the literature. Our results show that the Brown-Forsythe test and trimmed-mean-based Levene's test have good performance in testing for equal variance in our simulation studies and real data analysis.

Since the AW-type joint test is derived from the generalized linear regression model, it can be naturally generalized to incorporate covariates (Ahn and Wang, 2013). It has been demonstrated that many other factors, such as age, gender, race, can contribute to aberrant DNA methylation pattern (Phipson and Oshlack, 2014; Zhang et al., 2011). These factors can confound the effect of DNA methylation on diseases. To identify the true methylation sites associated with the disease of interest, we need to evaluate and adjust the effects of these confounding factors via appropriate methods (Teschendorff and Relton, 2018). Moreover, we can consider the correlations between CpG sites within a genomic region, such as the region nearby an important promoter. Under the framework of the generalized linear model, our proposed methods can extend beyond investigating one CpG site to the region-based analysis.

For studies of miRNA expression data, we propose a novel model-based clustering method (the gs method) to detect differentially variable miRNAs between cases and controls. We impose special structures on covariance matrices for each cluster of

miRNAs based on prior information about the relationship between variances in cases and controls. This method is different from probe-wise equal-variance tests in that it does not involve hypothesis testing. The simulation studies show that this method outperforms other model-based methods if the underlying assumption that the miRNA expression levels follow a mixture of multivariate normal distributions. The real data analysis shows that the gs method has a larger median bootstrap validation rate than the other model-based methods.

The proposed gs method has been demonstrated to rely on the normality assumption. Since it is common to have outliers in miRNA expression data, we intend to employ a broader family of distribution, Pearson type VII distribution (Pearson, 1916; Sun et al., 2010), to replace the multivariate normal distribution. Since the Pearson type VII distribution includes Student t-distribution and has heavy tails, the method based on the marginal model of Pearson type VII distribution can improve the detecting power when outliers are present (Sun et al., 2010). We could improve the gs method into a robust version against the violation of the normality assumption on the component distributions. We could also consider detecting differentially expressed (DE) and differentially variated (DV) miRNAs simultaneously using a nine-component multivariate normal mixture model.

# Bibliography

S. Ahn and T. Wang. A powerful statistical method for identifying differentially methylated markers in complex diseases. *Pacific Symposium on Biocomputing*, pages 69–79, 2013.

E. Bandrés, E. Cubedo, X. Agirre, R. Malumbres, R. Zarate, N. Ramirez, A. Abajo, A. Navarro, I. Moreno, M. Monzo, et al. Identification by real-time pcr of 13 mature microrna differentially expressed in colorectal cancer and non-tumoral tissues. *Molecular cancer*, 5(1):29, 2006.

H. Bar and E. D. Schifano. Differential variation and expression analysis. *bioRxiv*, page 276337, 2018.

H. Y. Bar, J. G. Booth, and M. T. Wells. A mixture-model approach for parallel testing for unequal variances. *Statistical applications in genetics and molecular biology*, 11(1):1–21, 2012.

H. Y. Bar, J. G. Booth, and M. T. Wells. A bivariate model for simultaneous testing in bioinformatics data. *Journal of the American Statistical Association*, 109(506): 537–547, 2014.

M. S. Bartlett. Properties of Sufficiency and Statistical Tests. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 160 (901):268–282, 1937. ISSN 1364-5021. doi: 10.1098/rspa.1937.0109. URL `http://www.jstor.org/stable/96803`.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

C. Bock. Analysing and interpreting dna methylation data. *Nature reviews. Genetics*, 13(10):705, 2012.

M. F. Boni. Vaccination and antigenic drift in influenza. *Vaccine*, 26:C8–C14, 2008.

A.-L. Boulesteix and K. Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, 8(1):32–44, 2006.

M. B. Brown and A. B. Forsythe. Robust Tests for Equality of Variances. *Journal of the American Statistical Association*, 69(May 2015):364–367, 1974. ISSN 00255718. doi: 10.2307/2285659.

F. Carrat and A. Flahault. Influenza vaccine: the challenge of antigenic drift. *Vaccine*, 25(39):6852–6862, 2007.

H. Chen, W. Cai, E. Chu, J. Tang, C. Wong, S. Wong, W. Sun, Q. Liang, J. Fang, Z. Sun, et al. Hepatic cyclooxygenase-2 overexpression induced spontaneous hepatocellular carcinoma formation in mice. *Oncogene*, 36(31):4415, 2017.

Y. Chen, Y. Ning, C. Hong, and S. Wang. Semiparametric tests for identifying differentially methylated loci with case-control designs using illumina arrays. *Genetic Epidemiology*, 38(1):42–50, 2014. ISSN 07410395. doi: 10.1002/gepi.21774.

W. J. Conover, M. E. Johnson, and M. M. Johnson. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23(4):351–361, 1981.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

B. R. Cullen. Herpesvirus microrna: phenotypes and functions. *Current opinion in virology*, 1(3):211–215, 2011.

C. C. Dang, Q. S. Le, O. Gascuel, and V. S. Le. Flu, an amino acid substitution model for influenza proteins. *BMC evolutionary biology*, 10(1):99, 2010.

S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks. Evaluation of the infinium methylation 450k technology. *Epigenomics*, 3(6):771–784, 2011.

T. Devosse, A. Guillabert, N. DHaene, A. Berton, P. De Nadai, S. Noel, M. Brait, J.-D. Franssen, S. Sozzani, I. Salmon, and M. Parmentier. Formyl peptide receptor-like 2 is expressed and functional in plasmacytoid dendritic cells, tissue-specific macrophage subpopulations, and eosinophils. *The Journal of immunology*, 182(8): 49744984, 2009.

A. Dobson. *An Introduction to Generalized Linear Models*. Chapman and Hall, 1990.

X. Du, A. A. King, R. J. Woods, and M. Pascual. Evolution-informed forecasting of seasonal influenza a (h3n2). *Science translational medicine*, 9(413):eaan5325, 2017.

R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

E. Elyakim, E. Sitbon, A. Faerman, S. Tabak, E. Montia, L. Belanis, A. Dov, E. G. Marcusson, C. F. Bennett, A. Chajut, et al. hsa-mir-191 is a candidate oncogene target for hepatocellular carcinoma therapy. *Cancer research*, 70(20):8077–8087, 2010.

B. Everitt and D. Hand. Finite mixture distribution, monograph on applied probability and statistics, 1981.

L. Fahrmeir. Asymptotic testing theory for generalized linear models. *Statistics: A Journal of Theoretical and Applied Statistics*, 18(1):65–76, 1987.

S. A. Farkas, N. Milutin-Gašperov, M. Grce, and T. K. Nilsson. Genome-wide dna methylation assay reveals novel candidate biomarker genes in cervical cancer. *Epigenetics*, 8(11):1213–1225, 2013.

A. P. Feinberg, R. A. Irizarry, D. Fradin, M. J. Aryee, P. Murakami, T. Aspelund, G. Eiriksdottir, T. B. Harris, L. Launer, V. Gudnason, et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. *Science translational medicine*, 2(49):49ra67–49ra67, 2010.

I.-C. Feng, M.-J. Hsieh, P.-N. Chen, Y.-H. Hsieh, H.-Y. Ho, S.-F. Yang, and C.-B. Yeh. Cantharidic acid induces apoptosis through the p38 mapk signaling pathway in human hepatocellular carcinoma. *Environmental toxicology*, 33(3):261–268, 2018.

C. Fernández-Hernando, C. M. Ramírez, L. Goedeke, and Y. Suárez. Micrornas in metabolic disease. *Arteriosclerosis, thrombosis, and vascular biology*, 33(2):178–185, 2013.

G. Forney. Generalized minimum distance decoding. *IEEE Transactions on Information Theory*, 12(2):125–131, 1966.

S. A. Frank. Evolution in health and medicine sackler colloquium: Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proceedings of the National Academy of Sciences of the United States of America*, 107(Suppl 1):1725–1730, 2010.

S. Gopalakrishnan, B. O. Van Emburgh, and K. D. Robertson. Dna methylation in development and human disease. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 647(1):30–38, 2008.

S. M. Hammond. An overview of microrna. *Advanced drug delivery reviews*, 87: 3–14, 2015.

K. D. Hansen, W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, et al. Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, 43(8):768–775, 2011.

L. He and G. J. Hannon. Microrna: small rna with a big role in gene regulation. *Nature Reviews Genetics*, 5(7):522, 2004.

Q. Huang, H. Cao, L. Zhan, X. Sun, G. Wang, J. Li, X. Guo, T. Ren, Z. Wang, Y. Lyu, et al. Mitochondrial fission forms a positive feedback loop with cytosolic calcium signaling pathway to promote autophagy in hepatocellular carcinoma cells. *Cancer letters*, 403:108–118, 2017.

P. Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 11(2): 37–50, 1912.

A. E. Jaffe, A. P. Feinberg, R. A. Irizarry, and J. T. Leek. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics*, 13(1):166–178, 2011.

H. Jin, Y. Zhang, H. You, X. Tao, C. Wang, G. Jin, N. Wang, H. Ruan, D. Gu, X. Huo, et al. Prognostic significance of kynurenine 3-monooxygenase and effects on proliferation, migration, and invasion of human hepatocellular carcinoma. *Scientific reports*, 5:10466, 2015.

H. Jin, M. Yu, Y. Lin, B. Hou, Z. Wu, Z. Li, and J. Sun. Mir-502-3p suppresses cell proliferation, migration, and invasion in hepatocellular carcinoma by targeting set. *OncoTargets and therapy*, 9:3281, 2016.

D. M. Knipe and P. M. Howley, editors. *Fields' Virology*, volume 1. Lippincott Williams & Wilkins, 5 edition, 2007.

A. Kozomara and S. Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73, 2013.

E. Kudela, V. Holubekova, A. Farkasova, and J. Danko. Determination of malignant potential of cervical intraepithelial neoplasia. *Tumor Biology*, 37(2):1521–1525, 2016.

P. T.-Y. Law and N. Wong. Emerging roles of microrna in the intracellular signaling networks of hepatocellular carcinoma. *Journal of gastroenterology and hepatology*, 26(3):437–449, 2011.

H. Levene. Robust tests for equality of variances1. *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, 2:278–292, 1960.

Y. Li, W. Tan, T. W. Neo, M. O. Aung, S. Wasser, S. G. Lim, and T. Tan. Role of the mir-106b-25 microrna cluster in hepatocellular carcinoma. *Cancer science*, 100(7):1234–1242, 2009.

R. C. Littell and J. L. Folks. Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association*, 66(336):802–806, 1971.

R. C. Littell and J. L. Folks. Asymptotic optimality of fisher's method of combining independent tests ii. *Journal of the American Statistical Association*, 68(341): 193–194, 1973.

J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, et al. Microrna expression profiles classify human cancers. *nature*, 435(7043):834, 2005.

M. Łuksza and M. Lässig. A predictive fitness model for influenza. *Nature*, 507 (7490):57–61, 2014.

J. C. Mar, N. A. Matigian, A. Mackay-Sim, G. D. Mellick, C. M. Sue, P. A. Silburn, J. J. McGrath, J. Quackenbush, and C. A. Wells. Variance of gene expression identifies altered network constraints in neurological disease. *PLoS genetics*, 7(8): e1002207, 2011.

A. Morishita, H. Iwama, S. Fujihara, T. Sakamoto, K. Fujita, J. Tani, H. Miyoshi, H. Yoneyama, T. Himoto, and T. Masaki. Microrna profiles in various hepatocellular carcinoma cell lines. *Oncology letters*, 12(3):1687–1692, 2016.

Y. Murakami, T. Yasuda, K. Saigo, T. Urashima, H. Toyoda, T. Okanoue, and K. Shimotohno. Comprehensive analysis of microrna expression patterns in hepatocellular carcinoma and non-tumorous tissues. *Oncogene*, 25(17):2537, 2006.

M. I. Nelson and E. C. Holmes. The evolution of epidemic influenza. *Nature reviews. Genetics*, 8(3):196, 2007.

Z. C. Nwosu, D. A. Megger, S. Hammad, B. Sitek, S. Roessler, M. P. Ebert, C. Meyer, and S. Dooley. Identification of the consistently altered metabolic targets in human hepatocellular carcinoma. *Cellular and molecular gastroenterology and hepatology*, 4(2):303–323, 2017.

E. S. Pearson and J. Neyman. On the problem of two samples, 1930.

K. Pearson. Ix. mathematical contributions to the theory of evolution.—xix. second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 216:429–457, 1916.

S. Perng and R. Littell. A test of equality of two normal population means and variances. *Journal of the American Statistical Association*, 71(356):968–971, 1976.

B. Phipson and A. Oshlack. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biology*, 15: 465, 2014.

R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Djik, B. Muhlhausler, C. Stirzaker, and S. J. Clark. Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. *Genome biology*, 17(1):208, 2016.

J. B. Plotkin, J. Dushoff, and S. A. Levin. Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. *Proceedings of the National Academy of Sciences*, 99(9):6263–6268, 2002.

C. C. Pritchard, H. H. Cheng, and M. Tewari. Microrna profiling: approaches and considerations. *Nature Reviews Genetics*, 13(5):358, 2012.

W. Qiu, W. He, X. Wang, and R. Lazarus. A marginal mixture model for selecting differentially expressed genes across two types of tissue samples. *The International Journal of Biostatistics*, 4(1), 2008.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `https://www.R-project.org/`.

K. E. Resnick, H. Alder, J. P. Hagan, D. L. Richardson, C. M. Croce, and D. E. Cohn. The detection of differentially expressed microrna from the serum of ovarian cancer patients using a novel real-time pcr platform. *Gynecologic oncology*, 112(1): 55–59, 2009.

H. S. Seol, S. E. Lee, J. S. Song, H. Y. Lee, S. Park, I. Kim, S. R. Singh, S. Chang, and S. J. Jang. Glutamate release inhibitor, riluzole, inhibited proliferation of human hepatocellular carcinoma cells by elevated ros production. *Cancer letters*, 382(2):157–165, 2016.

Y.-J. Shiah, M. Fraser, R. G. Bristow, and P. C. Boutros. Comparison of preprocessing methods for infinium humanmethylation450 beadchip array. *Bioinformatics*, 33(20):3151–3157, 2017.

L. H. Shoemaker. Fixing the f test for equal variances. *The American Statistician*, 57(2):105–114, 2003.

D. V. Silva Rodrigues, V. V. Silva Monteiro, K. C. Navegantes-Lima, A. L. de Brito Oliveira, S. L. de França Gaspar, L. B. Gonçalves Quadros, and M. C. Monteiro. Microrna in cell cycle progression and proliferation: molecular mechanisms and pathways. *Non-coding RNA Investigation*, 2(5), 2018.

G. K. Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–26, 2004. ISSN 1544-6115. doi: 10.2202/1544-6115.1027.

Z. Staneková and E. Varečková. Conserved epitopes of influenza a virus inducing protective immunity and their prospects for universal vaccine development. *Virology journal*, 7(1):351, 2010.

D. Strbenac, G. J. Mann, J. Y. Yang, and J. T. Ormerod. Differential distribution improves gene selection stability and has competitive classification performance for patient survival. *Nucleic acids research*, 44(13):e119–e119, 2016.

J. Sun, A. Kabán, and J. M. Garibaldi. Robust mixture clustering using pearson type vii distribution. *Pattern Recognition Letters*, 31(16):2447–2454, 2010.

K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28 (10):2731–2739, 2011.

A. Teschendorff and M. Widschwendter. Differential variability improves the identification of cancer risk markers in dna methylation studies profiling precursor cancer lesions. *Bioinformatics*, 28(11):1487–1494, 2012.

A. E. Teschendorff and C. L. Relton. Statistical and integrative system-level analysis of dna methylation data. *Nature Reviews Genetics*, 19(3):129, 2018.

A. E. Teschendorff, U. Menon, A. Gentry-Maharaj, S. J. Ramus, D. J. Weisenberger, H. Shen, M. Campan, H. Noushmehr, C. G. Bell, A. P. Maxwell, et al. Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome research*, 20(4):440–446, 2010.

D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*, volume 2. Wiley, New York, 1985.

D. L. Van den Hove, K. Kompotis, R. Lardenoije, G. Kenis, J. Mill, H. W. Steinbusch, K.-P. Lesch, C. P. Fitzsimons, B. De Strooper, and B. P. Rutten. Epigenetically regulated micrornas in alzheimer's disease. *Neurobiology of aging*, 35(4):731–745, 2014.

S. Wahl, N. Fenske, S. Zeilinger, K. Suhre, C. Gieger, M. Waldenberger, H. Grallert, and M. Schmid. On the potential of models for location and scale for genome-wide dna methylation data. *BMC bioinformatics*, 15(1):232, 2014.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

J. C. William. Practical nonparametric statistics, 1971.

M. C. Wu, B. R. Joubert, P.-f. Kuan, S. E. Håberg, W. Nystad, S. D. Peddada, and S. J. London. A systematic assessment of normalization approaches for the infinium 450k methylation platform. *Epigenetics*, 9(2):318–329, 2014.

M. Yazicioglu, L. Monaldini, K. Chu, F. Khazi, S. Murphy, H. Huang, P. Margaritis, and K. High. Cellular localization and characterization of cytosolic binding partners for Gla domain-containing proteins PRRG4 and PRRG2. *Journal of Biological Chemistry*, 288(36):2590825914, 2013.

F. F. Zhang, R. Cardarelli, J. Carroll, K. G. Fulda, M. Kaur, K. Gonzalez, J. K. Vishwanatha, R. M. Santella, and A. Morabia. Significant differences in global genomic dna methylation by gender and race/ethnicity in peripheral blood. *Epigenetics*, 6 (5):623–629, 2011.

L. Zhang, X. Xu, and G. Chen. The exact likelihood ratio test for equality of two normal populations. *The American Statistician*, 66(3):180–184, 2012.

P. Zhang, X. Wang, and P. X.-K. Song. Clustering categorical data based on distance vectors. *Journal of the American Statistical Association*, 101(473):355–367, 2006.

X. Zhang, Y. Hu, A. C. Justice, B. Li, Z. Wang, H. Zhao, J. H. Krystal, and K. Xu. Dna methylation signatures of illicit drug injection and hepatitis c are associated with hiv frailty. *Nature communications*, 8(1):2243, 2017.

# A Additional simulation results in Chapter 3

Table A.1: Ranks of the seven equal-variance tests in terms of power for simulation scenarios of sample size 200.

| n | Distribution | eqMean | Outlier | F | Bartlett | Levene | L.trim | BF | PO.AD | PO.SQ |
|---|---|---|---|---|---|---|---|---|---|---|
| 200 | N | yes | no | 2 | 2 | 5.5 | 5.5 | 5.5 | 5.5 | 2 |
| 200 | N | yes | yes | - | - | - | 2.5 | 2.5 | - | 1 |
| 200 | N | no | no | 2 | 2 | 5.5 | 5.5 | 5.5 | 5.5 | 2 |
| 200 | N | no | yes | - | - | - | 2.5 | 2.5 | - | 1 |
| 200 | c.N | no | no | 4 | 4 | 4.0 | 4.0 | 4.0 | 4.0 | 4 |
| 200 | c.N | no | yes | - | - | - | 2.0 | 2.0 | 2.0 | - |
| 200 | t | yes | no | - | - | 2.0 | 4.0 | 5.0 | 3.0 | 1 |
| 200 | t | yes | yes | - | - | 2.0 | 3.0 | 4.0 | 1.0 | 5 |
| 200 | t | no | no | - | - | - | 1.0 | 2.0 | - | 3 |
| 200 | t | no | yes | - | - | - | - | - | - | 1 |
| 200 | chisq | yes | no | - | - | - | - | 1.0 | - | 2 |
| 200 | chisq | yes | yes | - | - | - | 1.0 | 2.0 | - | 3 |
| 200 | chisq | no | no | - | - | - | - | - | - | 1 |
| 200 | chisq | no | yes | - | - | - | 1.0 | 2.0 | - | 3 |
| 200 | c.chisq | no | no | - | - | - | - | 1.0 | - | 2 |
| 200 | c.chisq | no | yes | - | - | - | 1.5 | 1.5 | - | 3 |

"-" : no power can be considered because the test has inflated type I error rates for all the scenarios in the situation.

Table A.2: Ranks of the seven equal-variance tests in terms of power for simulation scenarios of sample size 50.

| n | Distribution | eqMean | Outlier | F | Bartlett | Levene | L.trim | BF | PO.AD | PO.SQ |
|---|---|---|---|---|---|---|---|---|---|---|
| 50 | N | yes | no | 1.5 | 1.5 | - | 4.0 | 5.0 | - | 3.0 |
| 50 | N | yes | yes | - | - | 2 | 3.0 | 4.0 | 1 | 5.0 |
| 50 | N | no | no | 1.5 | 1.5 | - | 4.0 | 5.0 | - | 3.0 |
| 50 | N | no | yes | - | - | - | 2.0 | 3.0 | 1 | 4.0 |
| 50 | c.N | no | no | 1.5 | 1.5 | - | 4.0 | 5.0 | - | 3.0 |
| 50 | c.N | no | yes | - | - | - | - | 1.0 | - | - |
| 50 | t | yes | no | - | - | - | 1.0 | 2.0 | - | 3.0 |
| 50 | t | yes | yes | - | - | 2 | 3.5 | 3.5 | 1 | 5.0 |
| 50 | t | no | no | - | - | - | 1.0 | 2.0 | - | 3.0 |
| 50 | t | no | yes | - | - | - | 1.0 | 2.0 | - | 3.0 |
| 50 | chisq | yes | no | - | - | - | - | 1.5 | - | 1.5 |
| 50 | chisq | yes | yes | - | - | - | 1.0 | 2.0 | - | 3.0 |
| 50 | chisq | no | no | - | - | - | - | - | - | 1.0 |
| 50 | chisq | no | yes | - | - | - | 1.0 | 2.0 | - | 3.0 |
| 50 | c.chisq | no | no | - | - | - | - | 1.0 | - | 2.0 |
| 50 | c.chisq | no | yes | - | - | - | 1.0 | 2.0 | - | 3.0 |

"-" : no power can be considered because the test has inflated type I error rates for all the scenarios in the situation.

Table A.3: Ranks of the seven equal-variance tests in terms of power for simulation scenarios of sample size 20.

| n | Distribution | eqMean | Outlier | F | Bartlett | Levene | L.trim | BF | PO.AD | PO.SQ |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | N | yes | no | 1.5 | 1.5 | - | 4 | 5 | - | 3 |
| 20 | N | yes | yes | - | - | - | 2 | 3 | 1 | 4 |
| 20 | N | no | no | 1.5 | 1.5 | - | 4 | 5 | - | 3 |
| 20 | N | no | yes | - | - | - | 2 | 3 | 1 | 4 |
| 20 | c.N | no | no | 1.5 | 1.5 | - | 4 | 5 | - | 3 |
| 20 | c.N | no | yes | - | - | - | - | 1 | - | - |
| 20 | t | yes | no | - | - | - | 1 | 2 | - | 3 |
| 20 | t | yes | yes | - | - | 2 | 4 | 4 | 1 | 4 |
| 20 | t | no | no | - | - | - | 1 | 2 | - | 3 |
| 20 | t | no | yes | - | - | - | 2 | 3 | 1 | 4 |
| 20 | chisq | yes | no | - | - | - | - | 1 | - | - |
| 20 | chisq | yes | yes | - | - | - | 1 | 2 | - | 3 |
| 20 | chisq | no | no | - | - | - | - | - | - | - |
| 20 | chisq | no | yes | - | - | - | 1 | 2 | - | 3 |
| 20 | c.chisq | no | no | - | - | - | - | 1 | - | 2 |
| 20 | c.chisq | no | yes | - | - | - | 1 | 2 | - | 3 |

"-" : no power can be considered because the test has inflated type I error rates for all the scenarios in the situation.

Table A.4: Number of scenarios with inflated type I error rate and median of ranks for the seven tests from 48 simulated comparisons.

| | F | Bartlett | Levene | L.trim | BF | PO.AD | PO.SQ |
|---|---|---|---|---|---|---|---|
| $n_{reject}$ | 39 | 39 | 40 | 12 | 4 | 35 | 5 |
| $m$ | 1.5 | 1.5 | 2.5 | 2.0 | 2.0 | 1.0 | 3.0 |

$n_{reject}$ : number of scenarios where the test has inflated type I error rate;

$m$ : the median of the ranks of the power. For ranks with ties, average ranks were used.

# B Quality control and data preprocessing in Chapters 3 and 4

## QC and data preprocessing of HM27k data sets GSE30760 and GSE20080

GSE37020 contains a total of 48 samples, 24 of which have normal histology and the remaining are cervical intraepithelial neoplasia of grade 2 or higher (CIN2+). All of them are human papillomavirus (HPV) positive. Normal and CIN2+ samples are age-matched. GSE20080 also contains 48 samples. A total of 30 samples (11 HPV positive samples and 19 HPV negative samples) have normal cytology. The other 18 samples (all HPV positive) are with CIN2+. Moreover, normal and CIN2+ samples were age-matched.

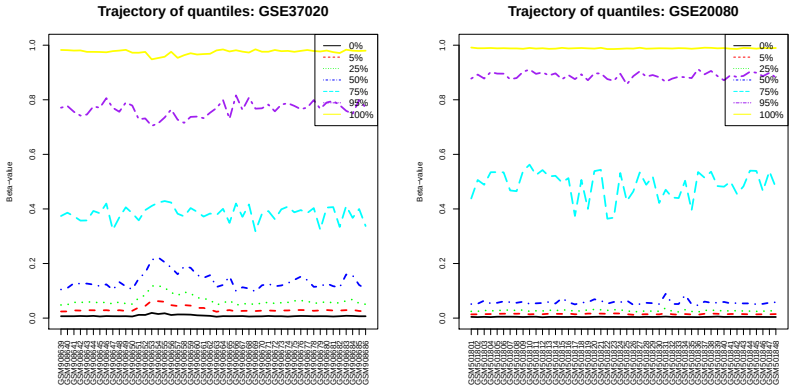


Figure B.1: The plot of quantiles across arrays.

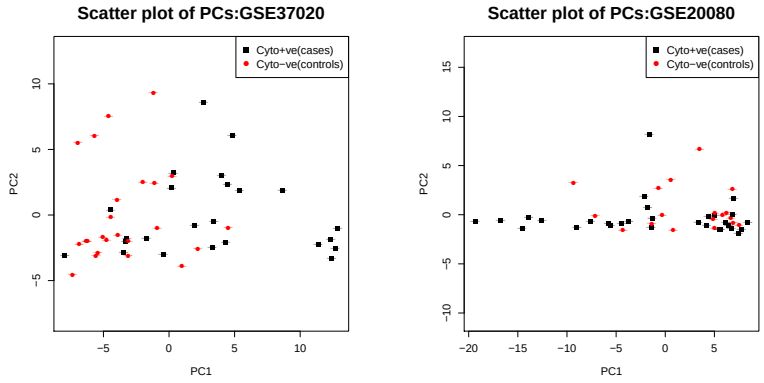


Figure B.2: The plot of the first principal component (PC1) versus the second principal component (PC2) for DNA methylation data.

For both of the data sets, we excluded some CpG sites residing on SNPs or with missing values. Quantile plots and principal component analysis did not show obvious patterns (see Figures B.1 and B.2). After quality control steps, GSE37020 has 23,066 CpG sites, and GSE20080 has 23,255 CpG sites. There are 22,859 CpG

sites in common between the two data sets. We used these 22,859 CpG sites in our real data analysis.

## QC and preprocessing of Illumina MethylationEPIC data

For EPIC data GSE107080, we downloaded the processed data set from GEO database (Zhang et al., 2017). We first removed the CpG sites with at least one missing value or with probe name using "ch" as the prefix. Second, CpG sites with detection p-values larger than or equal to $10^{-12}$ are discarded. There are $378,808$ CpG sites in the clean data set. We drew the plot of quantiles across arrays for the clean GSE107080 data set. The results did not show any obvious patterns (see Figure B.3).

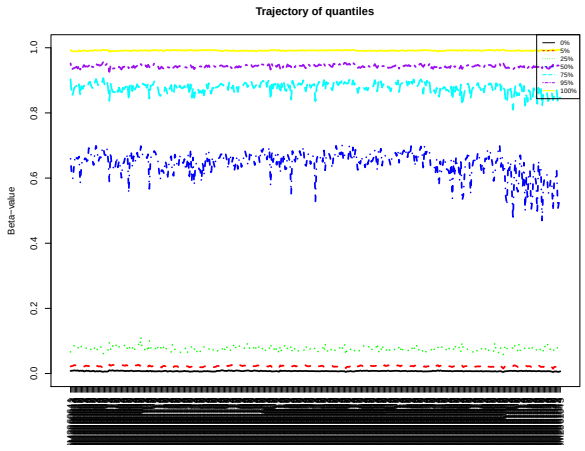


Figure B.3: The plot of quantiles across arrays for GSE107080.

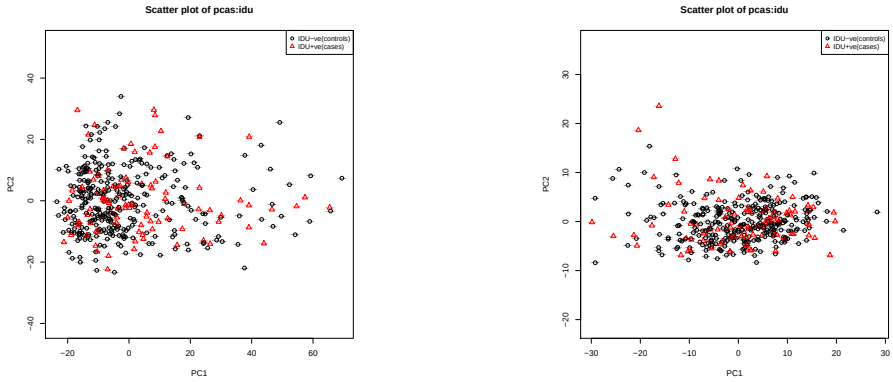


Figure B.4: The plot of the first principal component (PC1) versus the second principal component (PC2) for EPIC data GSE107080. Left panel is for unadjusted data; right panel is for adjusted data.

Next we did principal component analysis for the clean GSE107080 data set. The results did not show any obvious patterns (see the left panel in Figure B.4). Additionally, we regressed out the effects of age and cell type compositions and obtained the residuals. We re-did principle component analysis on the adjusted data and drew the plot of the first two principal components. No obvious patterns were found in the adjusted data (see the right panel in Figure B.4).

# C Additional simulation results in Chapter 4

Table C.1: The empirical type I error rates (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation values were generated from normal distributions without or with an outlier. The numbers of non-diseased and diseased samples are (50, 50).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 5.5 | 3.8 | 5.0 | 5.1 | 5.1 | 5.1 |
| (Type I error) | No | 1 | 1.1 | 0.5 | 0.8 | 0.9 | 0.9 | 0.9 |
| | No | 0.5 | 0.5 | 0.3 | 0.4 | 0.4 | 0.4 | 0.5 |
| diffM&eqV | No | 5 | 75.8 | 68.7 | 77 | 76.3 | 76.6 | 76.6 |
| | No | 1 | 53.1 | 40.8 | 53.5 | 52.9 | 53.2 | 53.6 |
| | No | 0.5 | 42.9 | 40.8 | 44.7 | 43.5 | 44.9 | 43.2 |
| eqM&diffV | No | 5 | 60.5 | 10.3 | 55.6 | 51.8 | 49.9 | 50.9 |
| | No | 1 | 35.5 | 2.0 | 25.9 | 27.7 | 26.1 | 27.6 |
| | No | 0.5 | 26.5 | 2.0 | 18.2 | 20.3 | 19.1 | 18.8 |
| diffM&diffV | No | 5 | 49.5 | 30 | 48.8 | 47.2 | 46.3 | 47.0 |
| | No | 1 | 27.1 | 10.0 | 25 | 23.6 | 23.2 | 23.9 |
| | No | 0.5 | 19.0 | 10.0 | 18.4 | 17.3 | 17.6 | 16.8 |
| eqM&eqV | Yes | 5 | 19.0 | 3.9 | 3.0 | 4.6 | 4.6 | 4.6 |
| (Type I error) | Yes | 1 | 6.4 | 0.5 | 0.4 | 0.7 | 0.7 | 0.7 |
| | Yes | 0.5 | 4.0 | 0.3 | 0.2 | 0.2 | 0.3 | 0.3 |
| diffM&eqV | Yes | 5 | 64.4 | 72.1 | 84.4 | 81.8 | 82.1 | 81.8 |
| | Yes | 1 | 33.9 | 44.3 | 63.9 | 59.7 | 59.5 | 60.0 |
| | Yes | 0.5 | 24.3 | 34.9 | 56.7 | 52.2 | 51.5 | 50.0 |
| eqM&diffV | Yes | 5 | 6.3 | 9.1 | 19.7 | 29.7 | 29.2 | 30.0 |
| | Yes | 1 | 1.1 | 2.0 | 7.0 | 12.3 | 11.8 | 12.6 |
| | Yes | 0.5 | 0.6 | 1.0 | 4.9 | 9.2 | 8.5 | 8.6 |
| diffM&diffV | Yes | 5 | 13.9 | 32.6 | 45.8 | 46.1 | 47.2 | 47.3 |
| | Yes | 1 | 3.4 | 11.0 | 23.7 | 23.2 | 24.1 | 24.5 |
| | Yes | 0.5 | 1.7 | 7.3 | 18.9 | 18.3 | 19.0 | 18.2 |

Table C.2: The empirical type I error rates (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation values were generated from Beta distributions without or with an outlier. The numbers of non-diseased and diseased samples are (50, 50).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 6.0 | 3.6 | 5.4 | 5.5 | 5.3 | 5.5 |
| (Type I error) | No | 1 | 1.3 | 0.4 | 1.0 | 1.2 | 1.2 | 1.2 |
| | No | 0.5 | 0.6 | 0.2 | 0.5 | 0.6 | 0.6 | 0.6 |
| diffM&eqV | No | 5 | 73.7 | 70.2 | 77.3 | 75.3 | 76.8 | 76.3 |
| | No | 1 | 50.8 | 41.6 | 54.1 | 50.6 | 51.7 | 50.8 |
| | No | 0.5 | 42.1 | 41.6 | 46.0 | 40.9 | 43.0 | 42.2 |
| eqM&diffV | No | 5 | 56.7 | 9.9 | 52.7 | 49.8 | 49.1 | 49.8 |
| | No | 1 | 31.2 | 1.5 | 24.0 | 23.9 | 21.9 | 22.6 |
| | No | 0.5 | 23.8 | 1.5 | 16.7 | 16.7 | 15.7 | 16.5 |
| diffM&diffV | No | 5 | 50.5 | 35.2 | 56.8 | 52.2 | 53.7 | 53.8 |
| | No | 1 | 24.6 | 13.1 | 31.2 | 25.7 | 26.8 | 26.8 |
| | No | 0.5 | 17.2 | 13.1 | 23.4 | 17.8 | 20.0 | 19.8 |
| eqM&eqV | Yes | 5 | 16.1 | 3.8 | 3.3 | 4.7 | 4.5 | 4.7 |
| (Type I error) | Yes | 1 | 4.3 | 0.5 | 0.5 | 0.8 | 0.8 | 0.8 |
| | Yes | 0.5 | 2.4 | 0.2 | 0.3 | 0.4 | 0.4 | 0.4 |
| diffM&eqV | Yes | 5 | 73.8 | 74.2 | 85.5 | 84.1 | 84.9 | 84.2 |
| | Yes | 1 | 48.7 | 46.1 | 65.1 | 61.3 | 62.4 | 61.4 |
| | Yes | 0.5 | 37.6 | 46.1 | 55.6 | 51.5 | 50.9 | 51.1 |
| eqM&diffV | Yes | 5 | 2.5 | 8.4 | 6.4 | 19.5 | 18.7 | 18.7 |
| | Yes | 1 | 0.5 | 1.2 | 1.7 | 6.7 | 6.2 | 6.3 |
| | Yes | 0.5 | 0.2 | 1.2 | 0.9 | 4.1 | 3.5 | 3.7 |
| diffM&diffV | Yes | 5 | 2.8 | 28.4 | 13.4 | 10.8 | 10.9 | 10.7 |
| | Yes | 1 | 0.7 | 9.7 | 4.8 | 2.4 | 2.5 | 2.4 |
| | Yes | 0.5 | 0.4 | 9.7 | 3.2 | 1.4 | 1.4 | 1.4 |

Table C.3: The empirical type I error rates (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation values were generated from mixtures of two normal distributions without or with an outlier. The numbers of non-diseased and diseased samples are (50, 50).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 2.4 | 4 | 4.5 | 8.9 | 5.3 | 10.0 |
| (Type I error) | No | 1 | 0.5 | 0.6 | 0.9 | 2.7 | 1.4 | 2.8 |
| | No | 0.5 | 0.3 | 0.3 | 0.4 | 1.6 | 0.7 | 1.5 |
| diffM&eqV | No | 5 | 9.8 | 29.8 | 44.2 | 26.5 | 49.9 | 27.7 |
| | No | 1 | 2.2 | 10.5 | 20.4 | 7.9 | 23.2 | 10.0 |
| | No | 0.5 | 1.1 | 7.0 | 14.8 | 5.0 | 15.0 | 6.3 |
| eqM&diffV | No | 5 | 18.5 | 68.3 | 36.4 | 59 | 35.3 | 46 |
| | No | 1 | 3.6 | 31.5 | 19.7 | 33.3 | 16.7 | 26.5 |
| | No | 0.5 | 1.7 | 22.1 | 15.7 | 27.1 | 11.3 | 20.3 |
| diffM&diffV | No | 5 | 21.2 | 71.9 | 41.1 | 63.6 | 41.6 | 51.2 |
| | No | 1 | 4.8 | 34.4 | 23.6 | 37.7 | 21 | 30.7 |
| | No | 0.5 | 2.3 | 25.6 | 18.3 | 30.3 | 14.9 | 24.6 |
| eqM&eqV | Yes | 5 | 47.0 | 3.9 | 2.2 | 4.4 | 4.2 | 5.0 |
| (Type I error) | Yes | 1 | 15.2 | 0.6 | 0.3 | 0.8 | 0.8 | 0.9 |
| | Yes | 0.5 | 9.2 | 0.3 | 0.1 | 0.3 | 0.4 | 0.4 |
| diffM&eqV | Yes | 5 | 3.2 | 31.4 | 11.3 | 10.2 | 39.5 | 15.3 |
| | Yes | 1 | 0.5 | 11.2 | 3.2 | 2.6 | 19.8 | 3.9 |
| | Yes | 0.5 | 0.3 | 7.4 | 1.8 | 1.7 | 13.0 | 2.5 |
| eqM&diffV | Yes | 5 | 0.2 | 66.1 | 7.4 | 39.6 | 24.7 | 33.1 |
| | Yes | 1 | 0.1 | 30.6 | 2.2 | 19.2 | 11.6 | 15.1 |
| | Yes | 0.5 | 0.0 | 21.8 | 1.3 | 14.8 | 7.7 | 11.5 |
| diffM&diffV | Yes | 5 | 0.3 | 69.8 | 10.6 | 44.9 | 32.8 | 39.4 |
| | Yes | 1 | 0.0 | 34.6 | 3.8 | 24.3 | 16.4 | 20.2 |
| | Yes | 0.5 | 0.0 | 25.2 | 2.4 | 19.2 | 11.0 | 15.3 |

Table C.4: The empirical type I error rates (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation values were generated from Chi-square distributions without or with an outlier. The numbers of non-diseased and diseased samples are (50, 50).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 13.8 | 3.7 | 4.8 | 6.1 | 5.1 | 5.1 |
| (Type I error) | No | 1 | 5.9 | 0.5 | 0.7 | 1.3 | 1.0 | 0.9 |
| | No | 0.5 | 4.3 | 0.2 | 0.4 | 0.8 | 0.5 | 0.5 |
| diffM&eqV | No | 5 | 50.1 | 88.1 | 91.8 | 85.7 | 94.2 | 93.6 |
| | No | 1 | 21.3 | 64.9 | 78.9 | 63.4 | 82.7 | 81.5 |
| | No | 0.5 | 12.8 | 64.9 | 71.6 | 50.5 | 75.5 | 72.2 |
| eqM&diffV | No | 5 | 12.0 | 6.9 | 15.9 | 16.3 | 18.5 | 18.3 |
| | No | 1 | 3.4 | 0.9 | 5.2 | 5.3 | 6.2 | 6.3 |
| | No | 0.5 | 2.0 | 0.9 | 3.0 | 3.0 | 3.9 | 3.6 |
| diffM&diffV | No | 5 | 11.4 | 21.7 | 34.5 | 28.0 | 39.7 | 38.7 |
| | No | 1 | 3.0 | 6.1 | 16.1 | 10.5 | 19 | 18.2 |
| | No | 0.5 | 1.5 | 6.1 | 10.6 | 6.2 | 13.4 | 11.7 |
| eqM&eqV | Yes | 5 | 23.1 | 3.6 | 3.3 | 6.4 | 5.0 | 5.0 |
| (Type I error) | Yes | 1 | 12.2 | 0.5 | 0.4 | 1.4 | 0.8 | 0.8 |
| | Yes | 0.5 | 9.3 | 0.2 | 0.2 | 0.6 | 0.4 | 0.3 |
| diffM&eqV | Yes | 5 | 16.4 | 85.2 | 93.3 | 82.5 | 93.1 | 92.6 |
| | Yes | 1 | 2.6 | 59.6 | 82.7 | 57.1 | 81.5 | 79.8 |
| | Yes | 0.5 | 1.6 | 50.1 | 76.8 | 47.3 | 75.5 | 72.8 |
| eqM&diffV | Yes | 5 | 19.8 | 6.3 | 19.4 | 24.7 | 24.0 | 23.8 |
| | Yes | 1 | 5.6 | 0.7 | 4.4 | 7.9 | 7.9 | 7.5 |
| | Yes | 0.5 | 3.6 | 0.4 | 2.5 | 4.6 | 4.9 | 5.0 |
| diffM&diffV | Yes | 5 | 14.2 | 18.8 | 39.9 | 34.3 | 43.5 | 42.3 |
| | Yes | 1 | 3.9 | 4.8 | 18.0 | 13.4 | 22.3 | 21.1 |
| | Yes | 0.5 | 2.2 | 2.8 | 12.6 | 8.9 | 16.3 | 14.9 |

Table C.5: The empirical type I error rates (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation values were generated from normal distributions without or with an outlier. The numbers of non-diseased and diseased samples are (20, 20).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 5.9 | 3.3 | 5.3 | 5.4 | 5.7 | 5.8 |
| (Type I error) | No | 1 | 1.4 | 0.4 | 0.9 | 0.9 | 1.2 | 1.0 |
| | No | 0.5 | 0.8 | 0.4 | 0.4 | 0.4 | 0.5 | 0.4 |
| diffM&eqV | No | 5 | 35.5 | 29.2 | 35.2 | 33.7 | 34.1 | 33.6 |
| | No | 1 | 15.2 | 9.1 | 16.0 | 14.7 | 14.5 | 14.6 |
| | No | 0.5 | 10.9 | 9.1 | 11.8 | 9.7 | 9.8 | 10.0 |
| eqM&diffV | No | 5 | 25.6 | 4.8 | 19.8 | 19.9 | 17.9 | 18.5 |
| | No | 1 | 8.4 | 0.7 | 4.6 | 6.2 | 4.6 | 5.8 |
| | No | 0.5 | 5.4 | 0.7 | 2.6 | 3.4 | 2.6 | 3.2 |
| diffM&diffV | No | 5 | 22.3 | 11.7 | 20.6 | 20.1 | 18.5 | 19.1 |
| | No | 1 | 6.7 | 2.7 | 6.5 | 6.4 | 5.5 | 6.2 |
| | No | 0.5 | 4.2 | 2.7 | 4.3 | 3.8 | 3.7 | 4.0 |
| eqM&eqV | Yes | 5 | 26.2 | 3.1 | 2.6 | 3.5 | 3.8 | 3.7 |
| (Type I error) | Yes | 1 | 10.6 | 0.4 | 0.3 | 0.3 | 0.5 | 0.5 |
| | Yes | 0.5 | 7.2 | 0.4 | 0.1 | 0.1 | 0.2 | 0.2 |
| diffM&eqV | Yes | 5 | 22.5 | 34.7 | 45.7 | 42.4 | 42.9 | 42.4 |
| | Yes | 1 | 6.4 | 11.7 | 21.1 | 19.3 | 19.2 | 19.0 |
| | Yes | 0.5 | 3.2 | 11.7 | 15.3 | 15.1 | 12.2 | 12.6 |
| eqM&diffV | Yes | 5 | 0.6 | 4.2 | 8.2 | 9.3 | 9.9 | 10.4 |
| | Yes | 1 | 0.1 | 0.5 | 2.1 | 2.9 | 2.9 | 3.1 |
| | Yes | 0.5 | 0.1 | 0.5 | 1.3 | 2.1 | 1.8 | 1.9 |
| diffM&diffV | Yes | 5 | 3.5 | 13.9 | 23.3 | 21.1 | 22.8 | 22.7 |
| | Yes | 1 | 0.5 | 3.6 | 8.3 | 7.9 | 8.4 | 8.6 |
| | Yes | 0.5 | 0.2 | 3.6 | 5.6 | 5.8 | 5.0 | 5.3 |

Table C.6: The empirical type I error rates (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation values were generated from Beta distributions without or with an outlier. The numbers of non-diseased and diseased samples are (20, 20).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 6.2 | 3.3 | 5.4 | 5.5 | 5.5 | 5.7 |
| (Type I error) | No | 1 | 1.5 | 0.4 | 0.9 | 0.8 | 1.0 | 1.0 |
| | No | 0.5 | 0.8 | 0.4 | 0.3 | 0.3 | 0.5 | 0.4 |
| diffM&eqV | No | 5 | 36 | 30.8 | 37.4 | 35.2 | 36.3 | 35.4 |
| | No | 1 | 14.8 | 9.7 | 17.8 | 16.8 | 16.7 | 16.7 |
| | No | 0.5 | 10.1 | 9.7 | 13.4 | 12.1 | 11.1 | 11.5 |
| eqM&diffV | No | 5 | 22.9 | 4.7 | 18.0 | 18.9 | 16.8 | 17.6 |
| | No | 1 | 7.2 | 0.4 | 4.9 | 6.3 | 4.6 | 5.3 |
| | No | 0.5 | 4.7 | 0.4 | 2.6 | 3.9 | 2.6 | 2.9 |
| diffM&diffV | No | 5 | 20.9 | 14.0 | 23.0 | 21.2 | 21.3 | 21.2 |
| | No | 1 | 6.3 | 3.2 | 8.2 | 8.1 | 8.1 | 8.1 |
| | No | 0.5 | 3.7 | 3.2 | 5.5 | 5.2 | 4.9 | 5.3 |
| eqM&eqV | Yes | 5 | 23.5 | 3.3 | 2.6 | 3.6 | 4.1 | 4.0 |
| (Type I error) | Yes | 1 | 6.4 | 0.4 | 0.2 | 0.3 | 0.5 | 0.5 |
| | Yes | 0.5 | 3.4 | 0.4 | 0.1 | 0.1 | 0.2 | 0.2 |
| diffM&eqV | Yes | 5 | 32.4 | 36.3 | 46.5 | 45.0 | 43.7 | 43.3 |
| | Yes | 1 | 13.7 | 12.1 | 22.4 | 20.1 | 19.1 | 19.3 |
| | Yes | 0.5 | 7.8 | 12.1 | 16.2 | 13.7 | 13.1 | 13.5 |
| eqM&diffV | Yes | 5 | 0.3 | 4.2 | 3.8 | 4.7 | 5.4 | 5.2 |
| | Yes | 1 | 0.0 | 0.3 | 1.1 | 1.3 | 1.6 | 1.7 |
| | Yes | 0.5 | 0.0 | 0.3 | 0.5 | 0.8 | 1.1 | 1.1 |
| diffM&diffV | Yes | 5 | 1.6 | 9.3 | 8.4 | 3.4 | 3.7 | 3.6 |
| | Yes | 1 | 0.3 | 1.8 | 3.1 | 0.9 | 1.2 | 1.1 |
| | Yes | 0.5 | 0.1 | 1.8 | 2.1 | 0.5 | 0.6 | 0.6 |

Table C.7: The empirical type I error rates (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation values were generated from mixtures of two normal distributions without or with an outlier. The numbers of non-diseased and diseased samples are (20, 20).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 3.4 | 3.3 | 4.5 | 8.1 | 5.1 | 6.7 |
| (Type I error) | No | 1 | 0.8 | 0.5 | 0.8 | 2.4 | 1.3 | 1.6 |
| | No | 0.5 | 0.5 | 0.5 | 0.4 | 1.3 | 0.6 | 0.9 |
| diffM&eqV | No | 5 | 5.8 | 11 | 17.7 | 10.8 | 17.5 | 12.6 |
| | No | 1 | 1.3 | 2 | 5.2 | 2.0 | 4.9 | 2.8 |
| | No | 0.5 | 0.7 | 2 | 3.1 | 0.9 | 2.4 | 1.1 |
| eqM&diffV | No | 5 | 11.2 | 24.4 | 26.2 | 28.5 | 23.1 | 27.3 |
| | No | 1 | 2.1 | 5.5 | 13.8 | 11.2 | 10 | 12.2 |
| | No | 0.5 | 0.9 | 5.5 | 10.2 | 6.3 | 6.9 | 7.5 |
| diffM&diffV | No | 5 | 11.8 | 25.4 | 29.1 | 31.1 | 25.9 | 30.4 |
| | No | 1 | 1.9 | 6.8 | 15.8 | 11.9 | 10.8 | 13.3 |
| | No | 0.5 | 0.8 | 6.8 | 11.5 | 6.2 | 7.2 | 8.1 |
| eqM&eqV | Yes | 5 | 60.5 | 3.4 | 1.9 | 2.9 | 2.2 | 2.3 |
| (Type I error) | Yes | 1 | 23.6 | 0.5 | 0.2 | 0.5 | 0.3 | 0.3 |
| | Yes | 0.5 | 14.2 | 0.5 | 0.1 | 0.2 | 0.1 | 0.1 |
| diffM&eqV | Yes | 5 | 2.7 | 12.7 | 9.0 | 7.0 | 17.3 | 10.4 |
| | Yes | 1 | 0.7 | 2.8 | 2.2 | 2.3 | 6.9 | 3.4 |
| | Yes | 0.5 | 0.2 | 2.8 | 1.2 | 1.2 | 4.0 | 2.0 |
| eqM&diffV | Yes | 5 | 0.3 | 22.8 | 8.4 | 13.6 | 16.0 | 16.4 |
| | Yes | 1 | 0.0 | 5.6 | 2.3 | 5.6 | 7.3 | 7.4 |
| | Yes | 0.5 | 0.0 | 5.6 | 1.6 | 3.5 | 5.0 | 5.1 |
| diffM&diffV | Yes | 5 | 0.7 | 26 | 10.4 | 16.3 | 19.7 | 20.3 |
| | Yes | 1 | 0.1 | 7.2 | 3.8 | 6.6 | 9.7 | 9.6 |
| | Yes | 0.5 | 0.1 | 7.2 | 2.6 | 4.2 | 6.9 | 6.7 |

Table C.8: The empirical type I error rates (×100) and power (×100) for the six tests evaluated at 5%, 1%, and 0.5% significance levels when methylation values were generated from Chi-square distributions without or with an outlier. The numbers of non-diseased and diseased samples are (20, 20).

| Scenarios | Out | Level(%) | jointLRT | KS | AW | iAW.Lev | iAW.BF | iAW.TM |
|---|---|---|---|---|---|---|---|---|
| eqM&eqV | No | 5 | 13.4 | 3.2 | 5.3 | 6.4 | 5.3 | 5.5 |
| (Type I error) | No | 1 | 5.2 | 0.4 | 0.9 | 1.3 | 1.0 | 0.9 |
| | No | 0.5 | 3.7 | 0.4 | 0.4 | 0.6 | 0.5 | 0.5 |
| diffM&eqV | No | 5 | 23.9 | 45.3 | 52.4 | 41.0 | 55.7 | 52.8 |
| | No | 1 | 7.4 | 17.0 | 26.6 | 17.5 | 30.4 | 28.3 |
| | No | 0.5 | 4.6 | 17 | 19.4 | 10.6 | 22.6 | 19.2 |
| eqM&diffV | No | 5 | 7.7 | 3.9 | 8.7 | 8.8 | 9.2 | 9.2 |
| | No | 1 | 2.0 | 0.6 | 1.7 | 2.2 | 2.2 | 2.4 |
| | No | 0.5 | 1.2 | 0.6 | 1.0 | 1.1 | 1.1 | 1.2 |
| diffM&diffV | No | 5 | 8.5 | 9.8 | 15.5 | 13.0 | 18.5 | 16.9 |
| | No | 1 | 2.0 | 1.6 | 4.3 | 3.5 | 6.2 | 5.7 |
| | No | 0.5 | 1.1 | 1.6 | 2.5 | 1.5 | 3.7 | 2.7 |
| eqM&eqV | Yes | 5 | 29.7 | 3.0 | 2.7 | 5.9 | 4.0 | 3.8 |
| (Type I error) | Yes | 1 | 16.1 | 0.4 | 0.3 | 0.7 | 0.7 | 0.5 |
| | Yes | 0.5 | 12.3 | 0.4 | 0.1 | 0.2 | 0.3 | 0.2 |
| diffM&eqV | Yes | 5 | 4.0 | 35.2 | 55.9 | 36.1 | 52.9 | 52.3 |
| | Yes | 1 | 0.8 | 11.2 | 30.9 | 17.8 | 29.3 | 29.8 |
| | Yes | 0.5 | 0.2 | 11.2 | 23.8 | 12.1 | 20.6 | 22.1 |
| eqM&diffV | Yes | 5 | 11.9 | 3.6 | 9.1 | 13.3 | 11 | 12.2 |
| | Yes | 1 | 3.6 | 0.4 | 2.1 | 3.9 | 2.3 | 2.9 |
| | Yes | 0.5 | 2.0 | 0.4 | 1.0 | 2.1 | 1.0 | 1.3 |
| diffM&diffV | Yes | 5 | 9.2 | 6.8 | 18.2 | 17.0 | 20.3 | 20.7 |
| | Yes | 1 | 2.5 | 0.9 | 5.4 | 5.5 | 6.7 | 7.2 |
| | Yes | 0.5 | 1.3 | 0.9 | 3.4 | 3.3 | 3.7 | 4.0 |

# D Parameter estimation in Chapter 5

In the first component, the miRNAs are over-variated in diseased samples, i.e. $\sigma_{1c}^2 > \sigma_{1n}^2$. In the third component, the miRNAs are under-variated in diseased samples, i.e. $\sigma_{3c}^2 < \sigma_{3n}^2$. Our prior belief is that the majority of miRNAs are usually non-differentially variated, so we assume $\pi_2 > \pi_1$ and $\pi_2 > \pi_3$. We re-parameterized variances as

$$s_k = \log(\sigma_k^2), \quad k = 1c, 2, 3c, 1n, 3n. \tag{D.1}$$

We re-parameterized variances again to make sure $\sigma_{1c}^2 > \sigma_{1n}^2$ and $\sigma_{3c}^2 < \sigma_{3n}^2$:

$$\begin{aligned} s_{1n} =& s_{1c} - \exp(\triangle_{1n}), \\ s_{3n} =& s_{3c} + \exp(\triangle_{3n}). \end{aligned} \tag{D.2}$$

To make sure the covariance matrix is positive definite, the correlation $\rho$ should satisfies the condition

$$-\frac{1}{m-1} < \rho < 1.$$

So we re-parameterized the correlation parameter

$$\rho = \frac{\exp(r) - 1/(m-1)}{1+\exp(r)}. \tag{D.3}$$

Then the log density for miRNAs in the first cluster:

$$\begin{aligned}
\log\left[f_1(\boldsymbol{x})|\boldsymbol{\theta}_1\right] =& -\frac{m_c}{2}\log(2\pi) - \frac{m_c}{2}s_{1c} - \frac{m_c}{2}\log(m_c) + \frac{m_c-1}{2}\log(m_c-1) + \frac{m_c}{2}\log\left[1+\exp(r_{1c})\right] - \frac{r_{1c}}{2} \\
&- \frac{\left[\boldsymbol{a}\left(\boldsymbol{x}_{1c},\mu_{1c}\right)\right]^T\left[\boldsymbol{a}\left(\boldsymbol{x}_{1c},\mu_{1c}\right)\right]}{2\exp(s_{1c})}\frac{m_c-1}{m_c}\left[1+\exp(r_{1c})\right] \\
&+ \frac{\left(\left[\boldsymbol{a}\left(\boldsymbol{x}_{1c},\mu_{1c}\right)\right]^T\mathbf{1}\right)^2\left[(m_c-2)+\exp(r_{1c})(m_c-1)-\frac{1}{\exp(r_{1c})}\right]}{2\exp(s_{1c})m_c^2} \\
&- \frac{m_n}{2}\log(2\pi) - \frac{m_n}{2}\left[s_{1c}-\exp(\triangle_{1n})\right] - \frac{m_n}{2}\log(m_n) + \frac{m_n-1}{2}\log(m_n-1) + \frac{m_n}{2}\log\left[1+\exp(r_{1n})\right] - \frac{r_{1n}}{2} \\
&- \frac{\left[\boldsymbol{a}\left(\boldsymbol{x}_{1n},\mu_{1n}\right)\right]^T\left[\boldsymbol{a}\left(\boldsymbol{x}_{1n},\mu_{1n}\right)\right]}{2\exp([s_{1c}-\exp(\triangle_{1n})])}\frac{m_n-1}{m_n}\left[1+\exp(r_{1n})\right] \\
&+ \frac{\left(\left[\boldsymbol{a}\left(\boldsymbol{x}_{1n},\mu_{1n}\right)\right]^T\mathbf{1}\right)^2\left[(m_n-2)+\exp(r_{1n})(m_n-1)-\frac{1}{\exp(r_{1n})}\right]}{2\exp([s_{1c}-\exp(\triangle_{1n})])m_n^2},
\end{aligned} \tag{D.4}$$

where

$$\boldsymbol{a}(\boldsymbol{x},\mu)^T\boldsymbol{a}(\boldsymbol{x},\mu) = \boldsymbol{x}^T\boldsymbol{x} - 2\mu\mathbf{1}^T\boldsymbol{x} + n\mu^2,$$

$$\left(\boldsymbol{a}(\boldsymbol{x},\mu)^T\mathbf{1}\right)^2 = \left(\mathbf{1}^T\boldsymbol{x}\right)^2 + n^2\mu^2 - 2n\mu\boldsymbol{x}^T\mathbf{1}.$$

The log density for miRNAs in the second cluster:

$$\begin{aligned}
\log\left[f_2(\boldsymbol{x})|\boldsymbol{\theta}_2\right] =& -\frac{m_c}{2}\log(2\pi) - \frac{m_c}{2}s_2 - \frac{m_c}{2}\log(m_c) + \frac{m_c-1}{2}\log(m_c-1) + \frac{m_c}{2}\log\left[1+\exp(r_{2c})\right] - \frac{r_{2c}}{2} \\
&- \frac{\left[\boldsymbol{a}\left(\boldsymbol{x}_{1c},\mu_{2c}\right)\right]^T\left[\boldsymbol{a}\left(\boldsymbol{x}_{1c},\mu_{2c}\right)\right]}{2\exp(s_2)}\frac{m_c-1}{m_c}\left[1+\exp(r_{2c})\right] \\
&+ \frac{\left(\left[\boldsymbol{a}\left(\boldsymbol{x}_{1c},\mu_{2c}\right)\right]^T\mathbf{1}\right)^2\left[(m_c-2)+\exp(r_{2c})(m_c-1)-\frac{1}{\exp(r_{2c})}\right]}{2\exp(s_2)m_c^2} \\
&- \frac{m_n}{2}\log(2\pi) - \frac{m_n}{2}\left[s_2\right] - \frac{m_n}{2}\log(m_n) + \frac{m_n-1}{2}\log(m_n-1) + \frac{m_n}{2}\log\left[1+\exp(r_{2n})\right] - \frac{r_{2n}}{2} \\
&- \frac{\left[\boldsymbol{a}\left(\boldsymbol{x}_{1n},\mu_{2n}\right)\right]^T\left[\boldsymbol{a}\left(\boldsymbol{x}_{1n},\mu_{2n}\right)\right]}{2\exp([s_2])}\frac{m_n-1}{m_n}\left[1+\exp(r_{2n})\right] \\
&+ \frac{\left(\left[\boldsymbol{a}\left(\boldsymbol{x}_{1n},\mu_{2n}\right)\right]^T\mathbf{1}\right)^2\left[(m_n-2)+\exp(r_{2n})(m_n-1)-\frac{1}{\exp(r_{2n})}\right]}{2\exp([s_2])m_n^2}.
\end{aligned} \tag{D.5}$$

The log density for miRNAs in the third cluster:

$$
\begin{aligned}
\log\left[f_3(\boldsymbol{x})|\boldsymbol{\theta}_3\right] = & -\frac{m_c}{2}\log(2\pi) - \frac{m_c}{2}s_{3c} - \frac{m_c}{2}\log(m_c) + \frac{m_c-1}{2}\log(m_c-1) + \frac{m_c}{2}\log\left[1+\exp(r_{3c})\right] - \frac{r_{3c}}{2} \\
& - \frac{\left[\boldsymbol{a}\left(\boldsymbol{x}_{3c},\mu_{3c}\right)\right]^T\left[\boldsymbol{a}\left(\boldsymbol{x}_{3c},\mu_{3c}\right)\right]}{2\exp(s_{3c})}\frac{m_c-1}{m_c}\left[1+\exp(r_{3c})\right] \\
& + \frac{\left(\left[\boldsymbol{a}\left(\boldsymbol{x}_{3c},\mu_{3c}\right)\right]^T\mathbf{1}\right)^2\left[(m_c-2)+\exp(r_{3c})(m_c-1)-\frac{1}{\exp(r_{3c})}\right]}{2\exp(s_{3c})m_c^2} \\
& - \frac{m_n}{2}\log(2\pi) - \frac{m_n}{2}\left[s_{3c}+\exp(\triangle_{3n})\right] - \frac{m_n}{2}\log(m_n) + \frac{m_n-1}{2}\log(m_n-1) + \frac{m_n}{2}\log\left[1+\exp(r_{3n})\right] - \frac{r_{3n}}{2} \\
& - \frac{\left[\boldsymbol{a}\left(\boldsymbol{x}_{3n},\mu_{3n}\right)\right]^T\left[\boldsymbol{a}\left(\boldsymbol{x}_{3n},\mu_{3n}\right)\right]}{2\exp([s_{3c}+\exp(\triangle_{3n})])}\frac{m_n-1}{m_n}\left[1+\exp(r_{3n})\right] \\
& + \frac{\left(\left[\boldsymbol{a}\left(\boldsymbol{x}_{3n},\mu_{3n}\right)\right]^T\mathbf{1}\right)^2\left[(m_n-2)+\exp(r_{3n})(m_n-1)-\frac{1}{\exp(r_{3n})}\right]}{2\exp([s_{3c}+\exp(\triangle_{3n})])m_n^2}.
\end{aligned}
\tag{D.6}
$$

## EM algorithm for updating the parameter sets

After the re-parameterizations, the likelihood for the complete data can be written as

$$
g(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_p|\boldsymbol{\Psi}') = \prod_{j=1}^{p}[\pi_1^{z_{1j}}f_1(\boldsymbol{x}_j|\boldsymbol{\theta}_1')^{z_{1j}}][\pi_2^{z_{2j}}f_2(\boldsymbol{x}_j|\boldsymbol{\theta}_2')^{z_{2j}}][\pi_3^{1-z_{1j}-z_{2j}}f_3(\boldsymbol{x}_j|\boldsymbol{\theta}_3')^{1-z_{1j}-z_{2j}}].
$$

And the log complete likelihood function is

$$
\ell_0(\boldsymbol{\Psi}') = \sum_{j=1}^{p}\boldsymbol{z}_j^T\boldsymbol{V}(\boldsymbol{\pi}) + \sum_{j=1}^{p}\boldsymbol{z}_j^T\boldsymbol{U}_j(\boldsymbol{\theta}'),
$$

where

$$\begin{aligned}
\boldsymbol{z}_j =&(z_{1j}, z_{2j}, 1 - z_{1j} - z_{2j})^T,\\
\boldsymbol{V}(\boldsymbol{\pi}) =&(\log(\pi_1), \log(\pi_2), \log(1 - \pi_1 - \pi_2))^T,\\
\boldsymbol{U}_j(\boldsymbol{\theta}') =&(\log(f_1(\boldsymbol{x}_j|\boldsymbol{\theta}'_1)), \log(f_2(\boldsymbol{x}_j|\boldsymbol{\theta}'_2)), \log(f_3(\boldsymbol{x}_j|\boldsymbol{\theta}'_3)))^T,\\
\boldsymbol{\theta}'_1 =&(s_{1c}, r_{1c}, \mu_{1c}, \triangle_{1n}, r_{1n}, \mu_{1n})^T,\\
\boldsymbol{\theta}'_2 =&(s_2, r_{2c}, \mu_{2c}, r_{2n}, \mu_{2n})^T,\\
\boldsymbol{\theta}'_3 =&(s_{3c}, r_{3c}, \mu_{3c}, \triangle_{3n}, r_{3n}, \mu_{3n})^T,\\
\boldsymbol{\Psi}' =&(\pi_1, \pi_2, s_{1c}, r_{1c}, \mu_{1c}, \triangle_{1n}, r_{1n}, \mu_{1n}, s_2, r_{2c}, \mu_{2c}, r_{2n}, \mu_{2n}, s_{3c}, r_{3c}, \mu_{3c}, \triangle_{3n}, r_{3n}, \mu_{3n})^T.
\end{aligned}$$

EM algorithm is initiated from $\boldsymbol{\Psi}'^{(0)}$ and generates a sequence of estimates $\{\boldsymbol{\Psi}'^{(t)}\}$. Each iteration consists of the following two step:

**E-step:** Evaluate $Q\left(\boldsymbol{\Psi}', \boldsymbol{\Psi}'^{(t)}\right) = \mathrm{E}\left[\log(g(\boldsymbol{y}|\boldsymbol{\Psi}')|\boldsymbol{x}, \boldsymbol{\Psi}'(t)\right]$.

**M-step:** Find $\boldsymbol{\Psi}' = \boldsymbol{\Psi}'^{(t+1)}$ to maximize $Q\left(\boldsymbol{\Psi}', \boldsymbol{\Psi}'^{(t)}\right)$.

To stabilize the estimates of $\pi_k$, $k = 1, 2, 3$, we assume that the vector of mixing proportions $(\pi_1, \pi_2, \pi_3)^T$ are Dirichlet distributed with parameters $b_1 = b_2 = b_3 = 3$.

**E-step:**

We can obtain

$$Q(\boldsymbol{\Psi}', \boldsymbol{\Psi}'^{(t)}) = \sum_{j=1}^{p}[\mathrm{E}\left(\boldsymbol{z}_j|\boldsymbol{x}, \boldsymbol{\Psi}'^{(t)}\right)]^T\boldsymbol{V}(\boldsymbol{\pi}) + \sum_{j=1}^{p}[\mathrm{E}\left(\boldsymbol{z}_j|\boldsymbol{x}, \boldsymbol{\Psi}'^{(t)}\right)]^T\boldsymbol{U}(\boldsymbol{\theta}').$$

Denote

$$\boldsymbol{w}_j(\boldsymbol{\Psi}'^{(t)}) = \mathrm{E}\,(\boldsymbol{z}_j|\boldsymbol{x}, \boldsymbol{\Psi}'^{(t)}) = \mathrm{E}\,(\boldsymbol{z}_j|\boldsymbol{x}_j, \boldsymbol{\Psi}'^{(t)}).$$

The last equality is because of the independence of data points. The $k$-th element of $\boldsymbol{w}_j(\boldsymbol{\Psi}'^{(t)})$ is

$$\begin{aligned} w_{jk}(\boldsymbol{\Psi}'^{(t)}) =& \mathrm{E}\,(z_{jk}|\boldsymbol{x}_j, \boldsymbol{\Psi}'^{(t)}) \\ =& Pr(z_{jk} = 1|\boldsymbol{x}_j, \boldsymbol{\Psi}'^{(t)}) \\ =& \frac{Pr(\boldsymbol{x}_j|z_{jk} = 1, \boldsymbol{\Psi}'^{(t)}) Pr(z_{jk} = 1|\boldsymbol{\Psi}'^{(t)})}{f(\boldsymbol{x}_j|\boldsymbol{\Psi}'^{(t)})} \\ =& \frac{f_k(\boldsymbol{x}_j|\boldsymbol{\theta}_j'^{(t)})\pi_k^{(t)}}{f(\boldsymbol{x}_j|\boldsymbol{\Psi}'^{(t)})}, \end{aligned}$$

where

$$f(\boldsymbol{x}_j|\boldsymbol{\Psi}'^{(t)}) = f_1(\boldsymbol{x}_j|\boldsymbol{\theta}_1'^{(t)})\pi_1^{(t)} + f_2(\boldsymbol{x}_j|\boldsymbol{\theta}_2'^{(t)})\pi_2^{(t)} + f_3(\boldsymbol{x}_j|\boldsymbol{\theta}_3'^{(t)})[1 - \pi_1^{(t)} - \pi_2^{(t)}].$$

These "weights" $(w_{jk}(\boldsymbol{\Psi}'^{(t)}),\ j = 1, \ldots, p,\ k = 1, 2)$ are therefore the probabilities of category membership for the $j$-th miRNA, conditional on $x_j$ and given that the parameter is $\boldsymbol{\Psi}'^{(t)}$.

$$
\begin{aligned}
Q(\boldsymbol{\Psi}', \boldsymbol{\Psi}'^{(t)}) =& \sum_{j=1}^{p} \Big\{ \mathrm{E}\left(z_{j1}|\boldsymbol{x}_j, \boldsymbol{\Psi}'^{(t)}\right) \log(\pi_1) + \mathrm{E}\left(z_{j2}|\boldsymbol{x}_j, \boldsymbol{\Psi}'^{(t)}\right) \log(\pi_2) \\
&+ \mathrm{E}\left(1 - z_{j1} - z_{j2}|\boldsymbol{x}_j, \boldsymbol{\Psi}'^{(t)}\right) \log(\pi_3) \\
&+ \mathrm{E}\left(z_{j1}|\boldsymbol{x}_j, \boldsymbol{\Psi}'^{(t)}\right) \log(f_1(\boldsymbol{x}_j)) + \mathrm{E}\left(z_{j2}|\boldsymbol{x}_j, \boldsymbol{\Psi}'^{(t)}\right) \log(f_2(\boldsymbol{x}_j)) \\
&+ \mathrm{E}\left(1 - z_{j1} - z_{j2}|\boldsymbol{x}_j, \boldsymbol{\Psi}'^{(t)}\right) \log(f_3(\boldsymbol{x}_j)) \Big\} \\
=& \sum_{j=1}^{p} \Big\{ w_{j1}\left(\boldsymbol{\Psi}'^{(t)}\right) \log(\pi_1) + w_{j2}\left(\boldsymbol{\Psi}'^{(t)}\right) \log(\pi_2) \\
&+ \left[1 - w_{j1}\left(\boldsymbol{\Psi}'^{(t)}\right) - w_{j2}\left(\boldsymbol{\Psi}'^{(t)}\right)\right] \log(1 - \pi_1 - \pi_2) \\
&+ w_{j1}\left(\boldsymbol{\Psi}'^{(t)}\right) \log(f_1(\boldsymbol{x}_j)) + w_{j2}\left(\boldsymbol{\Psi}'^{(t)}\right) \log(f_2(\boldsymbol{x}_j)) \\
&+ \left[1 - w_{j1}\left(\boldsymbol{\Psi}'^{(t)}\right) - w_{j2}\left(\boldsymbol{\Psi}'^{(t)}\right)\right] \log(f_3(\boldsymbol{x}_j)) \Big\}.
\end{aligned}
\tag{D.7}
$$

**M-step:**

We need to maximize the following function

$$
\begin{aligned}
Q(\boldsymbol{\Psi}'|\boldsymbol{\Psi}'^{(t)}) =& \log[\frac{\Gamma(b_1 + b_2 + b_3)}{\Gamma(b_1)\Gamma(b_2)\Gamma(b_3)}] + [w_1^{(t)} + (b_1 - 1)] \log(\pi_1) \\
&+ [w_2^{(t)} + (b_2 - 1)] \log(\pi_2) + [w_3^{(t)} + (b_3 - 1)] \log(1 - \pi_1 - \pi_2) \\
&+ \sum_{k=1}^{3} \sum_{j=1}^{p} w_{jk}^{(t)} \log[f_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k')],
\end{aligned}
$$

where

$$
\begin{aligned}
w_k^{(t)} =& \sum_{j=1}^{p} w_{jk}^{(t)}, \\
w_{jk}^{(t)} =& \frac{\pi_k f_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k'^{(t)})}{\sum_{k=1}^{3} \pi_k f_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k'^{(t)})},
\end{aligned}
$$

and the superscript $(t)$ indicates the number of iterations. Let the first derivative be zero, we can get

$$\pi_k^{(t+1)} = \frac{w_k^{(t)} + (b_k - 1)}{p + b_1 + b_2 + b_3 - 3}. \tag{D.8}$$

Denote

$$Q_k(\boldsymbol{\Psi}'|\boldsymbol{\Psi}'^{(t)}) = \sum_{j=1}^{p} w_{jk}^{(t)} \log[f_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k')].$$

The first derivatives are:

$$\frac{\partial Q_k}{\partial \boldsymbol{\theta}_k'} = \sum_{j=1}^{p} w_{jk}^{(t)} \frac{\partial \log[f_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k')]}{\partial \boldsymbol{\theta}_k'}, \quad k = 1, 2, 3.$$

We solve the above equations $\frac{\partial Q_k}{\partial \boldsymbol{\theta}_k'} = \mathbf{0}$ to obtain the update $\boldsymbol{\Psi}'^{(t+1)}$.

# E Additional simulation results in Chapter 5

Table E.1: The p-values of two-sided Wilcoxon signed-rank tests to evaluate whether the differences of median Jaccard indices obtained by the gs method and by each of other methods are significant.

| Method | simI | simII | simIII | simIV |
|---|---|---|---|---|
| L2N.pp | 3.96E-18 | 3.96E-18 | 3.96E-18 | 3.96E-18 |
| L2N.q | 3.96E-18 | 3.96E-18 | 3.96E-18 | 3.96E-18 |
| L2N.f | 3.96E-18 | 3.96E-18 | 3.96E-18 | 3.96E-18 |
| N3.pp | 3.96E-18 | 3.96E-18 | 3.96E-18 | 3.96E-18 |
| N3.q | 3.96E-18 | 3.96E-18 | 3.96E-18 | 3.96E-18 |
| N3.f | 3.96E-18 | 3.96E-18 | 3.96E-18 | 3.96E-18 |

Table E.2: The p-values of two-sided Wilcoxon signed-rank tests to evaluate whether the differences of median FPR obtained by the gs method and by each of other methods are significant.

| Method | simI | simII | simIII | simIV |
|---|---|---|---|---|
| L2N.pp | 8.72E-18 | 3.96E-18 | 2.02E-16 | 3.96E-18 |
| L2N.q | 5.72E-18 | 3.95E-18 | 2.27E-06 | 3.96E-18 |
| L2N.f | 2.24E-17 | 3.96E-18 | 7.16E-02 | 3.96E-18 |
| N3.pp | 3.94E-18 | 3.96E-18 | 9.74E-18 | 3.96E-18 |
| N3.q | 3.94E-18 | 3.95E-18 | 2.66E-16 | 3.95E-18 |
| N3.f | 3.92E-18 | 3.95E-18 | 1.14E-11 | 3.95E-18 |

Table E.3: The p-values of two-sided Wilcoxon signed-rank tests to evaluate whether the differences of median FNR obtained by the gs method and by each of other methods are significant.

| Method | simI | simII | simIII | simIV |
|---|---|---|---|---|
| L2N.pp | 3.94E-18 | 3.95E-18 | 3.95E-18 | 3.95E-18 |
| L2N.q | 3.95E-18 | 3.95E-18 | 3.95E-18 | 3.95E-18 |
| L2N.f | 3.95E-18 | 3.95E-18 | 3.95E-18 | 3.95E-18 |
| N3.pp | 3.93E-18 | 3.95E-18 | 3.95E-18 | 3.95E-18 |
| N3.q | 3.94E-18 | 3.95E-18 | 3.95E-18 | 3.95E-18 |
| N3.f | 3.95E-18 | 3.95E-18 | 3.95E-18 | 3.95E-18 |

Table E.4: The p-values of two-sided Wilcoxon signed-rank tests to evaluate whether the differences of median proportion of validation obtained by the gs method and by each of other methods are significant based on 100 bootstrap samples of real data.

| L2N.pp | L2N.q | L2N.f | N3.pp | N3.q | N3.f |
|---|---|---|---|---|---|
| 3.96E-18 | 4.33E-06 | 8.57E-08 | 3.96E-18 | 5.02E-08 | 2.98E-08 |