SCISSOR FOR FINDING OUTLIERS IN RNA-SEQ

Hyo Young Choi

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

Chapel Hill
2018

Approved by:

J. S. Marron

D. Neil Hayes

Yufeng Liu

Michael I. Love

Andrew Nobel

ABSTRACT

Hyo Young Choi: Scissor for finding outliers in RNA-seq
(Under the direction of J. S. Marron and D. Neil Hayes)

The impressive progress of high-throughput technologies has provided many interesting modern data types, which has tremendously increased the demand for Statistics. RNA-seq, in particular, allows a rich characterization of the genome with many exciting applications. This dissertation makes contributions to RNA-seq data analysis by addressing several statistical challenges especially characterized by high dimensionality.

The dissertation is composed of two major parts. The first part concerns the issue of high dimensional outliers which are challenging to distinguish from inliers due to the special structure of high dimensional space. We introduce a new notion of high dimensional outliers that embraces various types and provides deep insights into understanding the behavior of these outliers based on several asymptotic regimes. Using this new framework, we develop an outlier detection method called Scissor that aims to identify sample outliers with distinct forms or patterns of transcripts across RNA-seq cohorts. Scissor offers a novel approach to unsupervised screening of a variety of shape changes that are possibly associated with important genetic events. Scissor has been implemented in **R** and is available online.

The second part is motivated by a challenge raised by an application of PCA to RNA-seq data. A fundamental question using PCA is how many principal components are effective for reducing dimensions. Although several algorithms have been developed to address this question, it has been observed that these algorithms may not be appropriate for RNA-seq data due to its abnormal noise structure. We propose a new algorithm for determining an effective number of principal components in RNA-seq data assuming a flexible noise structure based on some fundamental results in random matrix theory. The proposed method also provides a visualization tool for assessing the

noise assumption. This methodology has been successful in offering more reasonable numbers of principal components for RNA-seq data and implemented in Scissor.

To 두열, who always believes in me;
To my family, for their consistent support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Through advancements in terms of computing speed, storage capability, and data-collection technologies, the scope of Data Science has profoundly broadened. An immeasurable amount of data have been generated at an unprecedented speed and a variety of new data types have emerged from numerous platforms such as social media sites, Internet of Things (IoT), and scientific research institutions. The emergence of such massive data, or Big Data, and new data structures has tremendously increased the demand of Statistics.

An important type of Big Data is often characterized by a large number of variables in the thousands, millions and more with only tens or hundreds of observations (large d, small n). In statistical terminologies, we often call such a data set "high dimensional" or "large dimensional". Among many statistical challenges raised by Big Data, high dimensionality is a major issue. In traditional data analysis, it is assumed that data consist of many observations and a relatively few variables (large n, small d). The asymptotic assumption of increasing n with a fixed d leads to nice theoretical properties such as consistency, efficiency, and asymptotic normality. However, such theories do not provide useful insights for large d data sets. For example, a sample covariance matrix is not a consistent estimator for the true one in the case of an increasing d even when n is large. This is mainly because the number of parameters to be estimated ($\frac{d(d+1)}{2}$) is much larger than the number of observations (nd) (See Section 1.1).

In situations where the classical large n, small d statistical theory does not give useful insight, methods based on that cannot be expected to work effectively. Many conventional statistics based on a sample covariance matrix, such as the sphericity tests based on the uniformity of the sample eigenvalues, are not consistent with an increasing d for example. To appropriately handle the issue of

dimensionality, tremendous efforts have been made in both theoretical and methodological statistics for the last few decades.

This dissertation makes further contributions to the high dimensional data analysis motivated by challenges arising from modern genomic data types. We explore several theoretical aspects of high dimensional data in some particular situations where regular assumptions are violated and develop new statistical tools based on the theories investigated.

1.1 High-dimensional asymptotics

The sample covariance matrix is one of the fundamental objects in multivariate data analysis. When the sample size tends to infinity and the dimension size is fixed, the sample covariance matrix can be used as a good estimate of the population covariance matrix. However, when the dimension size is large, the estimation of the population covariance matrix based on the sample covariance matrix can become very poor. To appreciate this issue, let $\mathbf{X}$ be a $d \times n$ data matrix whose columns are random vectors from $N_d(0, I_d)$ and denote a version of its sample covariance matrix by $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$. Note that we do not subtract the sample mean vector because the population mean is zero. In the classical domain when n tends to infinity and d is fixed, it is expected that the eigenvalues of $\mathbf{S}_n$ will be close to 1, which is the eigenvalue of the population covariance matrix, I_d, which is a sense that $\mathbf{S}_n$ is a good estimate of I_d. Under different conditions, however, the empirical distribution of the eigenvalues of $\mathbf{S}_n$ converges to the Marcenko-Pastur (M-P) distribution (Marčenko and Pastur, 1967), when d and n both tend to infinity such that $\frac{d}{n} \to c$, as illustrated in Figure 1.1. The histogram illustrates the distribution of sample eigenvalues of $\mathbf{S}_n$ with $d = 500$ and $n = 1000$. The curve indicates the theoretical M-P distribution with $c = 0.5$. The distribution strongly deviates from the distribution of the population eigenvalue, which is a point mass at 1. This supports the idea that the sample covariance matrix is not a good estimate of the population covariance matrix for large dimensions.

When we consider a different asymptotic domain, the limiting distribution of sample eigenvalues is different. For example, if $\frac{d}{n} \to \infty$ as d and n both grow, the limiting distribution of non-zero

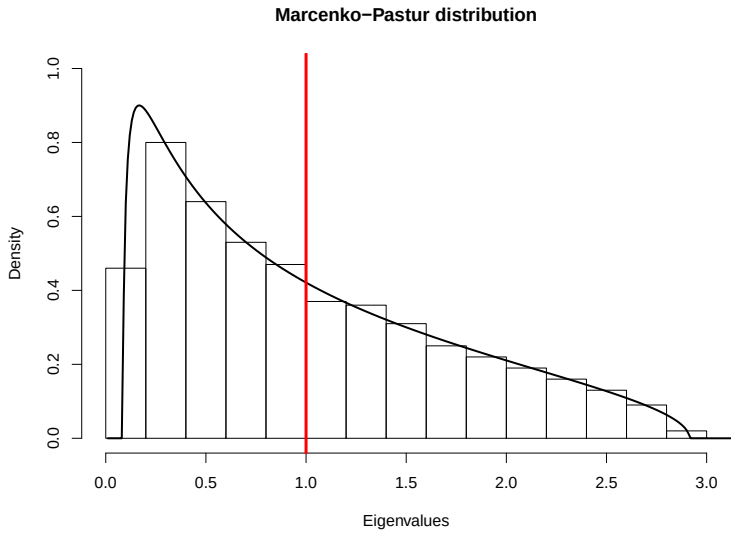


Figure 1.1: The histogram shows the empirical distribution of eigenvalues of $\mathbf{S}_n$ with $d = 500$ and $n = 1000$. The red vertical line indicates the distribution of the population eigenvalue, which is a point mass at 1. The curve shows the theoretical M-P distribution with an index $c = 0.5$. The empirical distribution well follows the theoretical M-P distribution whereas both distributions are strongly deviated from the population eigenvalue distribution.

sample eigenvalues is known to be a semi-circular law with an extreme point mass at zero. If $\frac{d}{n} \to 0$ as d and n both increase, the sample eigenvalues converge to one while the limiting distribution with suitable rescaling results in a semi-circular law. Thus, different asymptotic domains yield different limiting properties. Among various asymptotic regimes, this dissertation spotlights three asymptotic regimes:

- **Classical** asymptotic regime considers an increasing n ($n \to \infty$) and fixed d.
- **Random matrix theory (RMT)** asymptotic regime considers proportionally increasing n and d, i.e. $n \to \infty$ and $d \to \infty$ such that $\frac{d}{n} \to c$. Here, c is mostly assumed to be a constant within $0 < c < \infty$.
- **High dimensional low sample size (HDLSS)** asymptotic regime considers an increasing d ($d \to \infty$) with n being fixed.

A data set is often thought of as a collection of observations generated as a linear combination of a limited number of signals plus noise. This formulation allows us to confine our view to estimation

of such signals instead of the full recovery of a vast number of parameters. Obviously, this advantage is significantly augmented when dealing with large dimensional data sets. In this formulation, each observation vector X_j can be modeled as

$$X_j = \mu + Ay_j + \varepsilon_j \tag{1.1.1}$$

where μ is a mean vector, A is a $d \times K$ matrix representing source signals in its columns, y_j is an K-dimensional random vector, and ε_j is a d-dimensional vector of noise. Recent work based on the model (1.1.1) includes Kritchman and Nadler (2008); Passemier and Yao (2012); Ma et al. (2013); Shabalin and Nobel (2013); Choi et al. (2014); Yao et al. (2015); Fan and Wang (2015). For instance, in a signal detection model, X_j can be a vector of the recorded signals with noise at a certain time, where the columns of A are K unknown source signals, and the y_j's are emission levels of these signals. See e.g. Section 11.6 of Yao et al. (2015). In econometrics, X_j can be the returns of stocks at a certain time, with A being a matrix of latent common factors where the y_j's are unobservable random factors (Onatski, 2012; Ma et al., 2013; Fan and Wang, 2015). From now on, without loss of generality, we assume that the mean vector is zero, i.e. $\mu = 0$.

In many related works, the d-dimensional noise vector ε_j in (1.1.1) is modeled by $\varepsilon_j = \sigma z_j$ where $\sigma > 0$ is the noise level and z_j is a d-dimensional vector of white noise. Also, it is often assumed that y_j and z_j are independent. Then, the covariance matrix of X_j becomes

$$\mathbf{\Sigma} = A\mathrm{Cov}(y_j)A^T + \sigma^2 I_d.$$

Let $\alpha_1, \cdots, \alpha_K$ denote the eigenvalues of $A\mathrm{Cov}(y_j)A^T$. Since the rank of $A\mathrm{Cov}(y_j)A^T$ is at most K, the eigenvalues of Σ, i.e. spectrum, are

$$\mathrm{spec}(\mathbf{\Sigma}) = (\underbrace{\alpha_1 + \sigma^2, \alpha_2 + \sigma^2, \cdots, \alpha_K + \sigma^2}_{K}, \underbrace{\sigma^2, \cdots, \sigma^2}_{d-K}) \tag{1.1.2}$$

where $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_K \geq 0$. In this model, the signal component of $\mathbf{\Sigma}$ has K *spikes* $(\alpha_1, \alpha_2, \cdots, \alpha_K)$, so (1.1.2) is called the *spike covariance model* (Johnstone, 2001).

Under this spike model, it is of great interest to estimate the underlying spike signals, i.e. dimension reduction. One of the most popular dimension reduction techniques is Principal Component Analysis (PCA) (Jolliffe, 2002). PCA finds a low dimensional subspace maximizing the explained variation in data, which enables us to recover the underlying signals. To understand the performance of PCA, many of its interesting theoretical properties have been investigated. From now on, we assume $\alpha_K > 0$ in (1.1.2). In the classical domain, Anderson (1963) showed that the first K sample eigenvectors are consistent. However, when the dimension d is very large, the sample eigenvectors are no longer always consistent. In the RMT domain, a number of papers examine this aspect (Paul, 2007; Bai and Silverstein, 2010; Paul and Johnstone, 2012; Nadler et al., 2008; Johnstone and Lu, 2009; Benaych-Georges and Nadakuditi, 2011). Also, in the HDLSS domain, Ahn et al. (2007); Jung and Marron (2009); Jung et al. (2012); Shen et al. (2013); Aoshima et al. (2018) found mathematical conditions which characterize the consistency and the strong inconsistency of sample eigenvectors. Shen et al. (2016) developed a general asymptotic framework to explore interesting transitions among those three asymptotic domains.

In Chapter 1.3, we extend the previous asymptotic studies for high dimensional data to the case where there are a small number of outliers. First, we introduce a useful statistical framework modeling high dimensional outliers. We then explore the behavior of such high dimensional outliers and the consistency of the first few PC directions with a major interest in the recovery of underlying directions in which only a small number of outliers go. The challenges to our theories relative to the existing theory lie in that data observations may not follow an identical distribution, which makes some well-known theories established under the i.i.d. case no longer applicable. The results have an important application for detecting outliers in RNA-seq data as will be discussed in Chapter 2.7.2.

Under certain conditions, the studies mentioned above support the idea that the first few principal component (PC) directions provide crucial information on the signals in high dimensional data. Then, how many PC directions should be used? Interesting algorithms have been proposed

to address this question often under the assumption of $\varepsilon_j = \sigma z_j$. However, it has been observed that some modern data structures do not follow such assumption. Assuming more flexible noise structure, we propose in Chapter 3.6.2 an algorithm for determining an effective dimension size based on some fundamental results in RMT.

We also provide a visualization tool for assessing the distributional assumption made on the underlying eigenvalues. For example, the assumption, $\varepsilon_j = \sigma z_j$ with d-dimensional white noise z_j, indicates that the distribution of the underlying eigenvalues is a point mass at σ^2. As such, the proposed method is a visual alternative to testing underlying covariance structure.

1.2 RNA-seq data

Over the last decade, impressive advancements have been made in cancer genomics due to the progress of high-throughput or next generation sequencing (NGS) technologies. An important achievement is better understanding of the transcriptome, the set of transcripts in a cell, which is crucial for inferring the functions of genes and understanding human disease (Wang et al., 2009; Koboldt et al., 2013; Buermans and Den Dunnen, 2014; Van Dijk et al., 2014). In particular, massively parallel cDNA sequencing, or RNA-seq (RNA sequencing), enables the analysis of the entire transcriptome in a very high-throughput and quantitative manner.

RNA-seq uses deep-sequencing technologies. RNA-seq experiments mostly start with a population of RNA, convert it to cDNA fragments, and then obtain short reads (sequences) from each fragment from one end (single-end) or both ends (pair-end) (Wang et al., 2009; Oshlack et al., 2010). Millions of short (25-400bp) reads are generated from this process and then aligned to the reference genome or transcriptome. These aligned reads construct read count pile-ups for each gene locus, also known as a base-resolution expression profile, expression coverage, or per-base read depths. This single-base resolution for annotation has enabled an unprecedented landscape view of the transcriptional structure of genes by providing a microscopic examination of the transcriptome.

RNA-seq has several benefits over other existing NGS technologies. Specifically, it has high sensitivity by producing a huge range of expression levels with very low background noise (Wang

et al., 2009; Ozsolak and Milos, 2011). This enables more accurate quantification of RNA expression levels than microarrays. Also, its per-base resolution characterizes the precise exon-intron boundaries, allowing deep examination of splicing diversity. These advantages of RNA-seq have led to an enormous range of applications such as gene expression profiling, identification of novel transcripts, studies of DNA methylation and protein binding sites, the complete characterization of the genome of non-model organisms, and cancer gene identification (Pan et al., 2008; Keren et al., 2010; Eswaran et al., 2013).

Despite many advantages of RNA-seq, it also presents several technological and bioinformatic challenges. For example, spurious artifacts can be introduced from the experimental steps such as amplification, library construction, sequencing, and mapping (Conesa et al., 2016). This dissertation studies statistical challenges raised by RNA-seq data analysis. In general, RNA-seq expression coverage data for a single gene have measurements on the read-depths of thousands ($1,000 \sim 30,000$) of base-positions, but only with hundreds of individuals available. This is a good example of high dimensional data mentioned earlier, which motivates the theories and methodologies developed herein.

A substantial proportion of human genes differ in function in ways that are reflected through different forms of shape changes in expression coverage of RNA-seq. For example, very diverse splicing patterns and insertions/deletions have been observed. Standard genetic approaches use junction reads or single nucleotide changes. Taking a different approach to addressing the high-dimensionality issue, we propose a RNA-seq shape change detection method called Scissor (Shape changes in selecting sample outliers in RNA-seq). Particularly, Scissor depends on a new statistical model for outliers in high-dimensional settings that will be introduced in Chapter 1.3.

PCA has been an important exploratory step prior to the downstream analysis in the field of computational biology. For example, the first and second PCs from RNA-seq data often reveal the impacts of different library sizes and batch effects. Also, PCA has been extensively used for reducing dimensions in many genomic data types. An important challenge in the application of PCA to RNA-seq data is to determine the number of PCs that will be included in the analysis. This

is mainly because the noise structure of expression coverage differs from the regular white noise assumption. To address this challenge, we propose an algorithm in Chapter 3.6.2 as mentioned earlier. The proposed method has been successful in offering more reasonable numbers of PCs in RNA-seq data compared to other existing algorithms. Furthermore, this algorithm allows us to automatically choose different numbers of PCs for different genes.

1.3 Outline

This chapter presented the motivation and backgrounds of the theories and methodologies developed hereafter. The remainder of the dissertation is divided into three parts. The first presents some theoretical works on high-dimensional outliers motivated by RNA-seq data. The second part introduces Scissor. The last part presents an algorithm for determining the number of PC directions with application to RNA-seq data.

CHAPTER 2. THEORY OF HIGH-DIMENSIONAL OUTLIERS

2.1 Motivation

From a classical point of view, outliers have been considered as *bad* cases that may confound the statistical analysis. In this case, one may think the data are contaminated by a few outliers and those should be down-weighted or potentially removed from the dataset. Much work in this case has been done. See Hampel et al. (2011) and Huber (2011) for a good overview. On the other hand, there are situations where outliers can produce important and rich information. For example, aberrant observations of gene expression data can be highly related to important genetic phenomena such as mutations, abnormal splicing, and structural variations that are known to be strongly connected to cancer. In both cases, the study of outliers helps to better understand data.

Roughly speaking, in low dimensional space, a data point is an outlier if it does not fit the distribution that a majority of the data points come from. However, this definition is more challenging for high dimensional data due to the 'curse of dimensionality', i.e. the phenomenon where the data points tend to be more apart from each other as the dimension increases. As discussed in Section 2.4, when $d \gg n$, Hall et al. (2005) showed that data points tend to lie near the surface of a high-dimensional sphere and that, more surprisingly, all pairwise distances of points are approximately equal and all pairwise angles are approximately perpendicular. These geometrical properties indicate that data points in high dimensional space are very sparse, and thus they might be considered as *inliers*, which makes it challenging to distinguish outliers from them. Due to this curse of dimensionality, classical outlier detection methods such as distance-based or depth-based approaches (Barnett and Lewis, 1974; Hawkins, 1980; Stahel, 1981; Donoho and Gasko, 1992; Liu, 1992; Zuo and Serfling, 2000; Zuo, 2003; Dang and Serfling, 2010) do not work well for high dimensional data. Over the last decade, several alternative outlier detection methods have

been developed to tackle the challenge of high dimensionality. (Filzmoser et al., 2008; Ro et al., 2015; Rousseeuw et al., 2016; Ahn et al., 2018) However, there is no consensus on the definition of outliers and each method targets different types of outliers. For further discussion, see Section 3.2. In this chapter, we introduce a new notion of high dimensional outliers that embraces various types of outliers and provides deep insights into understanding the behaviors of outliers in high dimensions.

Often, the classical large sample theory does not provide good approximations to high dimensional data. For example, many statistics such as Hotelling's T^2-statistic, generalized variances, multiple correlation coefficients, and various statistics for sphericity tests are asymptotically consistent under a classical asymptotic regime, but those asymptotics are no longer valid with large d even $d < n$. To understand such different asymptotic behavior of high dimensional data, as mentioned earlier, tremendous efforts have been made over the last few decades under several different asymptotic regimes (Baik and Silverstein, 2006; Jung and Marron, 2009; Shen et al., 2016; Wang et al., 2013; Paul and Aue, 2014; Yao et al., 2015). However, the studies on limiting properties of high dimensional outliers are still lacking. Under the new notion of outliers, we investigate the conditions under which outliers can be distinguished from inliers as well as the conditions under which such outliers can be asymptotically well captured by a low dimensional subspace produced by PCA. Our theoretical results extend the previous asymptotic studies for high dimensional data to the case where there are a small number of outliers. The results have an important application for detecting outliers in RNA-seq data as will be discussed in Chapter 2.7.2.

The remainder of this chapter is organized as follows. In Section 2.2, we review related work. Section 2.3 introduces a model for an underlying distribution possibly generating outliers. Some geometrical properties of high dimensional outliers are explored in Section 2.4. Theoretical aspects related to the asymptotic behavior of sample eigenvalues and eigenvectors when there are two different types of signals, outlier signals as well as main signals, are investigated in Section 2.5. Section 2.6 provides a toy example to illustrate the theoretical results. The proofs of the theorems are given in Section 2.7.

2.2 Related work

Let X be a d-dimensional random vector with mean vector μ and covariance matrix $\mathbf{\Sigma}$. Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the d ordered eigenvalues of $\mathbf{\Sigma}$ and $U_1, \cdots, U_d$ be the corresponding eigenvectors. Let $X_1, \cdots, X_n$ be observations on X. Denote the sample covariance matrix by $\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})^T$ and its ordered sample eigenvalues and eigenvectors by $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ and $\hat{U}_1, \cdots, \hat{U}_d$, respectively. The asymptotic study of sample eigenvalues ($\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$) and sample eigenvectors ($\hat{U}_1, \cdots, \hat{U}_d$) has an interesting history and developed roughly in three different asymptotic domains: the classical domain, the RMT domain, and the HDLSS (See Chapter . In each domain, different asymptotic theories have been established.

In the classical domain, Girshick (1939) investigated the asymptotic properties of sample eigenvalues and eigenvectors in the case of all the eigenvalues of $\mathbf{\Sigma}$ being different. When the smallest $d-q$ eigenvalues of $\mathbf{\Sigma}$ are equal and the others are all different, Lawley (1953) investigated the asymptotic theories of sample eigenvectors. When $X_1, \cdots, X_n$ are from a multivariate normal distribution, Anderson (1963) has given the asymptotic distribution of $\hat{\lambda}_1, \cdots, \hat{\lambda}_d$, $\hat{U}_1, \cdots, \hat{U}_d$ in the case of $\lambda_1, \cdots, \lambda_d$ having any multiplicities. The asymptotic study of the eigenstructure of the sample covariance matrix in the classical domain essentially relies on the fact that the population covariance matrix is well approximated by the sample covariance matrix when the sample size is large with dimension fixed. When the dimension is also large, however, this is no longer the case.

In the RMT domain, these phenomena were explored in a large number of papers, e.g. Marčenko and Pastur (1967); Silverstein and Choi (1995); Silverstein (1995); Bai and Silverstein (1998); Baik et al. (2005); Paul (2007); Bai and Yao (2012). See also Bai (2008), Bai and Silverstein (2010), Paul and Aue (2014) and Yao et al. (2015) for useful overview. A well-known observation is that the empirical spectral distribution (ESD) of the sample covariance matrix converges almost surely to the Marcenko-Pastur distribution when the population covariance matrix is the identity and d and n proportionally grow to infinity. Combining the fact that the an eigenvalue is a continuous function of a matrix, this supports the idea that the sample covariance matrix is not a good estimate of the

population covariance matrix for large dimensions. However, many data sets in high dimensions involve quiet different eigenvalues, for instance, a few largest of those are much larger than the other eigenvalues. To understand these phenomena, the *spiked covariance model* was initially introduced by Johnstone (2001) and extensively studied. Baik et al. (2005) studied the conditions of the first m population eigenvalues that provided the corresponding sample eigenvalues being separate from the other small eigenvalues under the spike covariance model. They proved a transition phenomenon: the limits of the extreme sample eigenvalues depend on the critical value $1+\sqrt{c}$, i.e. a sample eigenvalue from a population eigenvalue that is greater than $1+\sqrt{c}$ is asymptotically isolated from the others, i.e. the *bulk* eigenvalues. Baik and Silverstein (2006) extended the results of Baik et al. (2005) to non-Gaussian variables and found that the limits of the extreme sample eigenvalues depend on the critical values $1+\sqrt{c}$ for the largest spike eigenvalues and on $1-\sqrt{c}$ for the smallest spike eigenvalues. Bai and Yao (2012) extended the results to a generalized spike covariance model that allows flexibility on the distribution of bulk population eigenvalues. The spike covariance model is closely related to the concept of small-rank perturbations, i.e. theories on perturbed random matrices. In a small-rank perturbation approach, convergence of the few largest sample eigenvalues and the corresponding sample eigenvectors are studied in Benaych-Georges and Nadakuditi (2011).

Note that underlying spike eigenvalues are constant in the classical domain and the RMT domain where the increasing sample size n boosts the consistency. On the other hand, in the HDLSS domain, underlying spike eigenvalues are allowed to increase, which encourages the PCA consistency for increasing dimension d and a fixed n (Ahn et al., 2007; Jung and Marron, 2009; Jung et al., 2012; Shen et al., 2016). Jung and Marron (2009) explored the asymptotic behaviors of the spike eigenvectors when the levels of spike eigenvalues increase at the rate d^{α}. In the case of $\alpha > 1$, they showed that the spike eigenvectors are subspace consistent, i.e. the subspace spanned by the sample spike eigenvectors consistently estimates the subspace spanned by the underlying population spike eigenvectors, and is strongly inconsistent for $\alpha < 1$, i.e. the angle between each sample eigenvector and the true one converges to 90 degrees. Jung et al. (2012) deeply explored the boundary case ($\alpha = 1$) and showed the convergence in distribution of the first spike eigenvector

under the normal assumption. Shen et al. (2016) have provided a general framework of the PCA consistency that nicely connected the existing results from different domains except for some boundary cases.

In this chapter, we deeply explore the behaviors of high dimensional outliers via geometrical representations in the HDLSS domain and asymptotic theories of sample eigenvalues and eigenvectors from the data containing a few outliers under the general framework studied in Shen et al. (2016). A major interest is the consistent estimation of underlying outlier directions in which only a small number of outliers go. We will provide for each scenario a condition that allows achievement of the PCA individual consistency or subspace consistency.

2.3 Model and Notations

In this section, we introduce a model that provides a new notion of high dimensional outliers. Figure 2.2 shows a motivating example with 30 normal RNA-seq data curves in grey color with two potential colored outliers. Each curve represents each observation in the genomic region around the gene TP53. The curves are read depth (or coverage), i.e. the number of reads aligned to each nucleoide, which are log10 read counts from RNA-seq experiments based on HNSCC (Head and Neck Squamous Cell Carcinoma) cancer tissue samples obtained from the TCGA Research Network. We use the terminology *sample* to indicate a patient. Exons, highlighted by colored background (except for pink), are regions of a gene that are annotated as the part of the messenger RNA region. By contrast, introns, highlighted by a white background together with on colored pink, are regions of the gene that are expected to be *spliced out*, i.e. not used in the RNA production. In the figure, the red and blue curves behave differently from the others in the sense that the red one retains an intron, as highlighted by the pink background, and the blue one skips several exons, as highlighted by the green background. Such abnormal splicing events are called *intron retention* and *exon skipping*, respectively. It has been observed that such events happen with a small chance at each gene, hence it makes sense to consider such samples as outliers.

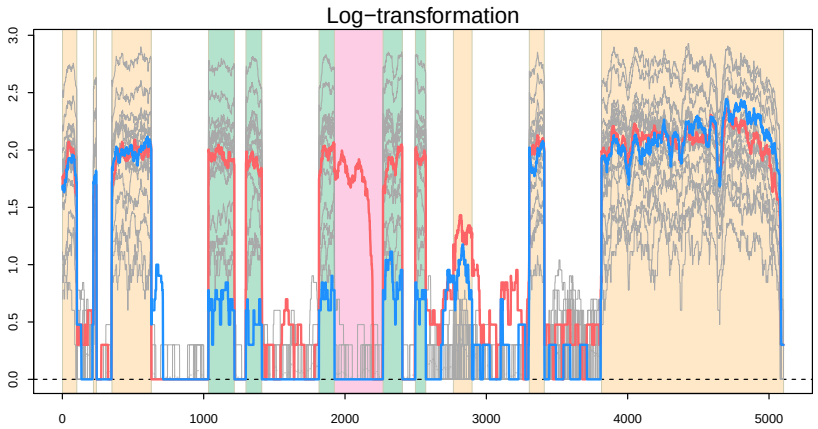


Figure 2.2: The 30 RNA-seq observations for the gene TP53 are plotted on the log-scale. Exons are highlighted by colored background and introns are indicated using mostly white background. The red and blue curves indicate biologically important outliers and the other gray curves indicate normal observations.

The two red and blue outliers show clearly different structure from the other curves, which implies that they show different underlying signals that do not fit together with the majority of the data. At the same time, interestingly, some of the main structures of the two outliers are shared with most of the data. This example motivated us to consider two different types of underlying directions in the data space, together with variation in those directions in describing outliers. Two important types are *outlier directions* that may lead to prominent high dimensional outliers and *main directions* whose variation is shared among all data points including outliers. The new proposed model incorporating these two components is now introduced in three parts.

Part 1. The classical way of describing underlying variations of a random vector using PCA is discussed in this paragraph. Let X be a random vector distributed as a d-dimensional multivariate normal distribution, $N_d(0, \mathbf{\Sigma}_d)$. The spectral decomposition of the population covariance matrix is

$$\mathbf{\Sigma}_d = \mathbf{U}_d \mathbf{\Lambda}_d \mathbf{U}_d^T$$

where $\mathbf{U}_d = [u_1, \cdots, u_d]$ contains the orthonormal eigenvectors of $\mathbf{\Sigma}_d$ in its columns and $\mathbf{\Lambda}_d = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$ is a diagonal matrix with the corresponding non-negative eigenvalues. Then, a random vector from $N_d(0, \mathbf{\Sigma}_d)$ can be expressed as

$$X = \mathbf{U}_d \mathbf{\Lambda}_d^{1/2} Z = \mathbf{U}_d Y$$

where $Z \sim N_d(0, I_d)$ and $Y \sim N_d(0, \mathbf{\Lambda}_d)$. That is, X is a linear combination of U_i with random coefficients y_i from $N(0, \lambda_i)$, i.e.,

$$X = \sum_{i=1}^{d} y_i U_i, \quad y_i \sim N(0, \lambda_i). \tag{2.3.1}$$

In the terminology of PCA, the y_i are the *principal components*, i.e. the scores or projection coefficients (Jolliffe, 2002). Intuitively, if X involves large y_i for some i, then the direction U_i is an important direction of variation of the underlying distribution of X, whereas if $y_i \approx 0$, X does not feel strongly the direction U_i.

Part 2. Distributions for modeling outliers are now considered. Based on the intuition behind the principal components, an outlier can be viewed as an observation that goes strongly in some directions that the bulk of data points do not. Denote one of those directions by U_i^* and the corresponding random coefficient by y_i^*. Then outliers that go in the direction U_i^* have large y_i^*'s whereas the other data points have small y_i^*'s in (2.3.1). We model this underlying variation of a random coefficient y_i^* by a scale mixture distribution with two different variances, $\tau_{i,2} \gg \tau_{i,1} > 0$, i.e.,

$$y_i^* \sim \begin{cases} \sqrt{\tau_{i,1}} z_i, & \text{w.p. } 1 - w_i \\ \sqrt{\tau_{i,2}} z_i, & \text{w.p. } w_i, \end{cases} \tag{2.3.2}$$

where the z_i's are i.i.d random variables with mean zero and variance one and $0 \leq w_i \leq 1$, with $w_i \approx 0$. The first part of the mixture distribution with the smaller variance, $\tau_{i,1}$, describes the behavior of the majority of data vectors with little variation in the direction U_i^*. The second part of the mixture

distribution with the larger variance, $\tau_{i,2}$, corresponds to outliers, and so we assume that w_i is small, e.g. less than 0.05. This mixture model well reflects an underlying mechanism generating outliers in the sense that "one person's noise could be another person's signal", as pointed out in Kamber and Han (2001).

Part 3. A new model for an underlying distribution embracing a small set of outliers is introduced based on the classical setting (2.3.1) together with the distribution (2.3.2) beyond the Gaussian models. Let $\mathbf{X} = [X_1, \cdots, X_n]$ be a data matrix whose columns are independent observation vectors distributed as a d-dimensional (perhaps non-Gaussian) multivariate distribution with a small number of aberrant vectors whose signals are different from the majority of the data. Let $\{U_i\}_{1 \le i \le d}$ be a set of underlying orthogonal vectors some of which are responsible for the potential outliers. Note that these vectors do not need to be the eigenvectors of the underlying covariance matrix. In the spirit of (2.3.1), an observation vector X_j can be expressed as a linear combination of the orthonormal direction vectors, $\{U_i\}_{1 \le i \le d}$, whose coefficients are independent random variables distributed as different mixture distributions, i.e.

$$X_j = \sum_{i=1}^{d} y_{ij} U_i, \quad \text{where} \quad y_{ij} \sim \begin{cases} \sqrt{\tau_{i,1}} z_{ij}, & \text{w.p. } 1 - w_i \\ \sqrt{\tau_{i,2}} z_{ij}, & \text{w.p. } w_i, \end{cases} \tag{2.3.3}$$

where the z_{ij}'s are assumed to be i.i.d. random variables with mean zero, variance one, and bounded fourth moment. Then, the random variables $\{y_{ij}\}_{1 \le j \le n}$ with $w_i > 0$ model how the direction U_i as an outlier component can generate outliers. Also, we will use $w_i = 0$ for other directions especially main components, which allows flexibility to include the classical way of describing the variation from underlying directions as in (2.3.1). To distinguish the two components, we let I_{main} denote a set of dimension indices that correspond to main components and I_{out} denote outlier components. That is, $I_{out} = \{1 \le i \le d \mid w_i > 0\}$ and $I_{main} = \{1 \le i \le d \mid w_i = 0\} = \{1, \cdots, d\} \backslash I_{out}$. Also, we denote the sample indices that are outlying in each outlier component, indexed by $i \in I_{out}$, by $s_i = \{1 \le j \le n \mid y_{ij} = \sqrt{\tau_{i,2}} z_{ij}\}$.

Under the model (2.3.3), outliers are allowed to share important features or background noise with normal data points. The model also allows an outlier to be associated with several outlier components, which offers flexibility in modeling the nature of outliers. Under this setting, a sample vector from (2.3.3) can be viewed as a random vector from a complicated mixture distribution whose components have different covariance structures.

As discussed earlier, still there is no consensus definition for outliers. Every procedure may target its own informal definition for outliers based on various goals. Here, we describe several types of outliers that are commonly used in various applications as special cases of the proposed model in (2.3.3).

- Variable-specific outliers: This type of outlier is different from the bulk of the data only at single variables. If an observation is an outlier with respect to the original variables, then it is usually extreme on these variables. Assuming there are d variables in the model, each sample can be modeled by (2.3.3) with $U_i = e_i$ for $i = 1, \cdots, d$. Here, e_i is a unit vector with 1 for the ith entry and 0 for the others. Then, an outlier X_j in the m-th variable can be described by $\tau_{m,2} > \tau_{m,1}$ and an underlying outlier proportion w_m.

- Scale mixture outliers: The outliers in this category exhibit a much different abberation, across all variables simultaneously, and are more scattered than the majority of data, and thus they are also known as *scatter outliers* (Filzmoser et al., 2008). Let $X_j \sim N_d(0, \sigma_1^2 \mathbf{\Sigma})$ with probability $1-p$ and $N_d(0, \sigma_2^2 \mathbf{\Sigma})$ with a small probability p and $\sigma_2^2 \gg \sigma_1^2$. This scale mixture model is a special subset of the model (2.3.3) with $w_i = p, \tau_{i,1} = \sigma_1^2, \tau_{i,2} = \sigma_2^2$ for all $i = 1, \cdots, d$, where the U_i's are a set of the orthogonal vectors, e.g. the eigenvectors of $\mathbf{\Sigma}$. Additionally, the I_{out} will include every index, $I_{main} = \emptyset$, and $s_1 = s_2 = \cdots = s_d$. That is,

$$X_j = \begin{cases} \sum_{i=1}^d y_{ij} U_i, \text{ where } y_{ij} = \sigma_1 z_{ij}, & \text{w.p. } 1-p \\ \sum_{i=1}^d y_{ij} U_i, \text{ where } y_{ij} = \sigma_2 z_{ij}, & \text{w.p. } p. \end{cases}$$

- Shifted outliers: The shifted outliers are those that are shifted globally to a common direction (Filzmoser et al., 2008; Ro et al., 2015; Dai and Genton, 2016). Often, these outliers share most of the variation with the the bulk of the data, but present abnormally high or low overall pattern, which is typically described by the mean vector denoted by μ. Let X_j be independent random vectors from $a_j\mu + Z_j$, where $Z_j \sim N_d(0,\Sigma)$, $a_j \sim N(0,\sigma_1^2)$ with probability $1-p$ and $N(0,\sigma_2^2)$ with probability p, and Z_j and a_j are independent. Assuming $\sigma_1 < \sigma_2$ and a small p, the random variable a_j describes how a small fraction of data points may be shifted. Define one of the underlying vectors, say U_1, to be the normalized mean vector, that is, $U_1 = \mu/\|\mu\|$, and the other underlying vectors to be orthogonal to each other. Then, the variation from the U_1 for normal samples and outliers are respectively $\sigma_1^2\|\mu\|^2 + U_1^T\mathbf{\Sigma}U_1$ and $\sigma_2^2\|\mu\|^2 + U_1^T\mathbf{\Sigma}U_1$. Thus, each data object can be modeled by

$$X_j = \sum_{i=1}^{d} y_{ij}U_i, \quad \text{where} \quad y_{1j} \sim \begin{cases} N(0,\sigma_1^2\|\mu\|^2 + U_1^T\mathbf{\Sigma}U_1), & \text{w.p. } 1-p \\ N(0,\sigma_2^2\|\mu\|^2 + U_1^T\mathbf{\Sigma}U_1), & \text{w.p. } p, \end{cases}$$

with the other y_{ij} from $N(0, U_i^T\mathbf{\Sigma}U_i)$ for $i = 2,\cdots,d$.

2.4 Geometrical representation

It is important to understand the behavior of outliers in high dimensional space. Roughly speaking, the distance between data points becomes heavily dominated by noise as dimension increases, resulting in a sparse data set where outliers are less distinguishable. Zhou and Marron (2016) studied the case where some outliers are too close to each other due to some common factors, e.g. family members, and thus unduly affect the conventional PCA and some robust methods. However, studies on the behavior of high dimensional outliers in a systematic manner are still lacking. This section explores the geometrical features of the high dimensional outliers based on the model (2.3.3).

It is of great interest to understand when outliers in high dimensions may deviate from the majority and when they may not. Intuitively, if $\tau_{i,2}$ in some outlier components are dramatically larger than $\tau_{i,1}$, then the relevant outliers are more likely to be separated from the bulk of the data. By contrast, if $\tau_{i,2}$ do not differ much from $\tau_{i,1}$, the corresponding outliers are expected to be harder to distinguish. As discussed below, an interesting observation in high dimensional data is that if an outlier is involved in a large fraction of outlier directions, d encourages the separability of the outlier from the other normal data points even when $\tau_{i,2}$ is not substantially large. On the other hand, if an outlier is involved in a limited number of outlier directions, d discourages the separability even for relatively large $\tau_{i,2}$'s. We study these phenomena using the geometrical representation of high dimensional outliers in the HDLSS context explored by Hall et al. (2005) and identify a condition when outliers may be distinguishable in such high dimensions.

We consider a simple scenario where data come from (2.3.3) with $\tau_{i,1} = \sigma^2$ and $\tau_{i,2} = \tau^{(d)}$ for all i under the normality assumption. In this section, we index the variation for outlier components by d, $\tau^{(d)}$, as an indication of increase with dimension. Then, our model can be expressed as

$$y_{ij} = \begin{cases} \sigma z_{ij}, & \text{w.p. } 1 - w_i \\ \tau^{(d)} z_{ij}, & \text{w.p. } w_i, \end{cases} \quad \text{for } i \in I_{out} \text{ and } y_{ij} = \sigma z_{ij} \text{ for } i \notin I_{out}. \tag{2.4.1}$$

Consider a non-outlier point X_j from (2.4.1) which can be expressed as $X_j = \sum_{i=1}^{d} \sigma z_{ij} U_i$ where the U_i are orthonormal underlying eigenvectors. As d increases, it follows by a law of large numbers that its squared Euclidean distance scaled by d converges to the constant σ^2 in the sense that

$$\begin{aligned} \frac{1}{d}\|X_j\|^2 &= \frac{1}{d}\sum_{i=1}^{d} \sigma^2 z_{ij}^2 \\ &\to \sigma^2 \end{aligned} \tag{2.4.2}$$

almost surely. Then, we might fairly say that a non-outlier point X_j lies approximately on the surface of a d-variate sphere, of radius $(\sigma^2 d)^{1/2}$, as $d \to \infty$. Similarly, we can obtain limiting behavior of distances between pairs of non-outlier points. The distance between two non-outlier points X_j and

X_l is approximately equal to $(2\sigma^2 d)^{1/2}$ as $d \to \infty$, in the sense that

$$\begin{aligned} \frac{1}{d}\|X_j - X_l\|^2 &= \frac{1}{d}\sum_{i=1}^{d} \sigma^2 (z_{ij} - z_{il})^2 \\ &\to 2\sigma^2 \end{aligned} \tag{2.4.3}$$

where the convergence is almost sure. These asymptotic results match with the results in Hall et al. (2005). As described in their paper, application of (2.4.3) to each pair (j,l) of non-outliers, and scaling all distances by the factor $d^{-1/2}$, shows that they asymptotically construct a polyhedron where each edge is of length $(2\sigma^2)^{1/2}$ and the vertices are the m non-outliers.

Similarly, we now explore the behavior of outliers in high dimensions. An outlier point $X_{j'}$ can be expressed as

$$X_{j'} = \sum_{i \in I_{out}^{j'}} \sqrt{\tau^{(d)}} z_{ij'} U_i + \sum_{i \notin I_{out}^{j'}} \sigma z_{ij'} U_i$$

where $I_{out}^{j'}$ is an index set for outlier components related to $X_{j'}$. Let $K_{j'}^{(d)} = |I_{out}^{j'}|$ be the cardinality of the set $I_{out}^{j'}$ for each d and $p_{out}^{j'} = \lim_{d\to\infty} \frac{K_{j'}^{(d)}}{d}$ be the fraction of the outliers components for a large d. The deviation of $X_{j'}$ from the majority depends on the levels of $K_{j'}^{(d)}$, that is, $p_{out}^{j'} > 0$, $p_{out}^{j'} = 0$ with $K_{j'}^{(d)} \to \infty$, and $p_{out}^{j'} = 0$ with $K_{j'}^{(d)}$ fixed. Each case requires the different levels of $\tau^{(d)}$ as will be discussed below.

Let us first consider the case of $p_{out}^{j'} > 0$ with $\tau = \lim_{d\to\infty} \tau^{(d)}$. It follows that if a law of large numbers applies to its squared distance divided by d, then

$$\begin{aligned} \frac{1}{d}\|X_{j'}\|^2 &= \frac{1}{d}\sum_{i \in I_{out}^{j'}} \tau^{(d)} z_{ij'}^2 + \frac{1}{d}\sum_{i \notin I_{out}^{j'}} \sigma^2 z_{ij'}^2 \\ &= \frac{K_{j'}^{(d)}}{d} \frac{1}{K_{j'}^{(d)}} \sum_{i \in I_{out}^{j'}} \tau^{(d)} z_{ij'}^2 + \frac{d - K_{j'}^{(d)}}{d} \frac{1}{d - K_{j'}^{(d)}} \sum_{i \notin I_{out}^{j'}} \sigma^2 z_{ij'}^2 \\ &\to p_{out}^{j'} \tau + (1 - p_{out}^{j'}) \sigma^2 \end{aligned} \tag{2.4.4}$$

almost surely as $d \to \infty$. This implies that an outlier point $X_{j'}$ is approximately of distance $(\sigma^2 d + p_{out}^{j'}(\tau - \sigma^2)d)^{1/2}$ from the origin. Also, the distance between an outlier $X_{j'}$ and a non-outlier X_j divided by $d^{1/2}$ converges almost surely to $(p_{out}^{j'}(\tau - \sigma^2) + 2\sigma^2)^{1/2}$ as $d \to \infty$:

$$\begin{aligned} \frac{1}{d}\|X_j - X_{j'}\|^2 &= \frac{1}{d}\sum_{i \in I_{out}^{j'}} (\sigma z_{ij} - \sqrt{\tau^{(d)}} z_{ij'})^2 + \frac{1}{d}\sum_{i \notin I_{out}^{j'}} \sigma^2 (z_{ij} - z_{ij'})^2 \\ &\to p_{out}^{j'}(\tau - \sigma^2) + 2\sigma^2. \end{aligned} \tag{2.4.5}$$

Therefore, a larger $p_{out}^{j'}$ or a larger τ help to better separate the outlier $X_{j'}$ from non-outliers provided that $\tau > \sigma^2$ and $p_{out} > 0$. In particular, this geometrical property shows that even when τ is not much bigger than σ^2, good separability still follows when $p_{out}^{j'}$ is sufficiently large for high dimensions whereas it tends to be less successful in low dimensions (Filzmoser et al., 2008).

The type of scale mixture outliers introduced in Section 2.3 is a special example of this case with $\sigma_1^2 = \sigma^2$ and $\sigma_2^2 = \tau$. For this particular type, all the resulting outliers have $p_{out} = 1$, which together with (2.4.2) and (2.4.4) leads to two d-variate spheres of different radii: a sphere of radius $(\sigma^2 d)^{1/2}$ on the surface of which the non-outliers approximately lie, and another sphere of radius $(\tau d)^{1/2}$ for the outliers. This geometrical representation is also associated with the unique spectrum limit of the sample covariance matrix of high dimensional scale mixture distributions as studied in Li and Yao (2018). They showed that the limit of the ESD from the scale mixture distribution can be viewed as a mix of the two separate ESD limits relevant to each mixture component, and the separation of these two limits becomes more distinct for a larger ratio of $\frac{d}{n}$. Roughly speaking, the part of the spectrum limit containing large eigenvalues is associated with the larger sphere of radius $(\tau d)^{1/2}$ and the other part involving smaller eigenvalues is associated with a smaller sphere of radius $(\sigma^2 d)^{1/2}$.

So far, we have observed that d encourages the geometrical separability of an outlier $X_{j'}$ if $p_{out}^{j'} > 0$. However, this is no longer the case for $p_{out}^{j'} = 0$ because the terms $p_{out}^{j'}\tau$ and $p_{out}^{j'}(\tau - \sigma^2)$ in (2.4.4) and (2.4.5), respectively, disappear for large d, which discourages the separability. In this more challenging situation, we need a $\tau^{(d)}$ much bigger than σ^2 to approximately models the

separability. So here we let $\tau^{(d)}$ increase as d increases. As mentioned earlier, the case with $p_{out}^{j'} = 0$ is further divided into two cases where $K_{j'}^{(d)}$ increases as d increases and where $K_{j'}^{(d)}$ is fixed. Let us first explore the case with increasing $K_{j'}^{(d)}$. We model the idea of a stronger outlier as

$$\frac{K_{j'}^{(d)}\tau^{(d)}}{d} \to r_{j'} \quad \text{as } d \to \infty. \tag{2.4.6}$$

Then, it is easy to show $\frac{1}{d}\|X_{j'}\|^2 \to r_{j'} + \sigma^2$ and $\frac{1}{d}\|X_j - X_{j'}\|^2 \to r_{j'} + 2\sigma^2$ as $d \to \infty$. This indicates that $r_{j'}$ plays an important role in separating $X_{j'}$ from non-outliers geometrically. If $r_{j'}$ is too small, and in particular if it equals 0, then the data points in the sample including outliers asymptotically behave as a regular data set with the absence of outliers. On the other hand, if the $r_{j'}$ is large enough, the outlier $X_{j'}$ tends to be distinguished from the sphere on the surface of which the majority of data points spread out.

The results above hold for increasing $K_{j'}^{(d)}$ as $d \to \infty$, for fixed sample size n. For the case of a limited number of outlier directions, i.e. $K_{j'}^{(d)} = K_{j'}$, a law of large numbers may not be applicable, and rather we employ the convergence in distribution. Then, we have $\frac{1}{d}\|X_{j'}\|^2 \to_d \frac{1}{K_{j'}}\sum_{i \in I_{out}^{j'}} r_{j'} z_{ij'}^2 + \sigma^2$ and $\frac{1}{d}\|X_j - X_{j'}\|^2 \to_d \frac{1}{K_{j'}}\sum_{i \in I_{out}^{j'}} r_{j'} z_{ij'}^2 + 2\sigma^2$. Still, we see that the level of $r_{j'}$ determines the separability of an outlier from the other normal data points. But here it is good to mention that $r_{j'}$ becomes the limit of $\frac{\tau^{(d)}}{d}$, which only depends on the level of $\tau^{(d)}$, because we fix the $K_{j'}$.

To sum up, our study in this section enables understanding of the transition phenomenon of high dimensional outliers from near the surface of a high dimensional sphere to being distant from the sphere. Our results indicate that there are two factors affecting this transition which are the proportion of outlier components involved in an outlier and the signals of those outlier directions.

2.5 PCA consistency

In a spike covariance model, a fixed number of population eigenvalues are assumed to be much larger than the others. This provides an important sense in which the signals corresponding to large

population eigenvalues are consistently estimated by PCA under some conditions that depend on various asymptotic domains (Shen et al., 2016). We employ the same concept of a spike covariance model here. Let K be the total number of different spike components among the covariance matrices in the mixture components. For convenience, we refer to $\{U_i\}_{1\le i\le K}$ as *spike directions* and $\{U_i\}_{K+1\le i\le d}$ as *non-spike directions*. The non-spike components are often considered as noise. In a modification of the definition in Section 2.3, denote the index sets for outlier spike components and main spike components by $I_{out} = \{1 \le i \le K | w_i > 0\}$ and $I_{main} = \{1, \cdots, K\} \backslash I_{out}$, respectively. That is, $\{U_i\}_{i\in I_{out}}$ is the set of outlier spike directions and $\{U_i\}_{i\in I_{main}}$ is the set of main spike directions. The inherent variation derived in each direction U_i can be expressed as $\lambda_i = (1-w_i)\tau_{i,1} + w_i\tau_{i,2}$ by the mixture distribution in (2.3.3) and such λ_i's are indeed the population eigenvalues corresponding to the direction U_i. This is because the covariance matrix of X_j from (2.3.3) can be written as $\mathbf{\Sigma} = \mathrm{Cov}(X_j) = \mathrm{Cov}(\mathbf{U}y_j) = \mathbf{U}\mathrm{Cov}(y_j)\mathbf{U}^T$ where $y_j = (y_{1j}, \cdots, y_{dj})^T$. Due to the independence of $\{y_{ij}\}_{1\le i\le d}$, $\mathrm{Cov}(y_j)$ is a diagonal matrix whose entries are $var(y_{ij}) = (1-w_i)\tau_{i,1} + w_i\tau_{i,2}$, and thus the λ_i's are the eigenvalues of $\mathbf{\Sigma}$ by the eigenvalue decomposition.

Let $X_1, \cdots, X_n$ be observations from (2.3.3) with the K spike components as described above. Denote the sample covariance matrix by $\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}\mathbf{X}^T = \frac{1}{n}\sum_{j=1}^n X_j X_j^T$ and its eigenvalue decomposition by $\hat{\mathbf{\Sigma}} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}^T$ with $\hat{\mathbf{U}} = [\hat{U}_1, \cdots, \hat{U}_d]$ and $\hat{\mathbf{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_d)$ where $\{(\hat{\lambda}_k, \hat{U}_k) : k = 1, \cdots, d\}$ are the pairs of eigenvalues and eigenvectors of $\hat{\mathbf{\Sigma}}$ such that $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_d$. In this section, asymptotic properties of $\hat{\lambda}_1, \cdots, \hat{\lambda}_d$ and $\hat{U}_1, \cdots, \hat{U}_d$ are analyzed under the general framework developed by Shen et al. (2016). As discussed in Section 2.2, this general framework includes several previously studied domains as special cases and allows one to understand interesting connections among the various domains. This section provides parallel asymptotic results for data from a complicated mixture distribution (2.3.3) and enables understanding of the behavior of outlier components in high dimensions. The main contribution of our theorems compared to the existing theories lie in that data observations do not follow the same distribution, which mean the well-known theories must be extended.

We consider increasing sample size n, increasing dimension d, and increasing spike signals. As an indication of increasing spike signals, we let λ_i, $\tau_{i,1}$, and $\tau_{i,2}$ be sequences indexed by n, that is, $\lambda_i^{(n)}$, $\tau_{i,1}^{(n)}$, and $\tau_{i,2}^{(n)}$. Consider the $M+1$ tiers where the first K eigenvalues, $\{\lambda_i^{(n)}\}_{1\le i\le K}$, are grouped such that q_m eigenvalues fall into the m-th tier where $\sum_{m=1}^{M} q_m = K$ and the rest of the eigenvalues are all grouped into the $M+1$-th tier. Define $q_0 = 0, q_{M+1} = d-K$, and the partial sums $p_m = \sum_{l=0}^{m} q_l$. Then, the index set of the eigenvalues in the m-th tier can be written as

$$H_m = \left\{ p_{m-1}+1, p_{m-1}+2, \cdots, p_{m-1}+q_m \right\} \quad \text{for } \ m = 1, \cdots, M+1.$$

Denote a linear subspace spanned by the components in the m-th tier by $S_m = \text{span}\{U_i, i \in H_m\}$ for $m = 1, \cdots, M+1$.

The following assumptions provide the conditions for the variances, $\tau_{i,1}^{(n)}$ and $\tau_{i,2}^{(n)}$, of the underlying mixtures. Several different conditions are assumed for main spike signals, outlier spike signals, and noise signals, which helps to distinguish spike components from non-spike components. There are two types of noise in our model. One type is noise for all data points that correspond to the non-spike components in the model. By contrast, the other type is noise for the majority but a signal for a few observations. The latter type of noise is modeled by the small variance part in the outlier components. The following two assumptions illustrate the variances for these two types of noise.

Assumption 2.5.1. $\lim_{n\to\infty} \tau_{i,1}^{(n)} = \lim_{n\to\infty} \tau_{i,2}^{(n)} = c_\lambda \ \ \textit{for } i \in H_{M+1}$.

Assumption 2.5.2. $\lim_{n\to\infty} \tau_{i,1}^{(n)} = c_\lambda \ \ \textit{for } i \in I_{out}$.

In a spike covariance model, noise signals are described in non-spike components and the corresponding underlying eigenvalues often are assumed to be constant for modeling white noise. This helps the bulk eigenvalues corresponding to the noise possess some known asymptotic properties. For instance, the distribution of the bulk eigenvalues converges to some well-known distributions, e.g. the Marcenko-Pastur law or the semi-circular law, and the extreme eigenvalues (the smallest and largest eigenvalues) are also known to be consistent to some values or asymptotically follow

the Tracy-Widom distribution (Marčenko and Pastur, 1967; Bai and Yin, 1988; Bai et al., 1988; Bai and Yin, 1993; Johnstone, 2001). In the same spirit, Assumption 2.5.1 describes the asymptotically equivalent noise signals for non-spike directions $\{U_i\}_{i\in H_{M+1}}$. Eventually, the underlying eigenvalues $\{\lambda_i^{(n)}\}_{i>K}$ are all equal to c_λ for large d. Assumption 2.5.2 describes noise variances ($\tau_{i,1}^{(n)}$) for the outlier spike components. Since the outlier spike components are nothing but noise for the majority of the data, the same level of variation assumed for the non-spike components can be assumed. Thus, the noise variances for outlier spike directions are also asymptotically equal to c_λ. This nicely connects the outlier model with the null model, i.e. the case with no outlier spike components, in the sense that the outlier components will merge with non-spike noise components.

In contrast to noise signals, we allow spike signals to be increasing in n. The intensity of each spike component is determined by the underlying variation that each component is involved in, which is equivalent to its corresponding eigenvalue. For large n, the underlying eigenvalues, $\lambda_i^{(n)}$, are simply $\tau_{i,1}^{(n)}$ for $i \in I_{main}$ whereas, for $i \in I_{out}$, the eigenvalues are $w_i\tau_{i,2}^{(n)}$ because variation from the larger variance component $\tau_{i,2}^{(n)}$ dominate variation from the smaller variance component $\tau_{i,1}^{(n)}$. The PCA consistency strongly depends on the magnitudes of spike eigenvalues, which are specified in a systematic manner in the following assumptions. Let $\delta_m^{(n)}$ for $m = 1, \cdots, M$ be sequences of constant values for index n.

Assumption 2.5.3. $\lim_{n\to\infty} \frac{\tau_{i,1}^{(n)}}{\delta_m^{(n)}} = 1$ *for* $i \in H_m \cap I_{main}$ *and* $\lim_{n\to\infty} \frac{\tau_{i,2}^{(n)}}{w_i\delta_m^{(n)}} = 1$ *for* $i \in H_m \cap I_{out}$, $m = 1, \cdots, M$.

Assumption 2.5.4. *As* $n \to \infty$, $\delta_1^{(n)} \succ \delta_2^{(n)} \succ \cdots \succ \delta_M^{(n)} \succ \lambda_{K+1}^{(n)}$ *where* $a_n \succ b_n$ *implies* $\lim_{n\to\infty} \frac{a_n}{b_n} > 1$.

Assumption 2.5.3 allows the components in the same tier to share asymptotically equivalent eigenvalues. We further assume different limiting coefficients for different tiers in Assumption 2.5.4, which enables the characterization of the M subspaces spanned by the directions in each tier.

Under Assumptions 2.5.1-2.5.4, we now investigate the asymptotic properties of the sample eigenvalues and eigenvectors. Even though we assume a complicated mixture distribution for the underlying structure and thus the data are not i.i.d., we obtain parallel asymptotic results to those in Shen et al. (2016). This is because although observations are not from an identical distribution,

they are allowed to share the same underlying eigenvectors from the model (2.3.3). Then, we have a simple integrated covariance matrix $\mathbf{\Sigma}$ so that a spike covariance model can be employed even when data come from multiple distributions.

In general, the strength of underlying spike signals and increasing sample size n encourage PCA consistency whereas increasing dimension d discourages consistency. When the underlying spike signals in the m-th tier with increasing n are asymptotically strong enough to prevail over the dimension d in the sense that $\frac{d}{n\delta_m^{(n)}} \to 0$, it follows that the estimates of the eigenvectors are subspace consistent in the m-th tier and the estimates of the eigenvalues are consistent as well. Theorems 2.5.1 and 2.5.2 demonstrate such asymptotic behavior in a concrete manner under different scenarios.

Theorem 2.5.1. *Under Assumptions 2.5.1-2.5.4,*

(a) if $\frac{d}{n\delta_M^{(n)}} \to 0$, *then*

(i) for $i \leq K$, $\frac{\hat{\lambda}_i}{\lambda_i^{(n)}} \to_{a.s.} 1$ *where* $\lambda_i^{(n)} = \tau_{i,1}^{(n)}$ *for* $i \in I_{main}$ *and* $\lambda_i^{(n)} = w_i \tau_{i,2}^{(n)}$ *for* $i \in I_{out}$;

(ii) for $i > K$,

- *if* $0 < c < \infty$, $c_\lambda(1-\sqrt{c})^2 \leq \hat{\lambda}_{n\wedge d} \leq \hat{\lambda}_1 \leq c_\lambda(1+\sqrt{c})^2$ *a.s.;*
- *if* $c = \infty$, $\frac{n\hat{\lambda}_i}{d} \to_{a.s.} c_\lambda$;
- *if* $c = 0$, $\hat{\lambda}_i \to_{a.s.} c_\lambda$;

(b) if $\frac{d}{n\delta_h^{(n)}} \to 0$ *where* $1 \leq h < M$ *and* $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$, *then*

(i) for $i \leq p_h$, $\frac{\hat{\lambda}_i}{\lambda_i^{(n)}} \to_{a.s.} 1$ *where* $\lambda_i^{(n)} = \tau_{i,1}^{(n)}$ *for* $i \in I_{main}$ *and* $\lambda_i^{(n)} = w_i \tau_{i,2}^{(n)}$ *for* $i \in I_{out}$;

(ii) for $i > p_h$, $\frac{n\hat{\lambda}_i}{d} \to_{a.s.} c_\lambda$.

Theorem 2.5.1 considers two scenarios: (a) when all spike signals are strong and (b) when strong population signals are assumed only up to the h-th tier and the other signals are dominated by the increasing dimension, i.e. $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$. It should be noted that these two different scenarios yield different asymptotic regimes: (a) considers all three cases for the limit of c, i.e. $0 < c < \infty$, $c = \infty$, and $c = 0$, whereas Theorem 2.5.1 (b) considers only the case of $c = \infty$. This is because the

condition $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$ of (b) and Assumption 2.5.4 together rule out the cases of $c < \infty$ as $\frac{d}{n\lambda_{K+1}^{(n)}} \to \infty$ can hold only when $\frac{d}{n} \to \infty$. In both cases, if the signal in a tier is strong enough so that $\frac{d}{n\delta^{(n)}} \to 0$, then the sample eigenvalues corresponding to the tier consistently estimate the true eigenvalues. On the other hand, if the spike signals are not that strong, then the corresponding sample eigenvalues tend to be swallowed by the small bulk eigenvalues.

Intuitively, although an underlying outlier component is dramatically intense, its realized signal is much weaker than the true one because it loses the power due to the small chance of participation. Assumption 2.5.3 reflects this intuition and gives a condition that the ith outlier spike signal should be $1/w_i$ times greater than the other main spike signals in the same tier to compensate for this loss of power. Based on this assumption, Theorem 2.5.1 demonstrates that such an outlier signal would asymptotically attain the same sample eigenvalues as the main signals in the same tier. In particular, the sample eigenvalue from an outlier signal converges to the dominating variance ($\tau_{i,2}^{(n)}$) multiplied by the corresponding proportion (w_i) in the underlying mixture distribution (2.3.3). Therefore, the true levels of outlier signals can be approximately estimated by dividing the corresponding eigenvalues by the proportion ($\approx w_i$) of the relevant outliers.

In many outlier detection methods, it is of great interest to choose the subspace that outlier components are involved in (Filzmoser et al., 2008; Ahn et al., 2018). Although Theorem 2.5.1 suggests that a few large sample eigenvalues may consistently estimate the true levels of the signals, it is not enough to say that the corresponding principal component directions construct a useful subspace for detecting outliers. This brings to the study of eigenvectors that is discussed in the following theorem. Let $\delta_0^{(n)} = \infty$ for all n.

Theorem 2.5.2. *Under Assumptions 2.5.1-2.5.4,*

(a) if $\frac{d}{n\delta_M^{(n)}} \to 0$, *and* $0 < c \leq \infty$, *then*

(i) $\hat{U}_i$ *are subspace consistent in the sense that the* $\text{angle}(\hat{U}_i, S_m) \to_{a.s.} 0$ *for* $i \in H_m$, $m = 1, \cdots, M+1$.

- *For* $m = 1, \cdots, M-1$, $\text{angle}(\hat{U}_i, S_m) = o(\{\frac{\delta_m^{(n)}}{\delta_{m-1}^{(n)}} \vee \frac{\delta_{m+1}^{(n)}}{\delta_m^{(n)}}\}^{1/2})$.

- *For* $m = M$, $\text{angle}(\hat{U}_i, S_m) = o(\{\frac{\delta_m^{(n)}}{\delta_{m-1}^{(n)}}\}^{1/2}) \vee O(\{\frac{d}{n\delta_m^{(n)}}\}^{1/2})$.
- *For* $m = M+1$, $\text{angle}(\hat{U}_i, S_m) = O(\{\frac{d}{n\delta_{m-1}^{(n)}}\}^{1/2})$.

(b) if $\frac{d}{n\delta_h^{(n)}} \to 0$ *where* $1 \le h < M$ *and* $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$, *then*

(i) $\hat{U}_i$ *for* $i \le p_h$ *are subspace consistent in the sense that the* $\text{angle}(\hat{U}_i, S_m) \to_{a.s.} 0$ *for* $i \in H_m$, $m = 1, \cdots, h$.

- *For* $m = 1, \cdots, h-1$, $\text{angle}(\hat{U}_i, S_m) = o(\{\frac{\delta_m^{(n)}}{\delta_{m-1}^{(n)}} \vee \frac{\delta_{m+1}^{(n)}}{\delta_m^{(n)}}\}^{1/2})$.
- *For* $m = h$, $\text{angle}(\hat{U}_i, S_m) = o(\{\frac{\delta_m^{(n)}}{\delta_{m-1}^{(n)}}\}^{1/2}) \vee O(\{\frac{d}{n\delta_m^{(n)}}\}^{1/2})$.

(ii) $\hat{U}_i$ *for* $i > p_h$ *are strongly inconsistent in the sense that* $| < \hat{U}_i, U_i > | = O(\{\frac{n\lambda_i^{(n)}}{d}\}^{1/2})$.

Under the same scenarios considered in Theorem 2.5.1, Theorem 2.5.2 studies the asymptotic behavior of the sample eigenvectors in terms of angles as studied in Jung and Marron (2009) and Shen et al. (2016). In each scenario, if the mth tier involves a strong signal such that $\frac{d}{n\delta_m^{(n)}} \to 0$, then the ith sample eigenvector, $\hat{U}_i$ for $i \in H_m$, tends to be in the subspace, S_m, which is spanned by the underlying directions in the mth tier. This holds for all $i \in H_m$, and thus it follows that the subspace spanned by the sample eigenvectors, $\{\hat{U}_i\}_{i \in H_m}$, converges to the S_m. This phenomenon is called the PCA subspace consistency. Also, the different levels of signals in different tiers assumed in Assumption 2.5.4 make the estimated subspaces become distinct for large n and large d, and more gaps between the levels accelerate this distinction as indicated by the different convergence rates obtained in the theorem.

In high-dimensional data, searching for outliers in a much lower dimensional subspace where the outliers are distinguishable is advantageous. The subspace often provides critical information for interpreting why an object is outlying and to what extent the object is an outlier (Kamber and Han, 2001). This is almost impossible using full dimensions because of the overwhelming noise. Once such a subspace is found, some appropriate conventional outlier detection methods may be applicable for the approximated low-dimensional data. Under the assumption that all outlier signals dominate the dimensions, combining Theorem 2.5.2 (a) with Theorem 2.5.1 (a) allows one to find

the outlier-relevant low dimensional subspace by using the first few PC directions whose sample eigenvalues are substantially large and thus separate from the other bulk eigenvalues.

On the other hand, Theorem 2.5.2 (b) together with Theorem 2.5.1 (b) shows that when only a subset of the outlier signals are strong enough to dominate the increasing dimensions, the first few PC directions with large sample eigenvalues provide a good subspace only for those strong outlier components. Not only this, the strong inconsistency suggests that it becomes very challenging to distinguish the outlier directions missing from the first few PCs from the non-spike directions. This is not simply because the sample eigenvalues from those weak outlier signals are not separable from the bulk sample eigenvalues. Once the spike samples eigenvalues are swallowed by the bulk, then it is likely that the corresponding directions are all mixed with non-spike directions, so any of the single sample eigenvectors may not be representative of those spike directions. Therefore, the approximation of the data matrix using the first few eigenvectors may miss some important information that are relatively weak but not noise. This implies that the outlier components with weak signals or with extremely small participation are harder to be separated from noise. Thus special care should be taken to find the hidden outlying structure.

Also, it should be noted that the theorem does not guarantee that the sample eigenvectors are individually consistent to the true ones. So looking at the individual PC directions may not be enough to detect outliers. To illustrate this situation, a toy example is given in Section 2.6. As one of the special and important cases, we now consider the case when all spike eigenvalues are separable, i.e. $q_1 = q_2 = \cdots = q_M = 1$ and $M = K$. Then, Assumption 2.5.4 becomes

Assumption 2.5.5. *As* $n \to \infty$, $\lambda_1^{(n)} \succ \lambda_2^{(n)} \succ \cdots \succ \lambda_K^{(n)} \succ \lambda_{K+1}^{(n)} > 0$.

This allows us to get the individual consistency of eigenvalues as well as eigenvectors instead of subspace consistency. The following corollaries of Theorem 2.5.1 and 2.5.2 describe such individual consistency under the same scenarios with the respective theorems.

Corollary 2.5.1. *Under Assumptions 2.5.1, 2.5.2, and 2.5.5,*

(a) if $\frac{d}{n\lambda_K^{(n)}} \to 0$, *then*

(i) for $i \le K$, $\frac{\hat{\lambda}_i}{\lambda_i^{(n)}} \to_{a.s.} 1$ *where* $\lambda_i^{(n)} = \tau_{i,1}^{(n)}$ *for* $i \in I_{main}$ *and* $\lambda_i^{(n)} = w_i \tau_{i,2}^{(n)}$ *for* $i \in I_{out}$;

(ii) for $i > K$,

- *if* $0 < c < \infty$, $c_\lambda(1-\sqrt{c})^2 \le \hat{\lambda}_{n\wedge d} \le \hat{\lambda}_1 \le c_\lambda(1+\sqrt{c})^2$ *a.s.*;
- *if* $c = \infty$, $\frac{n\hat{\lambda}_i}{d} \to_{a.s.} c_\lambda$;
- *if* $c = 0$, $\hat{\lambda}_i \to_{a.s.} c_\lambda$;

(b) if $\frac{d}{n\lambda_h^{(n)}} \to 0$ *where* $1 \le h < K$ *and* $\frac{d}{n\lambda_{h+1}^{(n)}} \to \infty$, *then*

(i) for $i \le h$, $\frac{\hat{\lambda}_i}{\lambda_i^{(n)}} \to_{a.s.} 1$ *where* $\lambda_i^{(n)} = \tau_{i,1}^{(n)}$ *for* $i \in I_{main}$ *and* $\lambda_i^{(n)} = w_i \tau_{i,2}^{(n)}$ *for* $i \in I_{out}$;

(ii) for $i > h$, $\frac{n\hat{\lambda}_i}{d} \to_{a.s.} c_\lambda$.

Corollary 2.5.2. *Under Assumptions 2.5.1, 2.5.2, and 2.5.5,*

(a) if $\frac{d}{n\lambda_K^{(n)}} \to 0$, *and* $0 < c \le \infty$, *then*

(i) $\hat{U}_i$ *are consistent with* U_i *in the sense that* $\text{angle}(\hat{U}_i, U_i) \to_{a.s.} 0$ *for* $i = 1, \cdots, K$.

- *For* $i = 1, \cdots, K-1$, $\text{angle}(\hat{U}_i, U_i) = o(\{\frac{\lambda_i^{(n)}}{\lambda_{i-1}^{(n)}} \vee \frac{\lambda_{i+1}^{(n)}}{\lambda_i^{(n)}}\}^{1/2})$.
- *For* $i = K$, $\text{angle}(\hat{U}_i, U_i) = o(\{\frac{\lambda_i^{(n)}}{\lambda_{i-1}^{(n)}}\}^{1/2}) \vee O(\{\frac{d}{n\lambda_i^{(n)}}\}^{1/2})$.
- *For* $i > K$, $\text{angle}(\hat{U}_i, S) = O(\{\frac{d}{n\lambda_{i-1}^{(n)}}\}^{1/2})$ *where* $S = \text{span}(U_i : i > K)$.

(b) if $\frac{d}{n\lambda_h^{(n)}} \to 0$ *where* $1 \le h < K$ *and* $\frac{d}{n\lambda_{h+1}^{(n)}} \to \infty$, *then*

(i) $\hat{U}_i$ *are consistent with* U_i *in the sense that* $\text{angle}(\hat{U}_i, U_i) \to_{a.s.} 0$ *for* $i = 1, \cdots, h$.

- *For* $i = 1, \cdots, h-1$, $\text{angle}(\hat{U}_i, U_i) = o(\{\frac{\lambda_i^{(n)}}{\lambda_{i-1}^{(n)}} \vee \frac{\lambda_{i+1}^{(n)}}{\lambda_m^{(n)}}\}^{1/2})$.
- *For* $i = h$, $\text{angle}(\hat{U}_i, U_i) = o(\{\frac{\lambda_i^{(n)}}{\lambda_{i-1}^{(n)}}\}^{1/2}) \vee O(\{\frac{d}{n\lambda_i^{(n)}}\}^{1/2})$.

(ii) $\hat{U}_i$ *for* $i > h$ *are strongly inconsistent in the sense that* $|<\hat{U}_i, U_i>| = O(\{\frac{n\lambda_i^{(n)}}{d}\}^{1/2})$.

2.6 Illustration using a toy example

We now illustrate the PCA subspace consistency with a toy example under the model (2.3.3), highlighting the situation where an outlier component is captured by the first few PC directions but none of the PC directions are individually representative of the outlier component.

First, let us describe the simulation setting. We generated $n = 200$ independent data vectors in $d = 3000$ dimensions based on our model described in (2.3.3). To generate such data, $\{z_{ij}\}_{1\le i\le d, 1\le j\le n}$ are assumed to be distributed as independent $N(0,1)$ and the standard basis vectors, $\{e_i\}_{1\le i\le d}$, are used as underlying eigenvectors $\{U_i\}_{1\le i\le d}$ with $e_1,\cdots,e_9$ being the main spike directions and e_{10} being an outlier spike direction. For the main spike directions, the underlying variations are assumed to be $\tau_{i,1} = 3000, 1000, 100, 90, 80, 70, 60, 50, 40$ for $i = 1,\cdots,9$. For the outlier spike direction e_{10}, we assume $\tau_{10,1} = 2000, \tau_{10,2} = 1$ and the outlier proportion $w_{10} = 0.02$. For the other non-spike directions $\{U_i\}_{11\le i\le d}$ corresponding to noise, $\tau_{i,1} = 1$ and $\tau_{i,2} = 1$ are assumed. A realization from this model had the 4 outliers, denoted by $X_1,\cdots,X_4$, with 196 normal data points, denoted by $X_5,\cdots,X_{200}$.

For this data set, we constructed a sample covariance matrix where PCA was applied and obtained a set of sample eigenvectors and eigenvalues. Since the true spike directions are $e_1,\cdots,e_{10}$, we can examine the contribution of each sample eigenvector onto the true spike directions simply by taking the squares of the entries. The sum of the squares of entries in each sample eigenvector is one and thus the squared values $\hat{u}_{ji}^2$, i.e. the squared jth entry of $\hat{U}_i$, can be regarded as the explained percentage of the underlying vector e_j in the direction $\hat{U}_i$. Table 2.1 gives the squares of the first 12 entries (in rows) of the first 11 eigenvectors $\hat{U}_1,\cdots,\hat{U}_{11}$ (in columns). The last two rows indicate the corresponding sample eigenvalues $\hat{\lambda}_i$ and the angles between the true outlier direction e_{10} and $\hat{U}_i$ for $i = 1,\cdots,11$. The largest value in each $\hat{U}_i$ is indicated using red, and if the red value, say $\hat{u}_{ji}^2$, is close to one and the other entries are close to zero, then the $\hat{U}_i$ is a good estimate of the e_j. For example, the first entry of $\hat{U}_1$ is approximately one with all the other entries of zero, indicating that the first underlying direction e_1 is well estimated by $\hat{U}_1$. Similarly, $\hat{U}_2$ is a good estimate of the e_2.

On the other hand, none of the $\hat{U}_3, \cdots, \hat{U}_{10}$ has an entry which is close to one. Instead, they have several nonzero entries, indicating that each of them has some correlation with several underlying directions. This can be understood that any of the underlying directions, $e_3, \cdots, e_{10}$ are well estimated by the single sample eigenvectors. Nonetheless, an important note is that, for each row $j = 3, \cdots, 10$, the sum of the squared jth entries of $\hat{U}_3, \cdots, \hat{U}_{10}$ is close to one. This supports the PCA subspace consistency in that each of the true eigenvectors $e_3, \cdots, e_{10}$ can be estimated by a linear combination of $\hat{U}_3, \cdots, \hat{U}_{10}$ rather than any individual directions. As described in Theorem 2.5.2, this is because the underlying variation in $e_3, \cdots, e_{10}$ are nearly in the same tier, which tends to somewhat discourage the individual consistency.

In particular, it should be noted that none of the first 10 eigenvectors alone provide good estimates for the outlier direction e_{10} as highlighted in lightblue. Specifically, there is no direction that describes the e_{10} more than 30%. The angles in the last row also reveal that none of those 10 sample eigenvectors are close to e_{10}. However, the sum of the squared 10th entries in the first 10 PC directions, $\sum_{i=1}^{10} \hat{u}_{i,10}^2$, is almost 0.88, indicating that the e_{10} may be well captured by the subspace spanned by the 10 sample eigenvectors. As discussed earlier, therefore, this supports the concept that although using individual PC directions for detecting outliers may be ineffective, this subspace does preserve the critical information for the outliers and thus may be used to detect those outliers.

	$\hat{U}_1$	$\hat{U}_2$	$\hat{U}_3$	$\hat{U}_4$	$\hat{U}_5$	$\hat{U}_6$	$\hat{U}_7$	$\hat{U}_8$	$\hat{U}_9$	$\hat{U}_{10}$	$\hat{U}_{11}$
1	0.996	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.987	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
3	0.000	0.000	0.657	0.001	0.072	0.173	0.004	0.000	0.000	0.002	0.000
4	0.000	0.000	0.020	0.665	0.111	0.023	0.055	0.011	0.001	0.003	0.000
5	0.000	0.001	0.029	0.031	0.330	0.422	0.048	0.008	0.001	0.013	0.000
6	0.000	0.000	0.071	0.006	0.198	0.023	0.064	0.501	0.007	0.003	0.000
7	0.000	0.000	0.036	0.002	0.080	0.009	0.408	0.201	0.053	0.059	0.000
8	0.000	0.000	0.000	0.007	0.004	0.021	0.094	0.047	0.243	0.382	0.000
9	0.000	0.001	0.003	0.000	0.012	0.001	0.010	0.000	0.467	0.292	0.000
10	0.000	0.000	0.101	0.182	0.079	0.215	0.169	0.096	0.030	0.007	0.000
11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\hat{\lambda}_i$	3519.209	996.408	123.856	99.055	91.825	86.694	71.458	70.393	52.236	42.087	16.850
angle	89.6	89.1	71.5	64.8	73.7	62.4	65.7	71.9	80.0	85.3	89.4

Table 2.1: This table shows the first 11 sample eigenvectors with squared entries. The largest value in each $\hat{U}_i$ is colored using a red font and the row corresponding to the outlier signal, e_{10}, is hightlighted using lightblue background. This row indicates how much each $\hat{U}_i$ explains the outlier signal. The last two rows respectively show the sample eigenvalues corresponding to $\{\hat{U}_i\}_{1\leq i\leq 11}$ and the angles between $\hat{U}_i$ and the true outlier signal, e_{10}.

2.7 Proofs

In this section, we provide proofs for the theorems in Section 2.5. The main steps in the proofs are similar to the proofs in Shen et al. (2016) but our different setting for the underlying distribution of data requires the addition of more detail.

Let $\mathbf{Y}$ be a $d \times n$ matrix whose column vectors are $Y_1, Y_2, \cdots, Y_n$ where $Y_j = (y_{1j}, \cdots, y_{dj})^T$ and y_{ij}'s are independent random variables in our model in (2.3.3). Then, switching the roles of columns and rows, we get the $n \times n$ dual (Gram) matrix of the sample covariance matrix $\hat{\mathbf{\Sigma}}$

$$\hat{\mathbf{\Sigma}}_D = \frac{1}{n}\mathbf{X}^T\mathbf{X} = \frac{1}{n}\mathbf{Y}^T\mathbf{U}^T\mathbf{U}\mathbf{Y} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y},$$

and it is well known that they share the same nonzero eigenvalues. Let us define two matrices that will be treated separately in the proof. Let

$$\mathbf{A} = \frac{1}{n}\sum_{i=1}^{K} \tilde{Y}_i\tilde{Y}_i^T, \quad \text{and} \quad \mathbf{B} = \frac{1}{n}\sum_{i=K+1}^{d} \tilde{Y}_i\tilde{Y}_i^T,$$

where $\tilde{Y}_i$ is the i-th row vector of $\mathbf{Y}$. Then,

$$\hat{\mathbf{\Sigma}}_D \;=\; \frac{1}{n}\sum_{i=1}^{d} \tilde{Y}_i\tilde{Y}_i^T = \mathbf{A} + \mathbf{B}.$$

Before the proof, we provide two popular lemmas. Lemma 2.7.1 provides the upper and lower bounds for the eigenvalues of a matrix that can be expressed as the sum of two symmetric matrices.

Lemma 2.7.1. *(Weyl inequality) Let A and B be $n \times n$ real symmetric matrices. Then, for all $j,k,l = 1, \cdots, n$,*

$$\begin{aligned} \lambda_k(A) + \lambda_l(B) &\leq \lambda_j(A+B) \quad \text{for } k+l = j+n \\ \lambda_k(A) + \lambda_l(B) &\geq \lambda_j(A+B) \quad \text{for } k+l = j+1 \end{aligned}$$

where $\lambda_j(A)$ is the j-th largest eigenvalue of a matrix A.

Next, Lemma 2.7.2 provides the convergence of the largest and smallest non-zero eigenvalues of a random matrix, which is known as Bai-Yin's law (Bai and Yin, 1993).

Lemma 2.7.2. *(Bai-Yin's law) Suppose $B = \frac{1}{q}VV^T$ where V is a $p \times q$ random matrix composed of i.i.d. random variables with zero mean, unit variance and finite fourth moment. As $q \to \infty$ and $\frac{p}{q} \to c \in [0,\infty)$, the largest and smallest non-zero eigenvalues of B converge almost surely to $(1+\sqrt{c})^2$ and $(1-\sqrt{c})^2$, respectively.*

2.7.1 Proof of Theorem 2.5.1

Proof. The proof consists of the following three steps:

1. Establish the convergence of $\lambda_k(\mathbf{A})$.
2. Establish the convergence of $\lambda_k(\mathbf{B})$.
3. Establish the convergence of $\lambda_k(\mathbf{A}+\mathbf{B})$.

Lemma 2.7.3 proves the first step.

Lemma 2.7.3. *As $n \to \infty$, we have*

$$\frac{1}{\lambda_k^{(n)}}\lambda_k(\mathbf{A}) \to 1 \text{ a.s. for } k = 1, \cdots, K.$$

where $\lambda_i^{(n)} = \tau_{i,1}^{(n)}$ for $i \in I_{main}$ and $\lambda_i^{(n)} = w_i\tau_{i,2}^{(n)}$ for $i \in I_{out}$.

Proof. Define $\mathbf{A}_k = \frac{1}{n}\sum_{i=k}^{K}\tilde{Y}_i\tilde{Y}_i^T$ with $\mathbf{A}_{k,D}$ being its dual matrix and $\mathbf{A}_{k,R} = \mathbf{A} - \mathbf{A}_k$. Then,

$$\lambda_1(\frac{1}{n}\tilde{Y}_k\tilde{Y}_k^T) + \lambda_n(\frac{1}{n}\sum_{i=k+1}^{K}\tilde{Y}_i\tilde{Y}_i^T) \leq \lambda_k(\mathbf{A}) \leq \lambda_1(\mathbf{A}_k) + \lambda_k(\mathbf{A}_{k,R}) \tag{2.7.1}$$

where the upper bound follows from Lemma 2.7.1 and the lower bound from the expression (5.9) in Jung and Marron (2009). Since the rank of $\frac{1}{n}\sum_{i=k+1}^{K}\tilde{Y}_i\tilde{Y}_i^T$ is less than $K < n$, $\lambda_n(\frac{1}{n}\sum_{i=k+1}^{K}\tilde{Y}_i\tilde{Y}_i^T)$

should be zero. Likewise, the rank of $\mathbf{A}_{k,R}$ is at most $k-1$, and thus $\lambda_k(\mathbf{A}_{k,R}) = 0$. Therefore, we get

$$\lambda_1(\frac{1}{n}\tilde{Y}_k\tilde{Y}_k^T) \leq \lambda_k(\mathbf{A}) \leq \lambda_1(\mathbf{A}_k). \tag{2.7.2}$$

By dividing (2.7.2) by $\lambda_k^{(n)}$, the inequality becomes

$$\frac{1}{\lambda_k^{(n)}}\lambda_1(\frac{1}{n}\tilde{Y}_k\tilde{Y}_k^T) \leq \frac{1}{\lambda_k^{(n)}}\lambda_k(\mathbf{A}) \leq \frac{1}{\lambda_k^{(n)}}\lambda_1(\mathbf{A}_k). \tag{2.7.3}$$

We now show that the left hand side converges to 1. Note that $\lambda_1(\frac{1}{n}\tilde{Y}_k\tilde{Y}_k^T) = \lambda_1(\frac{1}{n}\tilde{Y}_k^T\tilde{Y}_k)$ and thus we show the convergence of $\frac{1}{\lambda_k^{(n)}}\frac{1}{n}\tilde{Y}_k^T\tilde{Y}_k$. For $k = 1, \cdots, d$, let $s_k = \{1 \leq j \leq n : y_{kj} = \sqrt{\tau_{k,2}^{(n)}}z_{kj}\}$, which is an index set containing sample indices from the second component in the mixture model (2.3.3) corresponding to the direction U_k. Then,

$$\begin{aligned}
\frac{1}{n}\tilde{Y}_k^T\tilde{Y}_k &= \frac{1}{n}\sum_{j=1}^{n} y_{kj}^2 \\
&= \frac{1}{n}\sum_{j\in s_k^c}\tau_{k,1}^{(n)}z_{kj}^2 + \frac{1}{n}\sum_{j\in s_k}\tau_{k,2}^{(n)}z_{kj}^2 \\
&= \tau_{k,1}^{(n)}\frac{|s_k^c|}{n}\frac{1}{|s_k^c|}\sum_{j\in s_k^c} z_{kj}^2 + \tau_{k,2}^{(n)}\frac{|s_k|}{n}\frac{1}{|s_k|}\sum_{j\in s_k} z_{kj}^2
\end{aligned} \tag{2.7.4}$$

where $|s|$ is the cardinality of a set s. If $k \in I_{main}$, then $s_k = \emptyset$ and $\lambda_k^{(n)} = \tau_{k,1}^{(n)}$, and thus it follows from the law of large number that $\frac{1}{\lambda_k^{(n)}}\frac{1}{n}\tilde{Y}_k^T\tilde{Y}_k \to 1$ almost surely. If $k \in I_{out}$, then we have $\lambda_k^{(n)} = w_k\tau_{k,2}^{(n)}$, $\frac{|s_k^c|}{n} \to 1 - w_k$, $\frac{|s_k|}{n} \to w_k$. Also, since $\frac{\tau_{k,1}^{(n)}}{\tau_{k,2}^{(n)}} \to 0$ as $n \to \infty$, the convergence $\frac{1}{\lambda_k^{(n)}}\frac{1}{n}\tilde{Y}_k^T\tilde{Y}_k \to 1$ for $k \in I_{out}$ also follows from the law of large numbers. Hence we conclude that

$$\frac{1}{\lambda_k^{(n)}}\lambda_1(\frac{1}{n}\tilde{Y}_k\tilde{Y}_k^T) \to 1 \text{ a.s.} \tag{2.7.5}$$

for $k = 1, \cdots, K$.

Next, we show that the right hand side of (2.7.3) also converges to 1. Let $\mathbf{A}_{k,D}$ be the dual matrix of $\mathbf{A}_k$. Then, it can be written as

$$\mathbf{A}_{k,D} = \frac{1}{n} \begin{pmatrix} \tilde{Y}_k^T \tilde{Y}_k & \tilde{Y}_k^T \tilde{Y}_{k+1} & \cdots & \tilde{Y}_k^T \tilde{Y}_K \\ \vdots & \ddots & & \vdots \\ \tilde{Y}_K^T \tilde{Y}_k & \cdots & & \tilde{Y}_K^T \tilde{Y}_K \end{pmatrix}.$$

Then, one can show that

$$\frac{1}{\lambda_k^{(n)}} \mathbf{A}_{k,D} \longrightarrow \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & b_{k+1,k} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & b_{K,k} \end{pmatrix}$$

where $b_{l,k} = \lim_{n\to\infty} \frac{\lambda_l^{(n)}}{\lambda_k^{(n)}}$. The above convergence of each element of $\mathbf{A}_{k,D}$ can be obtained in a similar way as in (2.7.4) and thus is omitted. It follows from $b_{l,k} \leq 1$ for $l \geq k$ that the largest eigenvalue of $\frac{1}{\lambda_k^{(n)}} \mathbf{A}_{k,D}$ converges almost surely 1. Note that $\frac{1}{\lambda_k^{(n)}} \lambda_1(\mathbf{A}_k) = \frac{1}{\lambda_k^{(n)}} \lambda_1(\mathbf{A}_{k,D}) = \lambda_1(\frac{1}{\lambda_k^{(n)}} \mathbf{A}_{k,D})$ and thus it follows that

$$\frac{1}{\lambda_k^{(n)}} \lambda_1(\mathbf{A}_k) \to 1 \text{ a.s.} \tag{2.7.6}$$

By (2.7.5) and (2.7.6), we conclude that

$$\frac{1}{\lambda_k^{(n)}} \lambda_k(\mathbf{A}) \to 1 \text{ a.s.}$$

□

Lemma 2.7.4. *As $n \to \infty$, we have*

$$\lambda_{\max}(\mathbf{B}) \to c_\lambda(1+\sqrt{c})^2 \text{ and } \lambda_{\min}(\mathbf{B}) \to c_\lambda(1-\sqrt{c})^2 \text{ a.s. if } \frac{d}{n} \to c < \infty$$
$$\frac{n}{d}\lambda_{\max}(\mathbf{B}) \text{ and } \frac{n}{d}\lambda_{\min}(\mathbf{B}) \to c_\lambda \text{ a.s. if } \frac{d}{n} \to \infty. \tag{2.7.7}$$

Proof. Let $\tilde{Y}_{i,(1)} = \left(y_{ij} I_{\{j \in s_i^c\}}\right)_{1 \le j \le n}$ be an n-dimensional vector whose elements are y_{ij} if $j \in s_i^c$ and 0 if $j \in s_i$ and $\tilde{Y}_{i,(2)} = \tilde{Y}_i - \tilde{Y}_{i,(1)}$. Similarly, define $\tilde{Z}_{i,(1)}$ and $\tilde{Z}_{i,(2)}$ with $\tilde{Z}_i = (z_{i1}, \cdots, z_{in})^T$. Then,

$$\begin{aligned}
\mathbf{B} &= \frac{1}{n}\sum_{i=K+1}^{d} \tilde{Y}_i \tilde{Y}_i^T \\
&= \frac{1}{n}\sum_{i=K+1}^{d} \left(\tilde{Y}_{i,(1)}\tilde{Y}_{i,(1)}^T + \tilde{Y}_{i,(2)}\tilde{Y}_{i,(2)}^T\right) \\
&= \frac{1}{n}\sum_{i=K+1}^{d} \left(\tau_{i,1}^{(n)} \tilde{Z}_{i,(1)}\tilde{Z}_{i,(1)}^T + \tau_{i,2}^{(n)} \tilde{Z}_{i,(2)}\tilde{Z}_{i,(2)}^T\right)
\end{aligned} \tag{2.7.8}$$

Without loss of generality, assume that $\tau_{i,1}^{(n)} \le \tau_{i,2}^{(n)}$ for all n and all $i = K+1, \cdots, d$. Then,

$$\mathbf{B} + \frac{1}{n}\sum_{i=K+1}^{d} (\tau_{i,2}^{(n)} - \tau_{i,1}^{(n)}) \tilde{Z}_{i,(1)}\tilde{Z}_{i,(1)}^T = \frac{1}{n}\sum_{i=K+1}^{d} \tau_{i,2}^{(n)} \tilde{Z}_i \tilde{Z}_i^T.$$

It follows from the Weyl inequality that

$$\lambda_k(\mathbf{B}) + \lambda_n\left(\frac{1}{n}\sum_{i=K+1}^{d} (\tau_{i,2}^{(n)} - \tau_{i,1}^{(n)}) \tilde{Z}_{i,(1)}\tilde{Z}_{i,(1)}^T\right) \le \lambda_k\left(\frac{1}{n}\sum_{i=K+1}^{d} \tau_{i,2}^{(n)} \tilde{Z}_i \tilde{Z}_i^T\right).$$

Note that $\frac{1}{n}\sum_{i=K+1}^{d} (\tau_{i,2}^{(n)} - \tau_{i,1}^{(n)}) \tilde{Z}_{i,(1)}\tilde{Z}_{i,(1)}^T$ is nonnegative definite and thus we have

$$\lambda_k(\mathbf{B}) \le \lambda_k\left(\frac{1}{n}\sum_{i=K+1}^{d} \tau_{i,2}^{(n)} \tilde{Z}_i \tilde{Z}_i^T\right). \tag{2.7.9}$$

Also, since

$$\frac{1}{n}\sum_{i=K+1}^{d}\tau_{i,2}^{(n)}\tilde{Z}_i Z_i^T+\frac{1}{n}\sum_{i=K+1}^{d}(\max_{K+1\le i\le d}\tau_{i,2}^{(n)}-\tau_{i,2}^{(n)})\tilde{Z}_i\tilde{Z}_i^T=\frac{1}{n}\sum_{i=K+1}^{d}\max_{K+1\le i\le d}\tau_{i,2}^{(n)}\tilde{Z}_i\tilde{Z}_i^T.$$

Again, the Weyl inequality and the nonnegativity of $\frac{1}{n}\sum_{i=K+1}^{d}(\max_{K+1\le i\le d}\tau_{i,2}^{(n)}-\tau_{i,2}^{(n)})\tilde{Z}_i\tilde{Z}_i^T$ yield the following inequality

$$\lambda_k(\frac{1}{n}\sum_{i=K+1}^{d}\tau_{i,2}^{(n)}\tilde{Z}_i\tilde{Z}_i^T)\le\lambda_k(\frac{1}{n}\sum_{i=K+1}^{d}\max_{K+1\le i\le d}\tau_{i,2}^{(n)}\tilde{Z}_i\tilde{Z}_i^T).$$

Hence, it follows from (2.7.9) that

$$\lambda_k(\mathbf{B})\le\lambda_k(\frac{1}{n}\sum_{i=K+1}^{d}\max_{K+1\le i\le d}\tau_{i,2}^{(n)}\tilde{Z}_i\tilde{Z}_i^T). \tag{2.7.10}$$

To get a lower bound of $\lambda_k(B)$, we start with the following equality

$$\mathbf{B}=\frac{1}{n}\sum_{i=K+1}^{d}\tau_{i,1}^{(n)}\tilde{Z}_i\tilde{Z}_i^T+\frac{1}{n}\sum_{i=K+1}^{d}(\tau_{i,2}^{(n)}-\tau_{i,1}^{(n)})\tilde{Z}_{i,(2)}\tilde{Z}_{i,(2)}^T.$$

from (2.7.8). Then, we obtain

$$\lambda_k(\frac{1}{n}\sum_{i=K+1}^{d}\min_{K+1\le i\le d}\tau_{i,1}^{(n)}\tilde{Z}_i\tilde{Z}_i^T)\le\lambda_k(\mathbf{B}) \tag{2.7.11}$$

in a similar way to get (2.7.10).

By (2.7.10) and (2.7.11) and letting $\mathbf{B}^*=\frac{1}{n}\sum_{i=K+1}^{d}\tilde{Z}_i\tilde{Z}_i^T$, it follows that

$$\min_{K+1\le i\le d}\tau_{i,1}^{(n)}\times\lambda_k(\mathbf{B}^*)\le\lambda_k(\mathbf{B})\le\max_{K+1\le i\le d}\tau_{i,2}^{(n)}\times\lambda_k(\mathbf{B}^*). \tag{2.7.12}$$

By Lemma 2.7.2, the convergence of the extreme eigenvalues of $\mathbf{B}^*$ can be obtained as follows:

$$\begin{aligned} \lambda_{\max}(\mathbf{B}^*) \to (1+\sqrt{c})^2 \text{ and } \lambda_{\min}(\mathbf{B}^*) \to (1-\sqrt{c})^2 \quad \text{if} \quad \frac{d}{n} \to c < \infty \\ \frac{n}{d}\lambda_{\max}(\mathbf{B}^*) \text{ and } \frac{n}{d}\lambda_{\min}(\mathbf{B}^*) \to 1 \text{ a.s.} \quad \text{if} \quad \frac{d}{n} \to \infty. \end{aligned} \tag{2.7.13}$$

For details, see the proof of Lemma 6.4 in Shen et al. (2016).

Since $\min_{K+1\le i\le d} \tau_{i,1}^{(n)}$ and $\max_{K+1\le i\le d} \tau_{i,2}^{(n)} \to c_\lambda$ by Assumption 2.5.1, it follows from (2.7.12) and (2.7.13) that

$$\begin{aligned} \lambda_{\max}(\mathbf{B}) \to c_\lambda(1+\sqrt{c})^2 \text{ and } \lambda_{\min}(\mathbf{B}) \to c_\lambda(1-\sqrt{c})^2 \text{ a.s.} \quad \text{if} \quad \frac{d}{n} \to c < \infty \\ \frac{n}{d}\lambda_{\max}(\mathbf{B}) \text{ and } \frac{n}{d}\lambda_{\min}(\mathbf{B}) \to c_\lambda \text{ a.s.} \quad \text{if} \quad \frac{d}{n} \to \infty. \end{aligned}$$

□

Theorem 2.5.1 (a). So far, we proved the first and second steps and now we prove the last step that completes the proof. It follows from the Weyl inequality and (2.7.1) that

$$\lambda_i(\mathbf{A}) + \lambda_n(\mathbf{B}) \le \lambda_i(\hat{\mathbf{\Sigma}}_D) \le \lambda_i(\mathbf{A}) + \lambda_1(\mathbf{B}) \text{ for } i = 1, \cdots, \min(d,n). \tag{2.7.14}$$

Let us first consider the case when $0 < c < \infty$. Then, the condition $\frac{d}{n\delta_M^{(n)}} \to 0$ yields $\delta_M^{(n)} \to \infty$, and thus $\frac{1}{\lambda_i^{(n)}}\lambda_n(\mathbf{B})$ and $\frac{1}{\lambda_i^{(n)}}\lambda_1(\mathbf{B}) \to 0$ for $i \le K$ by Lemma 2.7.4. According to Lemma 2.7.3, we conclude that for $i \le K$

$$\frac{\hat{\lambda}_i}{\lambda_i^{(n)}} \to 1 \text{ a.s.} \tag{2.7.15}$$

because $\hat{\lambda}_i = \lambda_i(\hat{\mathbf{\Sigma}}) = \lambda_i(\hat{\mathbf{\Sigma}}_D)$. For $i > K$, $\lambda_i(\mathbf{A}) = 0$ since the rank of $\mathbf{A}$ is less than or equal to K. Then, by Lemma 2.7.4,

$$c_\lambda(1-\sqrt{c})^2 \le \liminf \hat{\lambda}_i \le \limsup \hat{\lambda}_i \le c_\lambda(1+\sqrt{c})^2 \text{ for } i > K. \tag{2.7.16}$$

Next, consider the case when $c = \infty$. By the condition $\frac{d}{n\delta_M^{(n)}} \to 0$, $\frac{1}{\lambda_i^{(n)}}\lambda_n(\mathbf{B}) = \frac{n}{d}\lambda_n(\mathbf{B}) \times \frac{d}{n\delta_M^{(n)}} \times \frac{\delta_M^{(n)}}{\lambda_i^{(n)}} \to 0$ for $i \leq K$. Similarly, $\frac{1}{\lambda_i^{(n)}}\lambda_1(\mathbf{B}) \to 0$ for $i \leq K$. Thus, we can conclude (2.7.15) for $i \leq K$. For $i > K$, it follows from $\lambda_i(\mathbf{A}) = 0$ and Lemma 2.7.4 that

$$\frac{n}{d}\hat{\lambda}_{K+1} \to c_\lambda \quad \text{and} \quad \frac{n}{d}\hat{\lambda}_n \to c_\lambda \ \text{ a.s.},$$

which gives

$$\frac{n}{d}\hat{\lambda}_i \to c_\lambda \ \text{ a.s. } \ \text{ for } \ i > K.$$

Lastly, consider the case when $c = 0$. In this case, the condition $\frac{d}{n\delta_M^{(n)}} \to 0$ does not guarantee $\delta_M^{(n)} \to \infty$ so that we divide the case into two sub-cases: $\delta_M^{(n)} \to \infty$ and $\delta_M^{(n)} < \infty$. When $\delta_M^{(n)} \to \infty$, (2.7.15) follows similarly to the case when $0 < c < \infty$. When $\delta_M^{(n)} < \infty$, according to Theorem 1 ($c = 0$) of Baik and Silverstein (2006) as mentioned in Shen et al. (2016), (2.7.15) still follows for $i \leq K$. For $i > K$, it is easy to see (2.7.16) with $c = 0$, that is,

$$c_\lambda \leq \liminf \hat{\lambda}_i \leq \limsup \hat{\lambda}_i \leq c_\lambda \ \text{ for } i > K.$$

Hence, we have $\hat{\lambda}_i \to c_\lambda$ for $i > K$. This completes the proof.

Theorem 2.5.1 (b). As mentioned earlier, the condition $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$ implies $\frac{d}{n} \to \infty$. So, we only consider the case $\frac{d}{n} \to \infty$.

(i) $i \leq p_h$: since $\frac{d}{n\delta_h^{(n)}} \to 0$, $\delta_h^{(n)} \to \infty$ and thus $\frac{\lambda_1(\mathbf{B})}{\lambda_i^{(n)}} \leq \frac{\lambda_1(\mathbf{B})}{\delta_h^{(n)}} \times \frac{\delta_h^{(n)}}{\lambda_{p_h}^{(n)}} \to_{a.s.} 0$. Similarly, $\frac{\lambda_n(\mathbf{B})}{\lambda_i^{(n)}} \to_{a.s.} 0$. Then, (2.7.15) follows from (2.7.14) for $i \leq p_h$.

(ii) $i > p_h$: (2.7.14) can be re-expressed as

$$\frac{n}{d}\lambda_i(\mathbf{A}) + \frac{n}{d}\lambda_n(\mathbf{B}) \leq \frac{n}{d}\hat{\lambda}_i \leq \frac{n}{d}\lambda_i(\mathbf{A}) + \frac{n}{d}\lambda_1(\mathbf{B}).$$

We can easily see that $\frac{n}{d}\lambda_i(\mathbf{A}) \leq \frac{n}{d}\lambda_{p_h+1}(\mathbf{A}) = \frac{n\delta_{h+1}^{(n)}}{d} \times \frac{\lambda_{p_h+1}(\mathbf{A})}{\delta_{h+1}^{(n)}} \to_{a.s.} 0$ by the condition $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$. Therefore, it follows from Lemma 2.7.4 that $\frac{n}{d}\hat{\lambda}_i \to_{a.s.} c_\lambda$ for $i > p_h$.

□

2.7.2 Proof of Theorem 2.5.2

Proof. For the subspace consistency of the sample eigenvectors, $\hat{U}_i$ to the S_m, we want to show

$$\text{angle}(\hat{U}_i, S_m) \to_{a.s.} 0. \tag{2.7.17}$$

This is equivalent to showing that $\cos(\text{angle}(\hat{U}_i, S_m)) = \left(\sum_{k\in H_m} \hat{U}_i^T U_k U_k^T \hat{U}_i\right)^{1/2} \to_{a.s.} 1$ for $i \in H_m$ since $S_m = \text{span}\{U_i\}_{i\in H_m}$. Without loss of generality, we can assume $U_k = e_k$ where the kth entry is 1 and the rest of entries are all zero. Then, $\hat{U}_i^T U_k$ is simply $\hat{u}_{ki}$ and therefore (2.7.17) is equivalent to

$$\sum_{k\in H_m} \hat{u}_{ki}^2 \to_{a.s.} 1 \qquad \text{for } i \in H_m. \tag{2.7.18}$$

In general, we will show (2.7.18) for the subspace consistency, but different convergence rates will be achieved under each scenario. For the strong inconsistency of $\hat{U}_i$ to the true eigenvector U_i, we will show that the angle$(\hat{U}_i, U_i) \to_{a.s.} \frac{\pi}{2}$, which is equivalent to showing that $\cos(\text{angle}(\hat{U}_i, U_i)) = \hat{u}_{ii} \to_{a.s.} 0$.

Before we prove the main parts, we first provide some important results that will be used in the proof of the main parts. Define $\mathbf{S} = \mathbf{\Lambda}^{-\frac{1}{2}}\hat{\mathbf{U}}\hat{\mathbf{\Lambda}}^{\frac{1}{2}}$ where $\mathbf{\Lambda} = \text{diag}(\lambda_1^{(n)}, \cdots, \lambda_d^{(n)})$, and then its element for is the kth row and ith column is $s_{ki} = \frac{\sqrt{\hat{\lambda}_i}}{\sqrt{\lambda_k^{(n)}}}\hat{u}_{ki}$. From (2.3.3), we obtain $\mathbf{S}\mathbf{S}^T = \frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{-\frac{1}{2}}$, and thus the kth diagonal element of $\mathbf{S}\mathbf{S}^T$, i.e.

$$(\mathbf{S}\mathbf{S}^T)_{(k,k)} = \sum_{i=1}^{d} s_{ki}^2 = \sum_{i=1}^{d} \frac{\hat{\lambda}_i}{\lambda_k^{(n)}}\hat{u}_{ki}^2 = \frac{1}{\lambda_k^{(n)}}\sum_{i=1}^{d} \hat{\lambda}_i \hat{u}_{ki}^2,$$

is equal to the kth diagonal element of $\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{-\frac{1}{2}} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\lambda_k^{(n)}}y_{kj}^2$. From (2.7.4) and (2.7.5), we get

$$\frac{1}{\lambda_k^{(n)}}\sum_{i=1}^{d}\hat{\lambda}_i\hat{u}_{ki}^2 = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\lambda_k^{(n)}}y_{kj}^2 \rightarrow_{a.s.} 1. \tag{2.7.19}$$

Since all diagonal values of a matrix should be less than its largest eigenvalue, it follows from $\lambda_i(\mathbf{S}\mathbf{S}^T) = \lambda_i(\mathbf{S}^T\mathbf{S})$ that

$$(\mathbf{S}^T\mathbf{S})_{(i,i)} = \sum_{k=1}^{d}s_{ki}^2 = \sum_{k=1}^{d}\frac{\hat{\lambda}_i}{\lambda_k^{(n)}}\hat{u}_{ki}^2 = \hat{\lambda}_i\sum_{k=1}^{d}\frac{1}{\lambda_k^{(n)}}\hat{u}_{ki}^2 \leq \lambda_1(\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{-\frac{1}{2}}). \tag{2.7.20}$$

To use the above inequality in the proof of the main parts, the largest eigenvalue of $\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{-\frac{1}{2}}$ is of interest. If there are no outlier components, $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}$ will consist of i.i.d. random variables with zero mean, unit variance, and finite fourth moment. Then, Lemma 2.7.2 implies that

$$\begin{aligned}
\lambda_1(\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{-\frac{1}{2}}) &\rightarrow_{a.s.} (1+\sqrt{c})^2 \quad \text{for } 0 \leq c < \infty \quad \text{and} \\
\lambda_1(\frac{1}{d}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{-\frac{1}{2}}) &\rightarrow_{a.s.} (1+\frac{1}{\sqrt{c}})^2 \quad \text{for } 0 < c \leq \infty.
\end{aligned}$$

In the case with outliers from the model (2.3.3), however, the entries of the $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}$ are not identically distributed any longer, which brings same challenges. Here, we prove that the maximum eigenvalue of $\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{-\frac{1}{2}}$ still has the same limit even though there are a few outliers from different distributions.

Lemma 2.7.5. *As n, $d \rightarrow \infty$ such that $\frac{d}{n} \rightarrow c$, we have*

$$\begin{aligned}
\lambda_1(\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{-\frac{1}{2}}) &\rightarrow_{a.s.} (1+\sqrt{c})^2 \qquad 0 \leq \mathrm{c} < \infty \\
\lambda_1(\frac{1}{d}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{-\frac{1}{2}}) &\rightarrow_{a.s.} (1+\frac{1}{\sqrt{c}})^2 \qquad 0 < \mathrm{c} \leq \infty.
\end{aligned}$$

Proof. Denote the ith row vector of $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Y}$ by $\breve{Y}_i$. Then, $\frac{1}{d}\mathbf{Y}^T\mathbf{\Lambda}^{-1}\mathbf{Y} = \frac{1}{d}\sum_{i=1}^{d}\breve{Y}_i\breve{Y}_i^T = \frac{1}{d}\sum_{i=1}^{K}\breve{Y}_i\breve{Y}_i^T + \frac{1}{d}\sum_{i>K}\breve{Y}_i\breve{Y}_i^T$. Recall that each element of $\breve{Y}_i$ is $\breve{y}_{ij} = \tau_{i,1}^{(n)} z_{ij}$ for $j \in s_i^c$ and $\breve{y}_{ij} = \tau_{i,2}^{(n)} z_{ij}$ for $j \in s_i$. Let $Z_i = (z_{i1}, \cdots, z_{in})^T$ for $i = 1, \cdots, d$, and then we have

$$\frac{1}{d}\mathbf{Y}^T\mathbf{\Lambda}^{-1}\mathbf{Y} = \frac{1}{d}\Big(\sum_{i=1}^{K}\breve{Y}_i\breve{Y}_i^T - \sum_{i=1}^{K}Z_iZ_i^T\Big) + \frac{1}{d}\Big(\sum_{i=1}^{K}Z_iZ_i^T + \sum_{i=K+1}^{d}\breve{Y}_i\breve{Y}_i^T\Big). \tag{2.7.21}$$

By the Weyl inequality, we have

$$\lambda_1(\frac{1}{d}\sum_{i=1}^{d}\breve{Y}_i\breve{Y}_i^T) \le \lambda_1\Big(\frac{1}{d}\Big(\sum_{i=1}^{K}\breve{Y}_i\breve{Y}_i^T - \sum_{i=1}^{K}Z_iZ_i^T\Big)\Big) + \lambda_1\Big(\frac{1}{d}\Big(\sum_{i=1}^{K}Z_iZ_i^T + \sum_{i=K+1}^{d}\breve{Y}_i\breve{Y}_i^T\Big)\Big). \tag{2.7.22}$$

Letting $\breve{\mathbf{Y}}_K = [\breve{Y}_1, \cdots, \breve{Y}_K]$ and $\mathbf{Z}_K = [Z_1, \cdots, Z_K]$, we get $\frac{1}{d}\Big(\sum_{i=1}^{K}\breve{Y}_i\breve{Y}_i^T - \sum_{i=1}^{K}Z_iZ_i^T\Big) = \frac{1}{d}\Big(\breve{\mathbf{Y}}_K\breve{\mathbf{Y}}_K^T - \mathbf{Z}_K\mathbf{Z}_K^T\Big)$. Then, we have

$$\begin{aligned}
\lambda_1\Big(\frac{1}{d}\Big(\breve{\mathbf{Y}}_K\breve{\mathbf{Y}}_K^T - \mathbf{Z}_K\mathbf{Z}_K^T\Big)\Big) &\le \operatorname{tr}\Big(\frac{1}{d}\Big(\breve{\mathbf{Y}}_K\breve{\mathbf{Y}}_K^T - \mathbf{Z}_K\mathbf{Z}_K^T\Big)\Big) \\
&= \frac{1}{d}\Big[\operatorname{tr}\Big(\breve{\mathbf{Y}}_K\breve{\mathbf{Y}}_K^T\Big) - \operatorname{tr}\Big(\mathbf{Z}_K\mathbf{Z}_K^T\Big)\Big] \\
&= \frac{1}{d}\Big[\operatorname{tr}\Big(\breve{\mathbf{Y}}_K^T\breve{\mathbf{Y}}_K\Big) - \operatorname{tr}\Big(\mathbf{Z}_K^T\mathbf{Z}_K\Big)\Big] \\
&= \frac{1}{d}\Big[\sum_{i=1}^{K}\sum_{j=1}^{n}\frac{1}{\lambda_i^{(n)}}y_{ij}^2 - \sum_{i=1}^{K}\sum_{j=1}^{n}z_{ij}^2\Big] \\
&= \frac{n}{d}\Big[\sum_{i=1}^{K}\frac{1}{n}\sum_{j=1}^{n}\frac{1}{\lambda_i^{(n)}}y_{ij}^2 - \sum_{i=1}^{K}\frac{1}{n}\sum_{j=1}^{n}z_{ij}^2\Big].
\end{aligned}$$

Because $\sum_{i=1}^{K}\frac{1}{n}\sum_{j=1}^{n}\frac{1}{\lambda_i^{(n)}}y_{ij}^2 - \sum_{i=1}^{K}\frac{1}{n}\sum_{j=1}^{n}z_{ij}^2 \to_{a.s.} 0$ and $\frac{n}{d} \to \frac{1}{c}$, we have

$$\lambda_1\Big(\frac{1}{d}\Big(\sum_{i=1}^{K}\breve{Y}_i\breve{Y}_i^T - \sum_{i=1}^{K}Z_iZ_i^T\Big)\Big) \to_{a.s.} 0 \quad \text{for } 0 \le \frac{1}{c} < \infty. \tag{2.7.23}$$

Based on techniques similar to the proof of Lemma 2.7.4, one can show that

$$\lambda_1\Big(\frac{1}{d}\Big(\sum_{i=1}^{K}Z_iZ_i^T + \sum_{i=K+1}^{d}\breve{Y}_i\breve{Y}_i^T\Big)\Big) \to_{a.s.} (1+\frac{1}{\sqrt{c}})^2. \tag{2.7.24}$$

By (2.7.22), (2.7.23), and (2.7.24), we have

$$\lambda_1(\frac{1}{d}\sum_{i=1}^{d}\check{Y}_i\check{Y}_i^T) \leq (1+\frac{1}{\sqrt{c}})^2 \quad \text{almost surely.} \tag{2.7.25}$$

For the lower bound, application of Weyl inequality on the other way gives

$$\lambda_n\Big(\frac{1}{d}\Big(\sum_{i=1}^{K}\check{Y}_i\check{Y}_i^T - \sum_{i=1}^{K}Z_iZ_i^T\Big)\Big) + \lambda_1\Big(\frac{1}{d}\Big(\sum_{i=1}^{K}Z_iZ_i^T + \sum_{i=K+1}^{d}\check{Y}_i\check{Y}_i^T\Big)\Big) \leq \lambda_1(\frac{1}{d}\sum_{i=1}^{d}\check{Y}_i\check{Y}_i^T).$$

Since the rank of $\frac{1}{d}\Big(\sum_{i=1}^{K}\check{Y}_i\check{Y}_i^T - \sum_{i=1}^{K}Z_iZ_i^T\Big)$ is less than or equal to K, we have $\lambda_n\Big(\frac{1}{d}\Big(\sum_{i=1}^{K}\check{Y}_i\check{Y}_i^T - \sum_{i=1}^{K}Z_iZ_i^T\Big)\Big) = 0$, which with (2.7.24) gives

$$(1+\frac{1}{\sqrt{c}})^2 \leq \lambda_1(\frac{1}{d}\sum_{i=1}^{d}\check{Y}_i\check{Y}_i^T) \quad \text{almost surely.} \tag{2.7.26}$$

A combination of (2.7.25) and (2.7.26) completes the proof. □

Now we start to prove Theorem 2.5.2. We will first prove (b) and move on to (a).

proof of (b). Assumption 2.5.4 and $\frac{d}{n\delta_{h+1}^{(n)}} \to \infty$ together imply $\frac{d}{n} \to \infty$. So here we only consider the case of $\frac{d}{n} \to \infty$. The proof consists of the following three steps:

1. Establish the convergence for the h-th tier.

2. Establish the convergence for the m-th tier sequentially from $m = h-1$ to 1.

3. Establish the strong inconsistency of the remaining sample eigenvectors.

We start with the first step.

1. Establish the convergence for the h-th tier, i.e. $\text{angle}(\hat{U}_i, S_h) = o(\{\frac{\delta_h^{(n)}}{\delta_{h-1}^{(n)}}\}^{1/2}) \vee O(\{\frac{d}{n\delta_h^{(n)}}\}^{1/2})$ for $i \in H_h$. As discussed earlier, we need to show the following:

$$\sum_{k\in H_h}\hat{u}_{ki}^2 = 1 + o\Big(\frac{\delta_h^{(n)}}{\delta_{h-1}^{(n)}}\Big) \vee O\Big(\frac{d}{n\delta_h^{(n)}}\Big) \quad \text{for } i \in H_h. \tag{2.7.27}$$

This can be proved by showing the following two equations:

$$\sum_{i=p_h+1}^{d} \hat{u}_{ki}^2 = O\Big(\frac{d}{n\delta_h^{(n)}}\Big) \quad \text{for } k \in H_h \tag{2.7.28}$$

and

$$\sum_{m=1}^{h-1} \sum_{i \in H_m} \hat{u}_{ki}^2 = O\Big(\frac{\delta_h^{(n)}}{\delta_{h-1}^{(n)}}\Big) \quad \text{for } k \in H_h. \tag{2.7.29}$$

Since $\sum_{i=p_h+1}^{d} \hat{u}_{ki}^2 \leq \sum_{k=1}^{p_h} \sum_{i=p_h+1}^{d} \hat{u}_{ki}^2 = \sum_{k=p_h+1}^{d} \sum_{i=1}^{p_h} \hat{u}_{ki}^2$ and p_h is finite, the first equation (2.7.28) is equivalent to

$$\sum_{k=p_h+1}^{d} \sum_{i=1}^{p_h} \hat{u}_{ki}^2 = O\Big(\frac{d}{n\delta_h^{(n)}}\Big) \quad \text{for } m = 1, \cdots, h \tag{2.7.30}$$

for $m = h$. The proof of (2.7.30) is equivalent to showing the following two equations:

$$\sum_{k=K+1}^{d} \sum_{i=1}^{p_m} \hat{u}_{ki}^2 = O\Big(\frac{d}{n\delta_m^{(n)}}\Big) \quad \text{for } m = 1, \cdots, h \tag{2.7.31}$$

and

$$\sum_{k=p_h+1}^{K} \sum_{i=1}^{p_m} \hat{u}_{ki}^2 = O\Big(\frac{\delta_{h+1}^{(n)}}{\delta_m^{(n)}}\Big) \quad \text{for } m = 1, \cdots, h. \tag{2.7.32}$$

For details, see Shen et al. (2016). Thus, in order to show (2.7.27), it is enough to show (2.7.29), (2.7.31), (2.7.32). We start with (2.7.31).

proof of (2.7.31). For $m = 1, \cdots, h$, from $\frac{\hat{\lambda}_{p_m}}{\lambda_{K+1}^{(n)}} \sum_{k=K+1}^{d} \sum_{i=1}^{p_m} \hat{u}_{ki}^2 \leq \sum_{k=K+1}^{d} \frac{1}{\lambda_k^{(n)}} \sum_{i=1}^{p_m} \hat{\lambda}_i \hat{u}_{ki}^2$,

$$\begin{aligned}
\sum_{k=K+1}^{d} \sum_{i=1}^{p_m} \hat{u}_{ki}^2 &\leq \frac{\lambda_{K+1}^{(n)}}{\hat{\lambda}_{p_m}} \sum_{i=1}^{p_m} \hat{\lambda}_i \sum_{k=K+1}^{d} \frac{1}{\lambda_k^{(n)}} \hat{u}_{ki}^2 \\
&\leq \frac{d}{n\delta_m^{(n)}} \frac{\delta_m^{(n)}}{\hat{\lambda}_{p_m}} \frac{n\lambda_{K+1}^{(n)}}{d} \sum_{i=1}^{p_m} \hat{\lambda}_i \sum_{k=1}^{d} \frac{1}{\lambda_k^{(n)}} \hat{u}_{ki}^2 \\
&\leq \frac{d}{n\delta_m^{(n)}} \frac{\delta_m^{(n)}}{\hat{\lambda}_{p_m}} \frac{n\lambda_{K+1}^{(n)}}{d} p_m \lambda_1 \left(\frac{1}{n} \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{Y}\mathbf{Y}^T \mathbf{\Lambda}^{-\frac{1}{2}}\right) \quad \text{by (2.7.20)} \\
&= \frac{d}{n\delta_m^{(n)}} \frac{\delta_m^{(n)}}{\hat{\lambda}_{p_m}} \lambda_{K+1}^{(n)} p_m \lambda_1 \left(\frac{1}{d} \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{Y}\mathbf{Y}^T \mathbf{\Lambda}^{-\frac{1}{2}}\right) \qquad (2.7.33)
\end{aligned}$$

By Theorem 2.5.1(b), Assumption 2.5.1, and Lemma 2.7.5, for large n and d, the last expression of (2.7.33) becomes a constant multiplied by $\frac{d}{n\delta_m^{(n)}}$ that converges to 0. Therefore, the proof of (2.7.31) is complete.

proof of (2.7.32). For $m = 1, \cdots, h$, from $\frac{\hat{\lambda}_{p_m}}{\lambda_{p_h+1}^{(n)}} \sum_{k=p_h+1}^{K} \sum_{i=1}^{p_m} \hat{u}_{ki}^2 \leq \sum_{k=p_h+1}^{K} \frac{1}{\lambda_k^{(n)}} \sum_{i=1}^{p_m} \hat{\lambda}_i \hat{u}_{ki}^2$,

$$\begin{aligned}
\sum_{k=p_h+1}^{K} \sum_{i=1}^{p_m} \hat{u}_{ki}^2 &\leq \frac{\lambda_{p_h+1}^{(n)}}{\hat{\lambda}_{p_m}} \sum_{k=p_h+1}^{K} \frac{1}{\lambda_k^{(n)}} \sum_{i=1}^{p_m} \hat{\lambda}_i \hat{u}_{ki}^2 \\
&\leq \frac{\lambda_{p_h+1}^{(n)}}{\hat{\lambda}_{p_m}} \sum_{k=p_h+1}^{K} \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\lambda_k^{(n)}} y_{kj}^2 \quad \text{by (2.7.19)} \\
&= \frac{\lambda_{p_h+1}^{(n)}}{\delta_{h+1}^{(n)}} \frac{\delta_{h+1}^{(n)}}{\delta_m^{(n)}} \frac{\delta_m^{(n)}}{\lambda_{p_m}^{(n)}} \frac{\lambda_{p_m}^{(n)}}{\hat{\lambda}_{p_m}} \sum_{k=p_h+1}^{K} \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\lambda_k^{(n)}} y_{kj}^2 \qquad (2.7.34)
\end{aligned}$$

For $m = h$, it follows from $\frac{\delta_{h+1}^{(n)}}{\delta_h^{(n)}} \to_{a.s.} 0$, Theorem 2.5.1(b), and (2.7.19), the proof of (2.7.32) is complete.

proof of (2.7.29). From (2.7.19), we obtain

$$\begin{aligned}
\frac{1}{\lambda_k^{(n)}} \sum_{i=1}^{d} \hat{\lambda}_i \hat{u}_{ki}^2 &= \frac{1}{\lambda_k^{(n)}} \sum_{i=1}^{p_{h-1}} \hat{\lambda}_i \hat{u}_{ki}^2 + \frac{1}{\lambda_k^{(n)}} \sum_{i \in H_h} \hat{\lambda}_i \hat{u}_{ki}^2 + \frac{1}{\lambda_k^{(n)}} \sum_{i=p_h+1}^{d} \hat{\lambda}_i \hat{u}_{ki}^2 \\
&\to_{a.s.} 1 \quad \text{for } k \in H_h. \qquad (2.7.35)
\end{aligned}$$

The third part of (2.7.35)

$$\begin{aligned}\frac{1}{\lambda_k^{(n)}}\sum_{i=p_h+1}^{d}\hat{\lambda}_i\hat{u}_{ki}^2 &\leq \frac{\hat{\lambda}_{p_h+1}}{\lambda_{p_h}^{(n)}}\sum_{i=p_h+1}^{d}\hat{u}_{ki}^2 \\ &= \frac{\delta_{p_h}^{(n)}}{\lambda_{p_h}^{(n)}}\frac{n\hat{\lambda}_{p_h+1}}{d}\frac{d}{n\delta_{p_h}^{(n)}}\sum_{i=p_h+1}^{d}\hat{u}_{ki}^2 \\ &\to_{a.s.} \quad 0 \quad \text{for } k\in H_h. \end{aligned} \tag{2.7.36}$$

The convergence follows from Theorem 2.5.1 (b) and (2.7.28). Thus, (2.7.35) becomes

$$\frac{1}{\lambda_k^{(n)}}\sum_{i=1}^{p_{h-1}}\hat{\lambda}_i\hat{u}_{ki}^2+\frac{1}{\lambda_k^{(n)}}\sum_{i\in H_h}\hat{\lambda}_i\hat{u}_{ki}^2\to_{a.s.}1 \quad \text{for } k\in H_h. \tag{2.7.37}$$

Because $\frac{\hat{\lambda}_i}{\lambda_k^{(n)}}\to_{a.s.}\frac{\delta_m^{(n)}}{\delta_h^{(n)}}$ for $i\in H_m$ and $k\in H_h$, (2.7.37) can be rewritten as

$$\sum_{m=1}^{h-1}\sum_{i\in H_m}\frac{\delta_m^{(n)}}{\delta_h^{(n)}}\hat{u}_{ki}^2+\frac{1}{\lambda_k^{(n)}}\sum_{i\in H_h}\hat{\lambda}_i\hat{u}_{ki}^2\to_{a.s.}1 \quad \text{for } k\in H_h. \tag{2.7.38}$$

Also, since $\sum_{i=1}^{d}\hat{u}_{ki}^2=1$ and $\sum_{i=p_h+1}^{d}\hat{u}_{ki}^2\to_{a.s.}0$ for $k\in H_h$ by (2.7.28), we have

$$\sum_{m=1}^{h-1}\sum_{i\in H_m}\hat{u}_{ki}^2+\frac{1}{\lambda_k^{(n)}}\sum_{i\in H_h}\hat{\lambda}_i\hat{u}_{ki}^2\to_{a.s.}1 \quad \text{for } k\in H_h. \tag{2.7.39}$$

From (2.7.38), (2.7.39), and $\frac{\delta_m^{(n)}}{\delta_h^{(n)}}\geq 1$, we get $\sum_{m=1}^{h-1}\sum_{i\in H_m}\hat{u}_{ki}^2\to_{a.s.}0$ and $\sum_{i\in H_h}\hat{u}_{ki}^2\to_{a.s.}1$. This already shows the subspace consistency, and for convergence rate of (2.7.29), $\sum_{m=1}^{h-1}\sum_{i\in H_m}\hat{u}_{ki}^2\leq\sum_{m=1}^{h-1}\sum_{i\in H_m}\frac{\delta_m^{(n)}}{\delta_h^{(n)}}\hat{u}_{ki}^2\leq(h-1)\frac{\delta_{h-1}^{(n)}}{\delta_h^{(n)}}\sum_{i\in H_m}\hat{u}_{ki}^2\to_{a.s.}0$ for $m=1,\cdots,h-1$. Hence, we complete the proof of the step 1. We move on to Step 2.

2. Establish the convergence for the m-th tier sequentially from $m=h-1$ to 1, i.e. $\text{angle}(\hat{U}_i,S_m)=o(\{\frac{\delta_m^{(n)}}{\delta_{m-1}^{(n)}}\vee\frac{\delta_{m+1}^{(n)}}{\delta_m^{(n)}}\}^{1/2})$ for $i\in H_m$ for each m.

We want to show

$$\sum_{k\in H_m} \hat{u}_{ki}^2 = 1 + o\Big(\frac{\delta_m^{(n)}}{\delta_{m-1}^{(n)}} \vee \frac{\delta_{m+1}^{(n)}}{\delta_m^{(n)}}\Big) \quad \text{for } i \in H_m,\ m = 1, \cdots, h-1,$$

which is equivalent to showing

$$\sum_{i\in H_m} \hat{u}_{ki}^2 = 1 + o\Big(\frac{\delta_m^{(n)}}{\delta_{m-1}^{(n)}} \vee \frac{\delta_{m+1}^{(n)}}{\delta_m^{(n)}}\Big) \quad \text{for } k \in H_m,\ m = 1, \cdots, h-1. \tag{2.7.40}$$

Let us start with $m = h-1$. We have

$$\sum_{i\in H_{h-1}} \hat{u}_{ki}^2 = 1 - \sum_{m=1}^{h-2} \sum_{i\in H_m} \hat{u}_{ki}^2 - \sum_{i=p_{h-1}+1}^{d} \hat{u}_{ki}^2, \tag{2.7.41}$$

and thus we will show $\sum_{m=1}^{h-2} \sum_{i\in H_m} \hat{u}_{ki}^2$ and $\sum_{i=p_{h-1}+1}^{d} \hat{u}_{ki}^2$ both converge to 0. Since $\sum_{i=p_{h-1}+1}^{d} \hat{u}_{ki}^2 \le \sum_{k=1}^{p_{h-1}} \sum_{i=p_{h-1}+1}^{d} \hat{u}_{ki}^2 = \sum_{k=p_{h-1}+1}^{d} \sum_{i=1}^{p_{h-1}} \hat{u}_{ki}^2 = \sum_{k=p_{h-1}+1}^{p_h} \sum_{i=1}^{p_{h-1}} \hat{u}_{ki}^2 + \sum_{k=p_h+1}^{d} \sum_{i=1}^{p_{h-1}} \hat{u}_{ki}^2$, it follows from $\sum_{k=p_{h-1}+1}^{p_h} \sum_{i=1}^{p_{h-1}} \hat{u}_{ki}^2 = o\Big(\frac{\delta_h^{(n)}}{\delta_{h-1}^{(n)}}\Big)$ by (2.7.29) and $\sum_{k=p_h+1}^{d} \sum_{i=1}^{p_{h-1}} \hat{u}_{ki}^2 = O\Big(\frac{d}{n\delta_{h-1}^{(n)}}\Big)$ by (2.7.30) that

$$\sum_{i=p_{h-1}+1}^{d} \hat{u}_{ki}^2 = o\Big(\frac{\delta_h^{(n)}}{\delta_{h-1}^{(n)}}\Big). \tag{2.7.42}$$

From $\sum_{m=1}^{h-2} \sum_{i\in H_m} \hat{u}_{ki}^2 + \sum_{i\in H_{h-1}} \hat{u}_{ki}^2 + \sum_{i=p_{h-1}+1}^{d} \hat{u}_{ki}^2 = 1$, we have

$$\sum_{m=1}^{h-2} \sum_{i\in H_m} \hat{u}_{ki}^2 + \sum_{i\in H_{h-1}} \hat{u}_{ki}^2 \to_{a.s.} 1 \tag{2.7.43}$$

by (2.7.42). Also, by (2.7.19), we have

$$\frac{1}{\lambda_k^{(n)}} \sum_{i=1}^{d} \hat{\lambda}_i \hat{u}_{ki}^2 = \frac{1}{\lambda_k^{(n)}} \sum_{i=1}^{p_{h-2}} \hat{\lambda}_i \hat{u}_{ki}^2 + \frac{1}{\lambda_k^{(n)}} \sum_{i\in H_{h-1}} \hat{\lambda}_i \hat{u}_{ki}^2 + \frac{1}{\lambda_k^{(n)}} \sum_{i=p_{h-1}+1}^{d} \hat{\lambda}_i \hat{u}_{ki}^2 \to_{a.s.} 1 \quad \text{for } k \in H_{h-1}.$$

By $\lim_{n\to\infty} \frac{\lambda_{p_{h-1}+1}^{(n)}}{\lambda_k^{(n)}} \to_{a.s.} 0$, we get $\frac{1}{\lambda_k^{(n)}} \sum_{i=p_{h-1}+1}^{d} \hat{\lambda}_i \hat{u}_{ki}^2 \leq \frac{\lambda_{p_{h-1}+1}^{(n)}}{\lambda_k^{(n)}} \sum_{i=p_{h-1}+1}^{d} \hat{u}_{ki}^2 \to_{a.s.} 0$, which leads to

$$\sum_{m=1}^{h-2} \sum_{i\in H_m} \frac{\delta_m^{(n)}}{\delta_{h-1}^{(n)}} \hat{u}_{ki}^2 + \sum_{i\in H_{h-1}} \hat{u}_{ki}^2 \to_{a.s.} 1 \quad \text{for } k \in H_{h-1} \tag{2.7.44}$$

since $\lim_{n\to\infty} \frac{\hat{\lambda}_i}{\lambda_k^{(n)}} \to \frac{\delta_m^{(n)}}{\delta_{h-1}^{(n)}}$ for $i \in H_m$ and $k \in H_{h-1}$. Combining (2.7.43) and (2.7.44) with $\frac{\delta_m^{(n)}}{\delta_{h-1}^{(n)}} > 1$ for $m = 1, \cdots, h-2$ yields $\sum_{m=1}^{h-2} \sum_{i\in H_m} \frac{\delta_m^{(n)}}{\delta_{h-1}^{(n)}} \hat{u}_{ki}^2 \to_{a.s.} 0$. Therefore, we have $\frac{\delta_{h-2}^{(n)}}{\delta_{h-1}^{(n)}} \sum_{i=1}^{p_{h-2}} \hat{u}_{ki}^2 \leq \sum_{i=1}^{p_{h-2}} \frac{\delta_m^{(n)}}{\delta_{h-1}^{(n)}} \hat{u}_{ki}^2 \to_{a.s.} 0$, which gives

$$\sum_{m=1}^{h-2} \sum_{i\in H_m} \hat{u}_{ki}^2 = o\Big(\frac{\delta_{h-1}^{(n)}}{\delta_{h-2}^{(n)}}\Big) \quad \text{for } k \in H_{h-1}. \tag{2.7.45}$$

Hence, (2.7.40) with $m = h-1$ follows from (2.7.41), (2.7.42) and (2.7.45) and the proofs for the other $m = h-1, \cdots, 1$ are similar to $m = h-1$, and thus omitted. Lastly, we complete the proof of Theorem 2.5.2 (b) by showing Step 3.

3. Establish the strong inconsistency of the remaining sample eigenvectors, i.e. $| < \hat{U}_i, U_i > | = O(\{\frac{n\lambda_i^{(n)}}{d}\}^{1/2})$ for $i > p_h$. From (2.7.19), we have

$$\max_{i>p_h} \frac{n\hat{\lambda}_i}{d} \frac{d}{n\lambda_i^{(n)}} \hat{u}_{ii}^2 = \max_{i>p_h} \frac{\hat{\lambda}_i}{\lambda_i^{(n)}} \hat{u}_{ii}^2 \leq \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\lambda_k^{(n)}} \hat{y}_{ki}^2 \to_{a.s.} 1.$$

Since $\frac{n\hat{\lambda}_i}{d} \to_{a.s.} c_\lambda$ for $i > p_h$ by Theorem 2.5.1 (b), we obtain $\max_{i>p_h} \hat{u}_{ii}^2 = O\Big(\frac{n\lambda_i^{(n)}}{d}\Big)$, which completes the proof of Theorem 2.5.2 (b).

□

CHAPTER 3. SCISSOR: SHAPE CHANGES IN SELECTING SAMPLE OUTLIERS IN RNA-SEQ

3.1 Motivation and Challenges

A substantial proportion of human genes differ in function in ways that are reflected through different forms of shape changes in read coverage of RNA-seq. For example, tumor-suppressor genes lose their function through various changes in expression such as aberrant splicing, frameshift indels, large deletions, or overexpression of noncoding RNAs. Although many recent reports (Zhang et al., 2014; Dvinge and Bradley, 2015; Jung et al., 2015; Wong et al., 2016) have addressed the importance of such shape changes in cancer, the systematic discovery of these events is still challenging.

This is not simply because shape changes are rare. As discussed in Chapter 1.3, high dimensionality of RNA-seq data often makes good outliers indistinguishable from inliers. To detect such high dimensional outliers, it is important to specify what type of aberration is of interest. However, changes in RNA-seq expression present many different forms each of which results from various reasons. For example, genetic mutations are responsible for some important shape changes. The splice site mutations, i.e. single-nucleotide variants (SNVs) at a splice site (exon-intron boundary), are known to cause abnormal splicing events such as intron retention or exon-skipping. This abnormal splicing often skips a whole exon or retains a whole intron, which tends to change a wide range of the gene. On the other hand, the frameshift mutations, i.e. indels (insertions or deletions) of nucleotides that can change the reading frame (a set of consecutive triplets in the sequence of nucleotides), are occasionally responsible for the deletion of read coverage in a relatively short region (30$\sim$200 bp) in the middle of an exon. It has been also observed that a number of shape changes occur in the absence of mutations.

Classically, differential expression (DE) analysis has been studied to identify changes in expression using properly normalized total read counts data (Anders and Huber, 2010; Robinson and Oshlack, 2010; Love et al., 2014). The DE analysis particularly compares gene-level expression between samples with known labels, e.g. treated versus untreated cells, cancer versus normal, wild-type and mutant (Oshlack et al., 2010). However, gene expression analysis provides a limited view of RNA-seq data by summarizing the fine architecture of transcriptome as single numbers for each gene. Some biologically important changes in expression that are too weak to be captured by single numbers can be missing using gene-level expression analysis. In this work, we introduce a new statistical method, Scissor (Shape Changes In Selecting Sample Outliers in RNA-seq) that uses base-resolution expression levels across a gene locus. The base-level RNA-seq data provide rich information beyond gene expression analysis, allowing us to detect fine structural changes such as frameshift indels besides landscape changes such as intergenic deletions. By using the complex and dynamic aspect of the transcriptome, Scissor offers a novel approach to unsupervised screening of a variety of shape changes that are possibly associated with important genetic events.

To identify various types of shape changes, we model the underlying mechanism possibly generating aberrant shapes by multiple unknown mixture distributions introduced in Section 2.3, recasting the problem in a high-dimensional latent variables framework. Then, the latent outlying structures described by outlier directions are distinctly or in combination associated with various shape changes. Based on this model, Scissor aims to identify sample outliers that have distinct forms or patterns of transcripts across RNA-seq cohorts from a single gene locus.

We have analyzed 522 TCGA head and neck squamous cell carcinoma (HNSCC) RNA-seq tumor samples. The in-depth analysis at frequently mutated tumor suppressor genes has detected several genes with novel shape changes including TP53, which appear to be associated with new splicing variants that were missing from the current variant callers. The genome-wide study has shown that some mutation types are strongly associated with shape changes. Moreover, the analysis has provided a set of key genes with strong evidence of shape variants, which shows that Scissor can

be used to prioritize genes for further investigation. We have also found that a substantial number of shape changes including intron retention and exon-skipping occur in the absence of mutations.

The remainder is organized as follows. In Section 3.2, we review several outlier detection methods associated with our study. Section 3.3 introduces a model for RNA-seq read counts using the underlying distribution possibly generating various types of RNA-seq shape outliers discussed in Chapter 2.3. The pre-processing part of Scissor including filtering and normalization steps are described in Section 3.4. Scissor is based on the theoretical results established in Section 2.5. In particular, we propose a new approach to high-dimensional outlier detection in Section 3.5. Then, we apply the proposed method to collections of HNSCC RNA-seq samples in Section 3.6.

3.2 Related work

Although numerous procedures have been suggested for identification of outliers, still there is no consensus definition for outliers. Every procedure may target its own informal definition of outliers based on various goals. Here, we describe several popular outlier detection methods highly related to our goal. Many classical methods for detecting outliers in low dimensional data (mostly for univariate data) are reviewed by Barnett and Lewis (1974) and Hawkins (1980). Chapter 10 of Jolliffe (2002) also reviews some outlier detection methods for multivariate data using PCA.

Two popular outlier-detection methods in the classical domain, are PCA-based methods proposed by Hawkins (1974, 1980). These two statistics for detecting outliers are based on the normalized PC scores corresponding to the last q principal components with low-variances because the low-variance PCs may be considered as the most effective directions in determining outliers, whereas the first few PCs are relatively insensitive to outliers and thus are discarded. The proposed statistics are

$$H_{1j}^2 = \sum_{k=d-q+1}^{d} \frac{\hat{y}_{kj}^2}{\hat{\lambda}_k}, \quad H_{2j} = \max_{d-q+1 \le k \le d} \frac{\hat{y}_{kj}}{\sqrt{\hat{\lambda}_k}}$$

where $\hat{y}_{kj}$ is the value of the kth PC score for the jth observation, $\hat{\lambda}_k$ is the kth eigenvalue, and d is the dimension size. These statistics perform very well for low dimensional data and have been widely used. In high dimensional space, however, the intuition behind the Hawkins' statistic might not work. One main reason is that the last few PCs may be just noise related directions and unrelated to outliers in high dimensional data. As discussed in Chapter 1.3, when outlier signals are fairly strong, the PC directions related to outliers may have large enough variation to be included in the first few PC directions. Although the intuition may not work, the Hawkins' statistic H_{1j}^2 based on other PC directions, instead of the last few, can give meaningful insights in some sense and we will revisit this issue in Section 3.5.1.

Another important outlier detection technique is based on the *projection depth function* that was introduced by Stahel (1981) and has been extensively studied in Donoho and Gasko (1992); Liu (1992); Zuo and Serfling (2000); Zuo (2003); Dang and Serfling (2010). The projection depth function generalizes a robust measure of outlyingness for one-dimensional datasets $\mathbf{x} = \{x_1, \cdots, x_n\}$, i.e. $o(x|\mathbf{x}) = \frac{x - \text{Med}_{1\le j\le n}(x_j)}{\text{MAD}_{1\le j\le n}(x_j)}$, to arbitrary dimension d, i.e.

$$o(X|\mathbf{X}) = \sup_{\|h\|=1} \Big| \frac{h^T X - \mu(F_{h^T\mathbf{X}})}{\sigma(F_{h^T\mathbf{X}})} \Big|, \tag{3.2.1}$$

where $F_{h^T\mathbf{X}}$ is the distribution of $h^T\mathbf{X}$ and $\mu(\cdot)$ and $\sigma(\cdot)$ are any univariate location and scale measures, respectively. Intuitively, the $o(X|\mathbf{X})$ is the most outlyingness of X over any one-dimensional projection of X with respect to the dataset $\mathbf{X}$ and thus it is also called a *projection outlyingness*. The projection depth function $o(X|\mathbf{X})$ with $(\mu, \sigma) = (\text{Med}, \text{MAD})$ is affine invariant and has a very high breakdown point. Dang and Serfling (2010) showed that the projection outlyingness is superior than some other important outlyingness functions (the half space, the Mahalanobis distance, and the Mahalanobis spatial) in the sense that it maintains a low false positive rate with a high masking breakdown point, which is a robustness criterion introduced in their paper. Although the projection outlyingness has received significant attention, it has been rarely used for detecting high dimensional

outliers because its optimization procedure is computationally so intensive that it is prohibitively costly to compute for high dimensional data.

When the underlying projected distribution is asymmetric, the projection depth function with the MAD may result in false positive errors for the skewed side and false negative errors on the other side since it does not capture the asymmetry. Rousseeuw et al. (2016) proposed a new measure of the projection outlyingness taking the skewness into account. They also propose a method to detect outliers in functional data by computing outlyingness at each gridpoint and taking a weighted average of them. This approach may perform well when the global structure of functions are similar and aberrations are only locally observed. However, if an outlier is different in its overall structure, then the method may miss it.

Dai and Genton (2016) introduced a graphical tool for detecting outliers for functional data. They proposed functional directional outlyingness by generalizing the statistical depth idea to multivariate functional data to capture the direction of outlyingness in addition to the point-wise scalar depth. Functional directional outlyingness is decomposed into two parts: magnitude outlyingness and shape outlyingness. Magnitude outlyingness quantifies the levels of overall shift whereas shape outlyingness measures the levels of deviation from the median shape pattern by aggregating the variation of outlyingness. The proposed magnitude-shape (MS) plot is a two-dimensional visualization of the two measures of outlyingness, allowing the inspection of shifted outliers and shape outliers. Compared to Rousseeuw et al. (2016), this method shows better performance for multivariate functional data or outliers that are outlying in a large part of the domain. For univariate and high-dimensional data, however, this method may also fail to detect important shape outliers as the shape outlyingness of the outliers in high-dimensional space are less distinguishable due to cumulative noise of the other curves.

In this chapter, we apply the projection depth function to an appropriate low rank subspace in high dimensional space. This is possible because our goal is to detect outliers whose signals are substantially different from the others and such signals tend to be well captured in a subset of the principal component directions under certain conditions as will be seen later. Our method

identifies outliers that are different globally as well as locally from the others. In addition, our method provides a direction for each observation which makes it maximally outlying.

3.3 Model and Notation

A variety of shape changes in mRNA product are reflected by the base-resolution read depths. In the light of this, Scissor benefits from the fine structure of the transcriptome by taking per-base read depths at a single gene locus as an input. In this section, we discuss the underlying statistical framework modelling the base-level expression coverage with possible outliers, which will be the basis of the outlier detection.

At a given gene, Scissor starts with a read count matrix R with one row for each base-position and one column for each sample. That is, the matrix entries R_{ij} indicate the number of reads mapped to the base position i ($1 \leq i \leq d$) in sample j ($1 \leq j \leq n$). Here, d denotes the total length of the transcript considered and n denotes the sample size. Samples at a given gene often share the overall expression patterns in a log-linear way, i.e. linearly associated after log-transformation. Therefore, for each gene, we fit a log-linear model as follows.

We model read counts R_{ij} as following a multiplicative framework in a high-dimensional setting:

$$R_{ij} \approx \mu_i^{a_j} \cdot m_{ij} \tag{3.3.1}$$

with mean $\mu_i > 0$, geometrically scaled by a normalization factor a_j, and the other sources of variation m_{ij} including alternative shape variants. The mean is the geometrical mean of the read counts at a base-position and the normalization factor accounts for differences in sequencing depth between samples. The part $\mu_i^{a_j}$ naturally models the different variability across the samples expressed at different levels. The remaining variation m_{ij} is further modeled by $\log(m_{ij}) = g(a_j)x_{ij}$ with a smooth curve $g(.)$ and high-dimensional random vectors $X_j \sim P_d$ where $X_j = (x_{1j}, \cdots, x_{dj})^T$. It has been observed that different levels of overall expression yield further dynamics that are not captured

solely by the part $\mu_i^{a_j}$. We model such dynamics by a smooth curve $g(a_j)$ as a function of an overall expression level satisfying a constraint $g(1) = 1$ for identifiability. Finally, the X_j's describe other sources of variation beyond overall expression levels, including diverse genetic events that may lead to shape changes in expression. Therefore, the X_j's involve the critical information for detecting outliers and are further modeled as follows.

As discussed in the previous chapters, high-dimensional random vectors are often represented by a linear combination of a set of underlying signal directions plus noise. In many cases, the signal directions capture most of the important variation in the data and the estimation of these signal directions are of great interest. Chapter 1.3 introduced an extended model (2.3.3) embracing some directions that possibly generate outliers. The variation from each of these latent outlier directions is modeled by a scale mixture distribution and an outlier is defined as a data point characterized by big contribution of some of these outlier components. As a motivating example, we already introduced how the underlying structure of RNA-seq data including outlying structure can be modeled by this complicated mixture model (See Section 2.3). From this, we model the variation of X_j by the high-dimensional mixture distribution (2.3.3).

Under this model, Scissor aims to detect sample outliers whose forms or patterns in expression coverage are significantly different from the majority of samples. For doing this, it is crucial to extract the X_j's that contain the crucial information of outlying structure by removing irrelevant variation from the raw coverage data. In the next section, we introduce the pre-processing procedure of Scissor to achieve this goal and illustrate each step with real data examples.

3.4 Pre-processing data

A collection of 522 HNSCC RNA-seq observations were obtained from the TCGA Research Network. The dataset was processed as described in Network (2012). As mentioned earlier, it is crucial to get a reliable data set that fits the statistical framework while reducing irrelevant variation. For the purpose of mining shape outliers, we assume that most samples at a single gene locus have similar expression patterns except for the ones with biologically important shape changes. However,

there are some genes and cases that violate this assumption. For example, degraded RNA samples often produce different shape patterns at a substantial number of genes, which can confound the real shape changes. Also, different sequencing depth between samples often produce extreme skewness and different levels of dynamics that should be adjusted to get reliable results.

As parts of the pre-processing steps of Scissor, we filter out some genes and cases that do not fit the statistical model and normalize the remaining samples to adjust sequencing depth. The main goal of the pre-processing steps of Scissor is to get x_{ij} from the raw read counts R_{ij} in (3.3.1). The pre-processing procedure is divided into several steps: filtering genes and cases, selecting intronic parts, transforming the data, and normalizing. The filtering step is based on genome-wide analysis and should be completed before the downstream per-gene analysis. Once the genes and cases are filtered out, the rest of the pre-processing steps and the outlier detection algorithm are performed for the remaining genes and cohorts.

A pipeline of the pre-processing procedure starting from the BAM files is given in Figure 3.3. The two filtering steps have a pink background because they are usually done before the per-gene analysis. To the right side of the pipeline, real data examples are shown to illustrate each step. In this figure, the focus is entirely on the gene TP53, which is known as the most frequently mutated tumor-suppressor gene. To construct the base-level RNA-seq data, read-depths for a single gene can be measured at each base-position along the length of the transcript, and the resulting data structure is an expression count matrix with one row for each position and one column for each patient. Here, we use the union of the existing transcripts to avoid missing regions not covered by some transcripts. The expression counts matrix can be obtained from BAM files using the SAMtools package Li et al. (2009). A set of 30 expression profiles were randomly chosen after excluding the suspect outliers out of 522 patients. These 30 samples are overlaid as curves colored as gray in Figure 3.3 (a). This figure also displays three examples of degraded RNA samples using different colors, showing completely distinct structure from the gray samples. The boundaries of each exonic region are indicated by the gray vertical lines where parts of introns will be included later. Notice that there is an exon with no coverage exon near 1000, and this is because we employ the TCGA

gene model which includes this exon even though this exon is not covered in this gene. The rest of the figures from (b)-(e) will be discussed in detail in the following subsections.

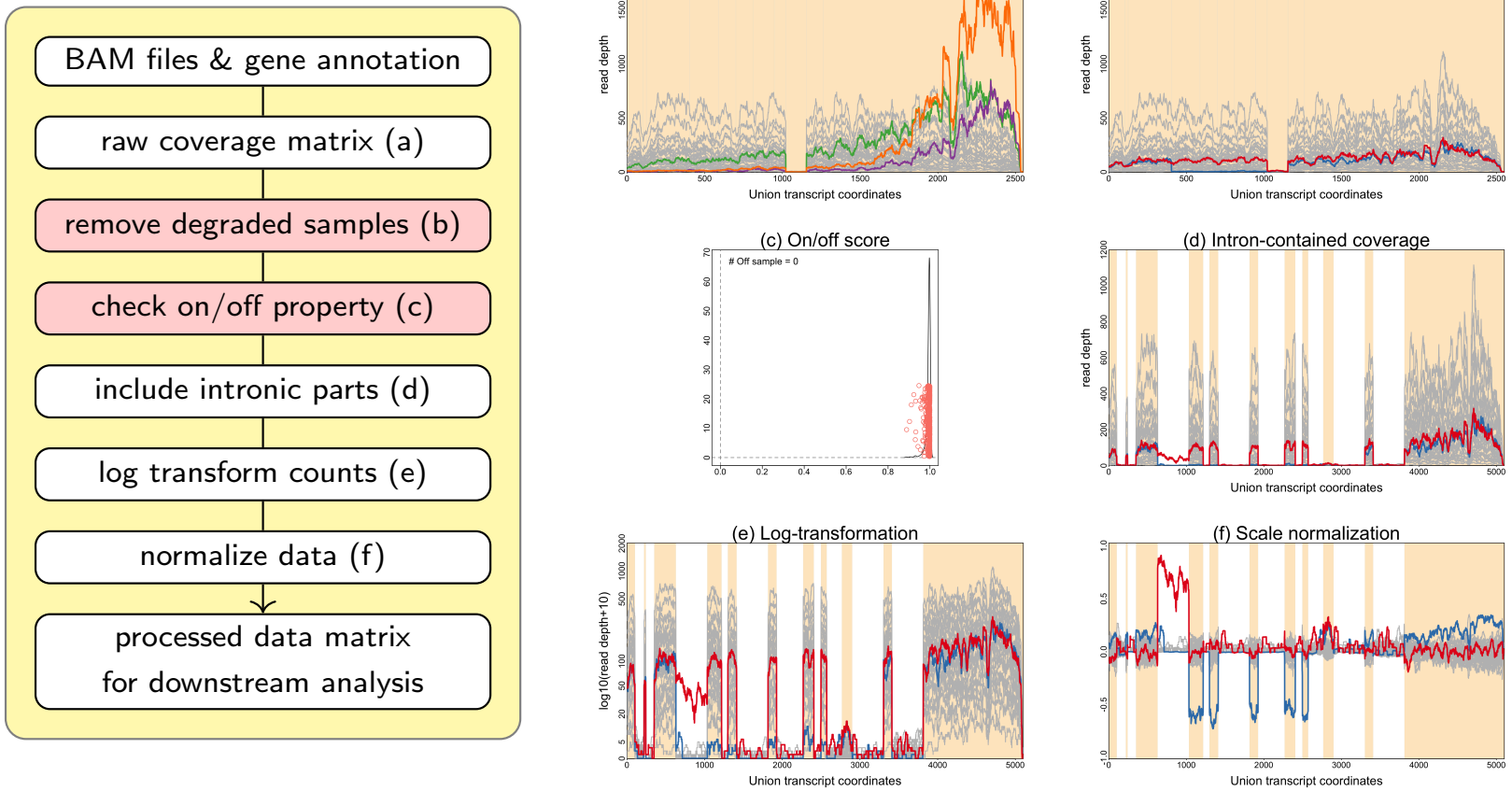


Figure 3.3: The pre-processing steps of Scissor at gene TP53 as an example. (Left) The pipeline for pre-processing data is shown starting with BAM files to get the data object that will be used for detecting outliers. The steps (b) and (c) are colored by pink background because these steps are usually completed before the single gene analysis. (Right) Each step from (a)-(f) is illustrated. (a) The raw coverage data obtained at each base-position along the union of existing transcripts. We randomly selected 30 cases (gray) after excluding the suspect outliers. Three degraded cases (colored) are included to indicate the different structure from degradation. (b) The intact cases after removing degraded cases. We included two suspect outliers colored by blue and red. (c) On/off scores for gene TP53. See Section 3.4.2 for details. (d) The intron-included raw coverage data. (e) Log transformed data using the chosen shift parameter 10. (f) Using the scale normalization, we obtained the normalized data. The red and blue outliers are more separate from the other grey non-outliers after the pre-processing procedure.

3.4.1 Filtering out degraded samples

Degradation of RNA transcripts could confound subsequent analyses, especially if data subjected to different amounts of degradation are naively compared against each other (Romero et al., 2014). Figure 3.3 (a) shows the three suspicious degraded RNA-seq observations indicated by different colors with the other non-degraded observations in gray. It is well known that sequencing degraded RNA samples often leads to less read coverage at the 5' end of the gene and negatively affect subsequent analyses such as transcript quantification, gene expression profiles, and fusion detection (Opitz et al., 2010; Romero et al., 2014; Davila et al., 2016). In particular, this leaves degraded RNA-seq samples susceptible to being considered as shape outliers, which could confound real biological aberrations from the high quality samples. In this subsection, we propose a method to quantify the level of degradation in each case at each gene and identify a group of cases that were globally degraded at a considerable number of genes.

A recent study (Davila et al., 2016) reports that the transcript coverage of degraded samples show exponential decrease as a function of the distance from the 3' end of mRNA that more highly degraded samples show a faster rate of decrease. This motivates us to measure the extent of degradation, also called *decay rate*, by the mean-corrected slope of log-transformed RNA-seq data. To accurately assess the decay rates, we first adjusted the different sequencing depths at each locus by using the first step of the scale normalization method. This procedure helps to remove the intrinsic slopes, allowing for high quality RNA-seq samples to be free of decreasing trend from the 3' end so that the remaining trend can be observed only in a set of degraded samples. Therefore, it enables more accurate comparison of the decay rates across genes by adjusting the other sources possibly affecting slopes. After the adjustment, we fitted a linear model to the mean corrected coverage with the ordered base positions as a covariate for each sample. Let $q_{ij} = q_j(i)$ be the mean corrected coverage for the jth observation ($1 \leq j \leq n = 522$) where i indexes base position at a given locus ($1 \leq i \leq d$) of which total length is d. The linear model

$$q_{ij} = q_j(i) = \alpha_j + \beta_j \times (\frac{i}{d}) + \varepsilon_{ij}$$

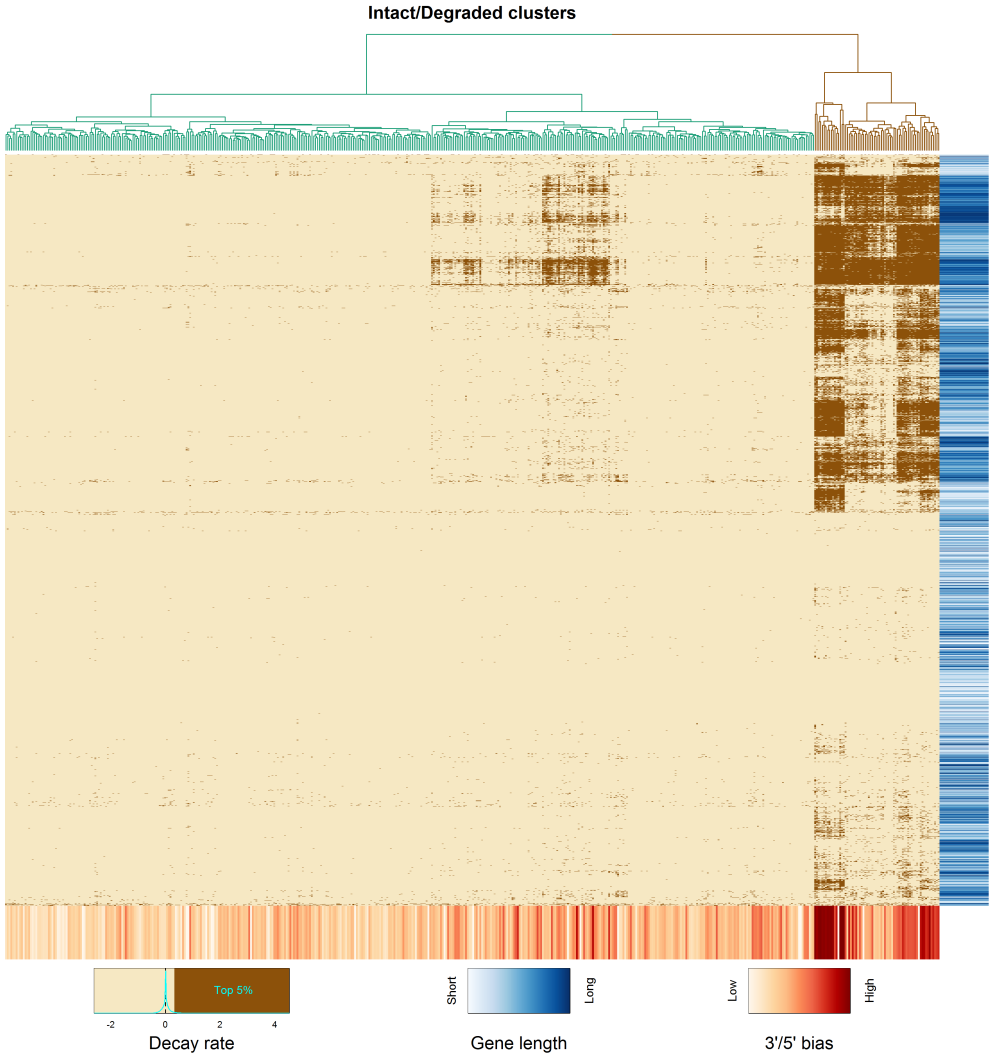


Figure 3.4: Heatmap of decay rates for all genes and all tumor cases, in which the brown color indicates the top 5% decay rates whereas the beige color indicates the lower values. The vertical color bar on the right of the heatmap encodes the gene lengths and a darker blue denotes a longer gene. The horizontal color bar on the bottom of the heatmap encodes the decay ranks from the 3'/5' bias method of Abeshouse et al. (2015) and a darker red denotes higher decay rank. Based on two groups from the hierarchical clustering on the top of the heatmap, the 70 identified degraded cases are in brown and the rest of the intact cases are in green.

was fitted and the least square estimates $\hat{\alpha}_j$ and $\hat{\beta}_j$ were obtained. Note that we divided the covariate i by d to correct the effect of gene length. Then, the $\hat{\beta}_j$ is the decay rate of the jth observation with a higher value of $\hat{\beta}_j$ indicating severe degradation.

We obtained $n = 522$ decay rates at each gene and collected those values across genes as a large matrix of which columns are samples and rows are genes. Unsupervised hierarchical cluster

analysis was performed with this matrix using hclust in R/Bioconductor with the complete linkage method. Figure 3.4 displays a heatmap of the decay rate matrix based on the order of unsupervised hierarchical clustering for both rows and columns. The samples were classified to two groups as shown in the dendrogram above the heatmap and samples in the second group indicated by the brown color clearly show higher decay rates at more than 50% of the genes. Based on this cluster analysis, we identified 70 RNA-seq samples with strong evidence of degradation and excluded them from the downstream analysis. The vertical color code on the right of the heatmap shows lengths of genes in which darker blue indicates longer genes. This color code shows that longer genes tend to undergo more degradation, which is expected since long genes are fully affected by degradation whereas short genes are less affected (Davila et al., 2016). As an illustration, a set of intact/degraded RNA-seq overlays for a short gene, LGALS1, and a long gene, FAT1, are displayed in Figure 3.5. The gene LGALS1 shows no clear changes in expression between the two groups whereas the gene FAT1 shows severe impact of degradation on the second group. We also found strong association between our decay rates and the 3'/5' biases from Abeshouse et al. (2015). The latter are shown using a color code at the bottom of the heatmap in Figure 3.4. A darker red represents severe degradation based on the 3'/5' bias and the samples with dark red color tend to have high decay rates.

The three suspicious degraded samples in Figure 3.3 (a) are all identified as degraded, and the remaining intact samples are shown in Figure 3.3 (b) with two colored suspicious shape outliers (red and blue).

3.4.2 Filtering out on/off genes

In most genes, RNA-seq coverage across samples share analogous global pattern with the exception of a small number of samples that we call shape outliers. On the other hand, there is a set of distinctive genes where overall structures of expression coverage dramatically differ between samples, which do not fit the underlying model and often confound biologically meaningful shape outliers. Here, we identify those genes and filter them out from downstream analysis. Although the

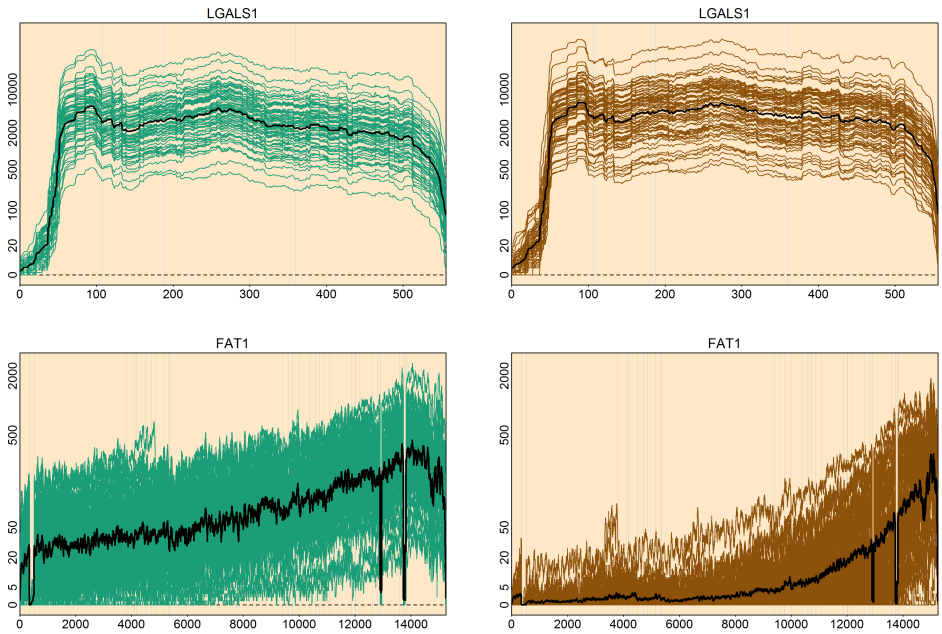


Figure 3.5: Intact and degraded samples are shown in two genes with different lengths. A set of intact samples are shown in green color on the left hand side and all 70 degraded samples identified in this section are shown in brown color on the right hand side. In each figure, a black curve represents the point-wise mean. The degraded samples (left) in the longer gene FAT1 (top row) show severe degradation whereas the intact and degraded cases in the shorter gene LGALS1 (bottom row) do not show a clear difference.

genes identified from this section are excluded from our main analysis, we provide interesting gene sets that are associated with head and neck cancer types.

It has been observed that the genes where samples do not share an analogous pattern are roughly categorized into two groups: (1) the genes with extremely low coverage across samples and (2) the genes where a group of samples are well expressed whereas the rest of samples are rarely expressed. These genes commonly have a considerable number of samples that appear to be off, and we call such genes *on/off genes*. To identify on/off genes, we measure the level of shape-similarity among samples in the context of angles in a high-dimensional space between each individual vector and the mean vector at a given gene. A larger angle from the mean vector indicates that the corresponding sample presents a higher dissimilarity from the other samples. The angle approach allows a lowly-expressed sample along with the global structure that is shared among

the other samples to be considered as being "on" at the given gene. This helps to distinguish the signal-involved low coverage from noise.

For easier interpretation, we take the cosine transformation of the angles, which results in the values ranging from 0 to 1 with a higher value indicating closer to "on" whereas a lower value indicates "off". Then, the resulting values can be considered as the explained amounts of individual samples by the mean vector. We call these values the "level of shape-similarity", denoted by lss_j where j indexes samples. The identification procedure is described as follows:

For each gene,

1. Transform the base-level coverage counts, R_{ij} for base i and sample j, by $z_{ij} = \log_{10}(R_{ij} + 10) - \log_{10}(10)$. This shift parameter of 10 gave good experimental performance.

2. Define the 0-adjusted mean as $m_i^0 = \frac{1}{\sharp(z_{ij}>0)} \sum_{\{j:z_{ij}>0\}} z_{ij}$ for base i. We estimate the overall shape structure by the 0-adjusted mean vector, $m^0 = (m_1^0, \cdots, m_d^0)^T$.

3. Compute the level of shape-similarity by $lss_j = \frac{z_j^T m^0}{\|z_j\| \|m^0\|}$, which is equivalent to the cosine of the angle between the vectors Z_j and m^0 where $Z_j = (z_{1j}, \cdots, z_{dj})^T$.

4. Collect lss_j $(j = 1, \cdots, n)$ for the given gene. If the percentage of samples with $lss_j < 0.6$ is greater than 20%, we declare the gene to be an on/off gene.

Results.

An on/off gene example is XIST in Figure 3.6. This gene is associated with gender because it is involved in the X-chromosome inactivation. The left figure shows coverage of all samples at XIST with the 0-adjusted mean vector in black. The identified on/off samples using the proposed algorithm are represented by pink/blue colors. In contrast to the on-samples expressed at high or moderate levels, off-samples are rarely expressed or just noise. The lss values used to separate on/off samples are shown on the left in Figure 3.6 with the symbols $+/\bigtriangleup$ for female/male. The two gender groups are almost perfectly separated except for a few cases and this shows the strong association between gender and the on/off property at this gene.

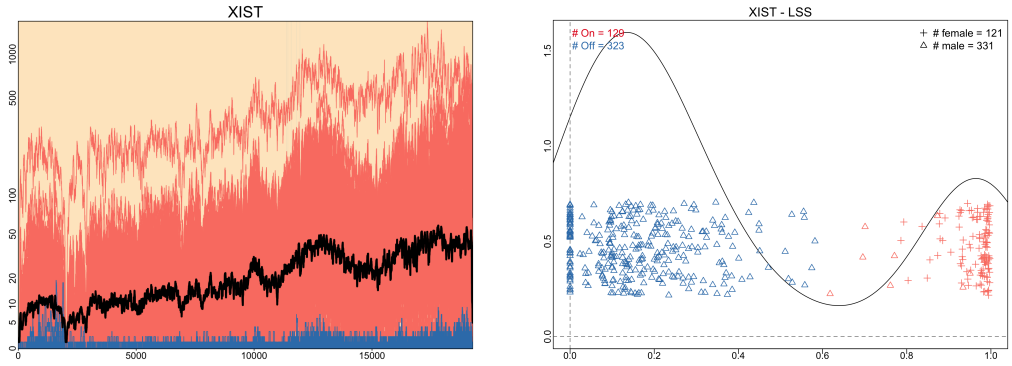


Figure 3.6: On/Off analysis on XIST. The left plot shows the overlays of RNA-seq coverage with red color for on-samples and blue for off-samples classified from the on/off analysis. The right plot shows the estimated lss values with a kernel density estimate (black curve).Based on the cutoff value 0.6, the on- and off-samples were classified as indicated by red and blue colors. The gender of samples are also indicated by the triangle ($\bigtriangleup$) for males and the cross symbol ($+$) for females. The on/off analysis clearly separates the data into two groups.

The on/off analysis at gene TP53 is shown in Figure 3.3 (c). In contrast to the above gene XIST, the lss values of TP53 are all close to 1, indicating that all samples are well expressed and share the overall pattern.

The goal of the on/off analysis is to identify a group of genes with this on/off property. The collection of the *lss* values for the whole genome are displayed in the left heatmap of Figure 3.7. We identified the 5802 on/off genes indicated by the green block of the first heatmap. These are the genes where more than 20% of the samples are off and are excluded from the downstream analysis of Scissor because they do not fit the statistical model for outlier detection. Nevertheless, it is interesting to see if there are any biologically meaningful features in the identified on/off genes.

The right heatmap is the zoom-in figure corresponding to the 5802 genes identified as on/off. The rows and columns are sorted by an unsupervised clustering method and the vertical color code on the left indicates the 5 different clusters for genes. On the right side of the heatmap, each tumor site is indicated by different colors. We grouped those tumor sites into four categories, Oral tongue (oral tongue), Larynx (larynx, hypopharynx), Oropharynx (oropharynx, tonsil, base of tongue), and Other (buccal mucosa, alveolar ridge, floor of mouth, hard palate, lip), and they are presented at the bottom of the heatmap to see if there is any association between tumor sites and the clusters from

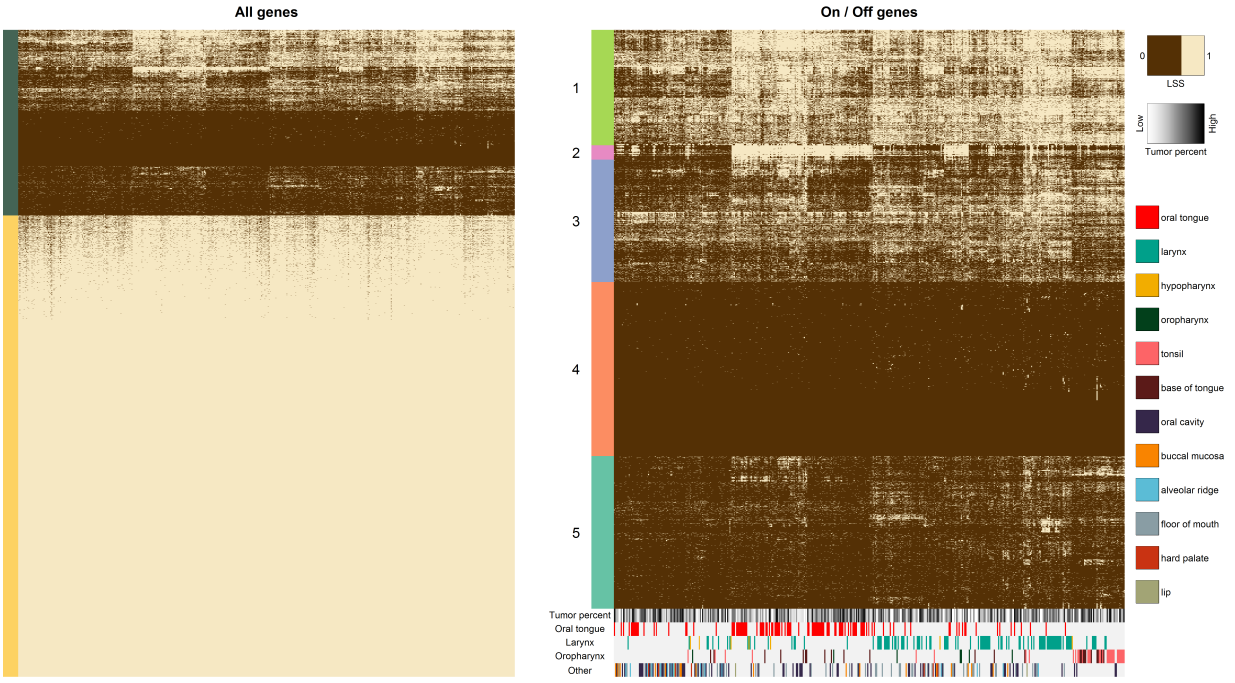


Figure 3.7: Heatmaps of the binary matrices from the level of shape-similarity (LSS). If the LSS is greater than 0.6, the value is 1. Otherwise, the values 0. (a) The heatmap for all genes/cases. Rows are genes and columns are samples. The top block of the heatmap with the dark green color code on the left indicates the identified 5802 on/off genes where more than 20% of samples are noisy or extremely low-expressed, that is, more than 20% of values are 0. The rows and columns are ordered by unsupervised hierarchical clustering. The bottom block with the yellow color code corresponds to the rest of genes that are included in the main statistical procedures for detecting shape variants. The rows are ordered by the number of values of 0. The gene with the most 0 values is displayed at the top of the bottom block. (b) The zoom-in heatmap for the identified 5802 on/off genes from the green color code in Figure 1 (a). The genes were classified into 5 groups as indicated by the color codes on the left. The top color code below the heatmap indicates the tumor percentage and the remaining four color codes indicate the tumor sites: Oral tongue, Larynx (larynx, hypopharynx), Oropharynx (oropharynx, tonsil, base of tongue), Other (buccal mucosa, alveolar ridge, floor of mouth, hard palate, lip). The legend for colors are displayed on the right.

the heatmap. Interestingly, some tumor sites are concentrated together, indicating some association between gene clusters and the tumor sites.

In particular, the zoom-in heatmap for Cluster 2 is shown in the left panel of Figure 3.8. An unsupervised clustering method was reapplied to Cluster 2, based on which the rows (genes) and and columns (samples) of the heatmap are sorted. The sorted heatmap well separates on- and off-samples, indicated by beige and dark brown, and these two sample groups are distinguished by the dendrogram above the heatmap (brown branches for the off-group and green branches for

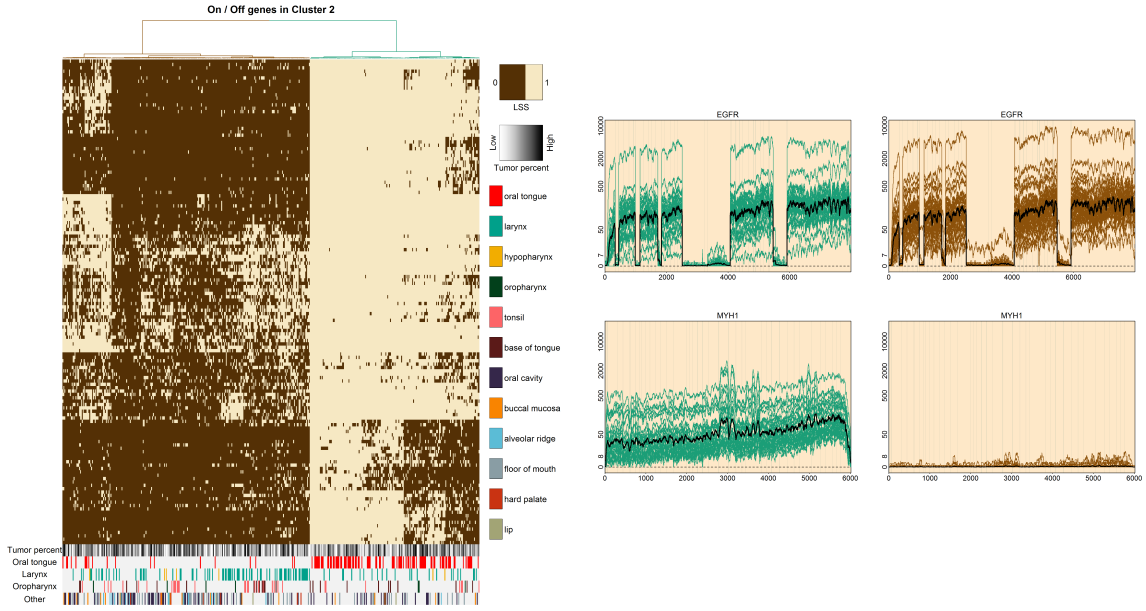


Figure 3.8: The heatmap (left panel) for Cluster 2 is presented using the same color codes as in Figure 3.7. The rows and columns were sorted by an unsupervised clustering algorithm applied to both samples (columns) and genes (rows). Two sample clusters (on/off) were identified as in the dendrogram (green/brown). The sample overlays for these two clusters are compared in two genes, EGFR and MYH1, which are a normal gene and on/off gene in Cluster 2, respectively. The two clusters do not show clear difference in the gene EGFR (top) whereas the gene MYH1 (bottom) shows distinct on/off features of the two clusters.

the on-group). The right panel of Figure 3.8 compares the behavior of these identified two sample groups at different genes, the gene MYH1 from Cluster 2 and a normal gene EGFR. The left column represents the on-sample group and the right column represents the off-sample group using the same colors as in the dendrogram. While there is no clear difference between two groups in the gene EGFR, the on/off gene MYH1 clearly shows that one group (green) is on whereas the other group (brown) is off.

The color code at the bottom of the heatmap demonstrates that tumor samples from oral tongue are mostly on at these on/off genes. To interpret these on/off genes in Cluster 2, we performed Gene Ontology (GO) enrichment analysis to this gene set. The results are summarized in Table 3.2. Here, we only presented the first few GO terms with the smallest p-values. Most of the pathways are associated with muscle, which is roughly because when tumors around tongue were extracted, there is a residual muscle tissue carried by some samples. This shows that these identified on/off genes can reveal presence of the muscle around tongue.

Pathway	% of genes (P-value)	Genes
Muscle contraction	26.2 (6×10^{-36})	CAV3, MYBPC2, MYL1, TRIM72, MYOT, MYOM2, CKMT2, MYOM1, LMOD2, LMOD3, HRC, ACTA1, MYH1, MYH2, MYH4, MYLPF, MYH7, MYH6, CACNG1, TRIM63, CACNA1S, MYH8, TRDN, MYH13, TMOD4, CHRND, SGCA, SCN4A, CHRNG
Muscle filament sliding	13.6 (1×10^{-20})	ACTC1, MYL2, MYBPC2, ACTA1, MYL3, MYBPC1, MYH2, MYL1, MYH4, MYH7, ACTN2, MYH6, MYH8, TNNI1
Cardiac muscle contraction	11.7 (11.7×10^{-15})	ACTC1, MYL2, MYL3, MYL1, MYLK2, MYH7, MYH6, ATP1A2, CSRP3, SCN5A, CASQ2, TNNI1
Sarcomere organization	7.8 (2×10^{-10})	MYPN, LDB3, ACTN2, ANKRD1, MYH6, LMOD2, CASQ1, CASQ2
Regulation of striated muscle contraction	5.8 (7×10^{-10})	MYL2, MYL3, MYH7, MYH6, ATP1A2, CSRP3
Myofibril assembly	5.8 (7×10^{-9})	TMOD4, MYOZ1, MYOZ2, MYH6, LMOD2, MYOZ3, LMOD3
Regulation of the force of heart contraction	5.8 (6×10^{-8})	MYL2, MYL3, MYH7, MYH6, ATP1A2, CSRP3
Striated muscle contraction	4.9 ($7 \times 10-7$)	MYLK2, SMPX, MYH7, MYH6, CASQ2
Regulation of heart rate	5.8 (1×10^{-6})	CAV3, MYH7, MYH6, SCN5A, CASQ2, HRC
Skeletal muscle contraction	4.9 (1×10^{-5})	CHRND, MYH7, CHRNA1, MYH8, TNNI1

Table 3.2: GO enrichment analysis was performed to the gene set of Cluster 2. The table summarizes the several top pathways (the left column) showing the smallest p-values shown in the second column. The genes associated with each pathway are listed in the right column.

3.4.3 Inclusion of intronic part

Although exon-based analysis is commonly applied for RNA-seq expression data (Kimes et al., 2014), it can miss some important abnormal shapes. For example, it can fail to detect intron retention caused by splice-site mutations since the region where the overexpression happened cannot be observed in exon-only expression values. Thus, we include some part of the intron expression values as well as exon expression values as our data object. Since it is common that lengths of transcripts corresponding to introns are much longer than those of exons, inclusion of all bases of introns may yield a data object which is dominated by intronic regions making it hard for some important features of the exons to be distinct. Also, read counts at introns are often very noisy, which could lead to unreliable results since the true signals may be concealed by that noise. For these reasons, we employed a rule determining which parts of introns will be included. Basically, every intron between exons will be fully or partly included. In general, because the coverage around splice sites provide useful information on important shape changes such as intron retention and exon skipping, we include the parts of introns that adjoin exons. In order to make variations of expression levels at exonic regions and intronic regions comparable, we have the total lengths of bases for all exons and all introns at a gene to be as equal as possible. The lengths of intron that will be included between exons are determined by the following rule: for each intron between exons,

1. if its length is less than or equal to a threshold (L), it is fully included.

2. if its length is greater than L, then the part of the intron is taken to be the union of two subsets of length $\frac{L}{2}$ from both ends of that intron and the rest of it is discarded.

Here, the threshold L is chosen such that the difference between total lengths of exons and introns is minimized. As a result, the RNA-seq data that will be considered throughout this work consist of exons and introns with equal weights. Figure 3.3 (d) shows the intron-contained coverage. The chosen intron subsets by the above rule are highlighted by a white background and exons are highlighted by a colored background as before. While the red outlier does not stick out in Figure 3.3 (b), it reveals its abnormality at the 6th intron in Figure 3.3 (d) because the aberration does not

appear at exons. Thus, intron-contained data allow us to discover interesting aberrant events that cannot be observed in exon-only analysis.

3.4.4 Data transformation

The raw coverage data often show heterogeneous variation within and between samples. As an example, Figure 3.9 (a) shows the raw coverage overlays of every sample at the gene CASP1 and we can easily notice unstable dynamics such as strong skewness and different variabilities across bases as well as across samples. For example, highly expressed cases have much more fluctuation than lowly expressed cases and further, coverage variation in exonic regions are much higher than in intronic regions. This implies that such high variation from highly expressed samples or exonic regions may dominate the other interesting variation. A typical remedy to adjust such heterogeneity of counts data is log-transformation (Anders and Huber, 2010), which helps to get more stable variation as shown in Figure 3.9 (b). It is common to add a shift parameter, also known as a pseudo-count, before the log-transformation to avoid the undefined zone of the log function and to control unwanted biases from low versus high. However, there is no consensus on what value of a shift parameter should be used. Here, we propose a procedure to determine a shift parameter that provides more reliable results for the downstream outlier detection.

In the outlier detection procedure of Scissor, it is crucial for the data distribution to have a roughly symmetric distribution to avoid confounding outliers from high skewness. The log-transformation often reduces the extreme skewness of raw counts, but different log shift parameters result in different levels of skewness and thus a careful selection of the parameter helps to obtain a less skewed and roughly symmetric data distribution. Also, different genes present different dynamics in coverage, so it is desirable to apply a suitable parameter for each gene.

Feng et al. (2016) proposed an algorithm to select a shift parameter that minimizes the skewness for one-dimensional data. They used the Anderson-Darling (A-D) statistic to measure the skewness in data and found the parameter minimizing the A-D statistic by a grid search method. However, the pileup data at a given gene has dimension d, the length of the gene, while better interpretation

comes from using only one shift parameter for each gene. It's very unlikely that there is a common value that minimizes the skewness at every base position and so we want a parameter that fits best in some overall sense. Therefore, we propose to select a parameter that minimizes the overall skewness of the data.

How can we measure the overall skewness? As mentioned earlier in Section 3.3, the overall mean expression is shared among samples in a log-linear way and the scale factor a_j in the model describes a relative amount of overall mean expression that is involved in the jth sample. From this point of view, the a_j can be considered as a representative value of the d relative expression values of the sample from all base positions. Thus, we propose to measure the overall skewness of the data based on the A-D statistic of the a_j's ($j = 1, \cdots, n$). Since the a_j's are unknown, we should first estimate them. The estimation procedure is simply based on the linear model with the estimated mean vector as a covariate and the log transformed read depth vectors as response vectors. Theoretically, if the underlying mean vector is zero, then the response values are also expected to be zero, and thus we use the linear model without intercept. Then, the estimate of the a_j can be obtained by the estimate of the coefficient in the linear model and we call the estimated values the *mean scale factors* (msf). See Section 3.4.5 for details. Note that the msfs depend on the log shift parameter. Thus, the resulting A-D statistic using msfs also depend on the parameter, and thus we can choose the parameter by comparing the A-D statistics from different values of the parameter. By a grid search based on a given range for the parameter, we select the parameter that achieves the minimum A-D statistic. The proposed algorithm is summarized as follows:

1. For a range of the parameter, say $1, \cdots, 10$, consider candidates (e.g. integers) of the shift parameter that will be considered in the grid search and compute the A-D statistic for each candidate based on the following steps.

   (a) Take the log-transformation of the data with the given parameter.

   (b) Obtain msf_j ($j = 1, \cdots, n$).

   (c) Compute the A-D statistic of $msf_1, \cdots, msf_n$.

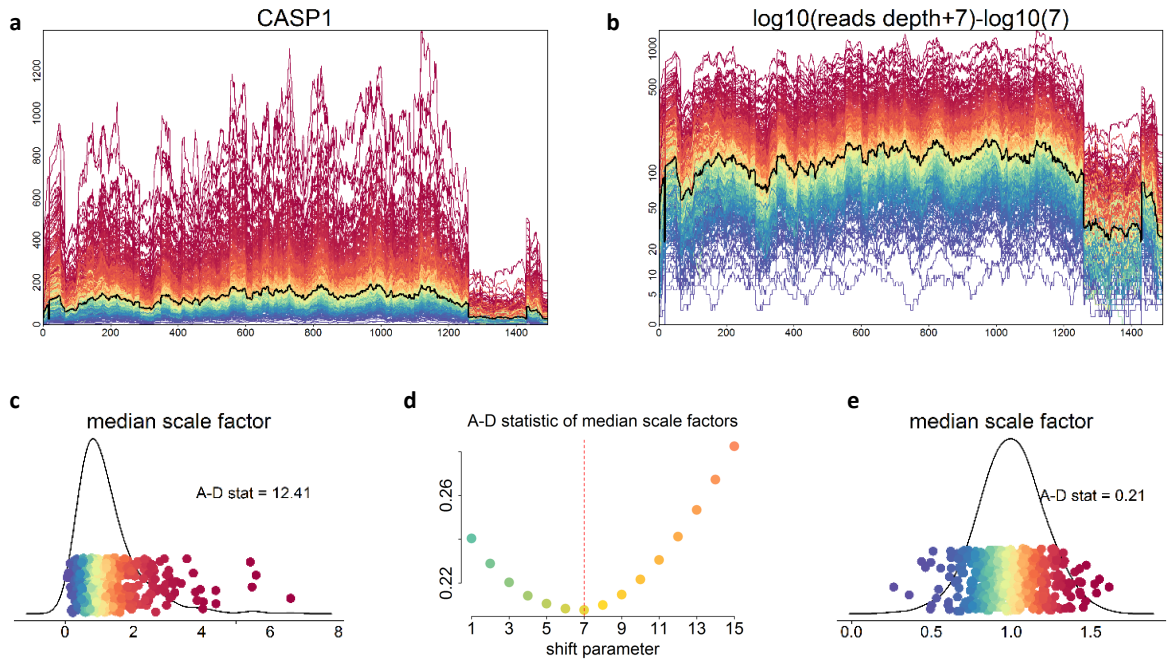


Figure 3.9: Automatic data transformation at gene CASP1. (a) The raw coverage data for 481 patients are shown using rainbow colors, which correspond to the mean scale factors in (c). The red color indicates highly expressed samples and the purple color for lowly expressed samples. The black curve represents the trimmed mean across samples at each base. The data show severe skewness and heterogeneous variation across samples and bases. (b) The log transformed data with the automatically selected shift parameter of 7 are shown with the same rainbow color scale as (a). The black curve represents the trimmed mean across samples at each base. The transformed data are roughly symmetric about the mean curve and show stable variation across samples and bases. (c) The mean scale factors from raw coverage data are shown with the kernel density estimate using the black curve. The distribution is skewed and the A-D statistic of 12.41 indicates strong skewness. (d) The A-D statistics with respect to a shift parameter from 1 to 15 are shown. The value 7 produces the smallest A-D statistic. (e) The mean scale factors with the transformed data using the parameter value 7 are shown with the kernel density estimate. The distribution seems roughly symmetric about the mean 1. The A-D statistic also shows significantly reduced skewness compared to (c).

2. Select the parameter which gives the smallest A-D statistic.

Results.

Figure 3.9 (c) shows the mean scale factors of the raw counts data in (a) with the kernel density curve. The skewed curve as well as the corresponding A-D statistic of 12.41 support that the distribution is substantially skewed. Figure 3.9 (d) shows the A-D statistics with respect to the parameter candidates $(1, \cdots, 15)$. Although any positive values can be used, we considered only

positive integer values for the faster computation and easier interpretation. Also, for the genome-wide analysis, we limited the range of parameters up to 10. This is because it has been observed that the shift parameter values beyond 10 often produce too little variation for low counts compared to high counts, which may result in shape changes at low expression levels being less detectable. Figure 3.9 (d) indicates 7 as the chosen shift parameter, and the log transformed coverage data with the chosen parameter are shown in Figure 3.9 (b). It can be seen that the variation is now nicely stabilized within samples as well as between samples. The corresponding msfs and the A-D statistic shown in Figure 3.9 (e) also attains much less skewness of the distribution than before the log transformation.

3.4.5 Data normalization

Normalization is a critical step of the processing pipeline in the analysis of short read sequencing data. Specifically, normalization aims to reduce the impact of unwanted technical variability on the downstream results by removing this from the data. To detect differential expression, several normalization approaches of RNA-seq data have been proposed to adjust sequencing depth from different libraries, allowing the accurate comparison of expression levels between genes (Anders and Huber, 2010; Robinson and Oshlack, 2010; Bullard et al., 2010; Li and Dewey, 2011; Dillies et al., 2013; Hicks and Irizarry, 2014; Zyprych-Walczak et al., 2015; Reddy, 2015). However, these approaches were designed to standardize gene expression data at the genome-wide level often by scaling the total number of reads across samples. In contrast, to identify base-level shape changes, the proposed normalization aims to remove systematic technical effects that may affect base-level read coverage data at individual gene loci.

Kimes et al. (2014) proposed a normalization procedure for read-depth data to identify clusters of differential isoform usage. They filtered out low-expression samples and performed count normalizing followed by log-transformation. However, biologically important shape changes are still observed in low-expression samples, and simply removing low-expressed samples may lose important genetic events. Also, scaling each remaining sample by its total expression can distort its

genuine shape architecture by yielding unwanted fluctuation especially in lowly-expressed regions. Such artificial variation may confound other important variation.

Here, we propose an effective way to standardize base-level read coverage data between samples at a given gene, while preserving important shape deviation. As discussed above, the model (3.3.1) accounts for the systematic variation from technical biases via the normalization factors a_j and a smooth curve $g(a_j)$. The a_j's describe the different expression levels and $g(a_j)$ accounts for the overall variation subject to the expression levels. While the log-transformation step helps to stabilize the different levels of variation and significantly reduce extreme skewness in the data, it is observed in a substantial number of genes that log-transformation does not fully remove the systematic variation. Such systematic variation often leads to biased results and therefore should be adjusted for. The normalization step of Scissor aims to reduce this technical bias and obtain the fundamental variations x_{ij} that are essentially associated with outlying signals. In practice, we only observe R_{ij} and the other unknown elements μ_i, a_j and $g(.)$ can be estimated as described below.

First, we estimate μ_i by a trimmed mean at each base-position. This simple method effectively down-weights the outliers and helps to robustly estimate the true mean expression levels. As introduced in the previous section, the majority of samples share the overall shape patterns with different expression levels and thus, the a_j can be estimated by using the linear model. And then, we equate the overall expression levels of samples to zero by subtracting out the sample-specific overall expression as follows:

$$r_j = \log R_j - \hat{a}_j \log \hat{\mu}.$$

where $\hat{a}_j$ and $\hat{\mu}$ are the estimates of a_j and μ.

Now, we propose a procedure to estimate the unknown function $g(a)$ that describes overall variation subject to the expression levels a_j. Suppose that a_j and μ are known and they are normalized out. Then, our model becomes

$$g(a_j)X_j = \log R_j - a_j \log \mu.$$

Letting $v_j = \log R_j - a_j \log \mu$, we have the following relationship:

$$\sum_i v_{ij}^2 = g^2(a_j) \sum_i x_{ij}^2.$$

By taking the expectation of both sides, we have

$$\begin{aligned} E[\sum_i v_{ij}^2] &= g^2(a_j) E[\sum_i x_{ij}^2] \\ &= g^2(a_j)\tau \end{aligned}$$

where $\tau = E[\sum_i x_{ij}^2]$. This implies that if $g(.)$ is independent of the scale factor, i.e. $g(.)$ is a constant function, then $\sum_i v_{ij}^2$ should be centered at τ with some variation with no systematic pattern. Conversely, if $\sum_i v_{ij}^2$ shows a certain systematic pattern with respect to a_j, then it can be an evidence of variation subject to overall expression. The characterization of this pattern thus can be used for the estimation of $g(.)$. However, there is an unknown factor τ, indicating that $g(.)$ is unidentifiable solely based on $\sum_i v_{ij}^2$. Let $f(a) = g^2(a)\tau$. By giving a constraint $g(1) = 1$ as stated before, we have $f(1) = g^2(1)\tau = \tau$, allowing $g(a)$ to be identifiable by $\sqrt{f(a)/f(1)}$. The remaining part is then the estimation of $f(a)$.

Using the estimates from above, we can replace the v_j by $r_j = \log R_j - \hat{a}_j \log \hat{\mu}$ and thus $r_j^T r_j = \sum_i r_{ij}^2$ can be used to estimate $f(a)$. Simply taking the sum of squared values from data may not be a good estimate because it can easily break down due to some extreme values. This limitation can be overcome by emplying other robust approaches. Here, we use Tukey's bisquare method to downweight get extreme values, allowing a more stable estimation, but other methods can be also used. Based on Tukey's bisquare function ρ, i.e.

$$\rho(r) = \begin{cases} \frac{k^2}{6}\left\{1 - [1 - (\frac{r}{k})^2]^3\right\} & \text{for } |r| \leq k \\ \frac{k^2}{6} & \text{for } |r| > k, \end{cases}$$

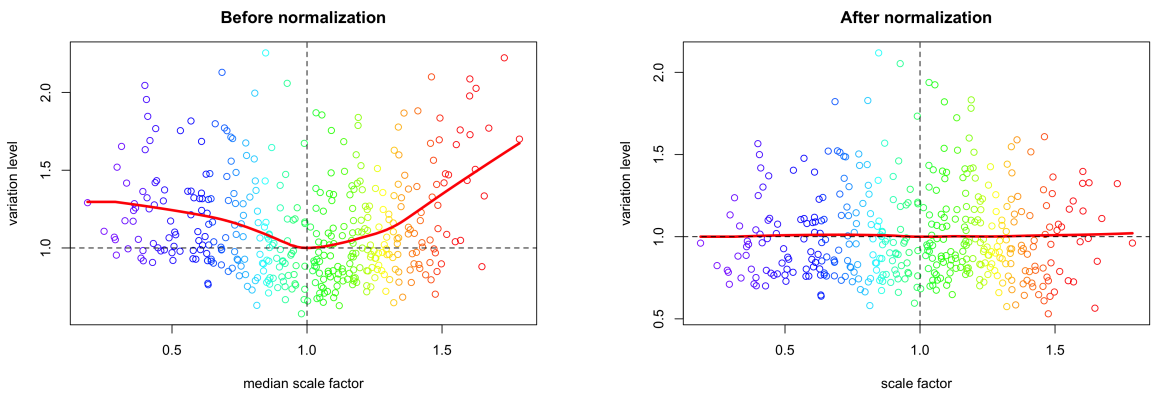


Figure 3.10: Variation levels of the data before and after normalization for gene TP53. In both plots, each point represents each sample and is colored based on the magnitude of the estimated scale factors $\hat{a}_j$ using a rainbow color scale. The red points indicate highly expressed samples and purple points indicate lowly expressed samples. The red curves in each plot are the estimates using a popular smoothing technique LOWESS with a smoothing parameter 0.7. The left plot presents the variation levels computed from the log-transformed data shown in Figure 3.3 (e). The right plot shows the variation levels computed from the normalized data shown in Figure 3.3 (f). After the proposed normalization, the systematic pattern involved in the overall variation with respect to expression levels are effectively removed.

we estimate the $E[\sum_i v_{ij}^2]$ by

$$\sum_i \rho(r_{ij})$$

with the tuning parameter $k = 4.685\sigma$ where σ is the sample standard deviation of r_{ij}.

The left plot of Figure 3.10 shows the properly scaled $\sum_i \rho(r_{ij})$ with respect to $\hat{a}_j$ at gene TP53 such that the center of the points is $(1,1)$. The points exhibit a systematic pattern, which supports the existence of non-negligible impact of overall expression onto the variation r_{ij}. The function f can be estimated by pooling information across points and fitting a smooth curve to capture the global trend. Then, the function g is estimated by the fitted curve $\hat{f}$ scaled by $\hat{f}(1)$ so that $\hat{g}(1) = 1$. The estimated function $\hat{g}$ for the gene TP53 is shown by the red curve. We divide r_{ij} by $\hat{g}(\hat{a}_j)$, i.e. $\frac{r_{ij}}{\hat{g}(\hat{a}_j)}$, to remove the overall expression-dependence. We repeat this procedure until we get the data object whose pattern is not systematic any longer. At gene TP53, the normalized data are presented in Figure 3.3 (f) and the adjusted variation levels after the normalization are shown in the right plot

of Figure 3.10. We now denote the normalized data by x_{ij} and this will be used for the downstream outlier detection analysis.

3.5 Detecting shape changes

We are now in a position to present the method. The shape outliers of interest in our study are the samples whose RNA-seq coverage structure are abnormal from the majority of the data. The shape changes mostly appear as atypical enrichment or scarcity of expression in some area of a given gene. The pre-processing procedure introduced in the previous section helps outliers be more distinctive in that the abnormal gain or loss of expression stick out as in Figure 3.3 (f). Such abnormality is reflected in the proposed mixture model in Section 3.3 using a set of unknown outlier directions. If we can estimate these outlier directions, they would be very useful for identifying outliers as well as interpreting the outliers. In this section, we propose a new approach to approximating underlying outlier directions with the goal of selecting important RNA-seq shape variants.

Extraction of latent outlier directions from a high dimensional data set is not simple because many features are involved in an overwhelming number of dimensions. So it is advantageous to make use of the knowledge about what types of aberrant features should be considered. In this work, we propose a two step procedure each step of which is designed to reveal particular types of aberration. The first step aims to find some abnormal features that involve strong underlying signals so that the features can be captured by the first few PC directions. While this type of abnormal feature does not need to be involved in some specific RNA-seq structure, the outliers in this category typically present global or landscape shape changes that appear in a wide range within a gene. This is roughly because for such wide shape changes to be distinguished, the levels of the underlying signals should be compelling. Based on this observation, we call this type of aberration *global shape changes*. For example, altered large exons or large introns, intergenic deletions, and fusions often manifest these global shape changes in various forms.

On the other hand, the second step aims to find some particular structure of RNA-seq shape changes. The shape changes missing from the first step are typically characterized by abnormal gain

or loss at some restricted regions. These local shape changes are associated with finer variation such as alternative splicing at short exons or short introns and small deletions in the middle of an exon. (Examples) We call this type of aberration *local shape changes*. These often have too weak signals to be consistently estimated in the first few PCs. To identify these more challenging outliers, we limit our view to some important regions within a given gene. This helps to find the directions we should look at to discover meaningful outliers.

3.5.1 Global shape change detection

The global shape change detection aims to identify the shape changes that are involved in some strong outlier directions so that the first few principal components may capture the underlying abnormal structure. This enables us to narrow down to the first few principal components instead of dealing with full large dimensions. With the normalized data vectors, X_j, we denote the data matrix by $\mathbf{X} = [X_1, X_2, \cdots, X_n]$ and the sample covariance matrix by $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$. We do not subtract the mean vector $\bar{X}$ because the X_j's are already centered during the normalization procedure. Let $(\hat{U}_i, \hat{\lambda}_i)$ from PCA denote the pairs of sample eigenvectors (PC directions) and sample eigenvalues of $\mathbf{S}_n$ sorted by the eigenvalues in decreasing order.

3.5.1.1 The most outlying direction

Denote the PC score, i.e. the projection of the jth point onto the ith sample eigenvector, by $\hat{y}_{ij}$. Suppose a hypothetical scenario that an outlier signal $U_{i'}$ is consistently estimated by one of the sample eigenvectors, say $\hat{U}_{i'}$. Then, the outliers that involve $U_{i'}$ are expected to stick out in the direction $\hat{U}_{i'}$ in that the $\hat{y}_{i'j}$ corresponding to those outliers are much larger than the others. That is, the normalized PC scores, defined as

$$\hat{v}_{ij} = \frac{\hat{y}_{ij}}{\sqrt{\hat{\lambda}_i}}, \tag{3.5.1}$$

can be used to detect outliers by choosing cases that have substantially large $\hat{v}_{i'j}$. As we discussed in Section 2.5, however, the PCA subspace consistency does not guarantee that outlier signals are estimated individually. Therefore, the PC scores corresponding to the single K sample eigenvectors might be unsuitable for quantifying the outlyingness and less powerful for detecting outliers.

Then, what is a more informative direction for outliers than the individual PC directions? The type of structure of interest here is *outlying* and there is no guarantee that the large-variance PC directions from PCA will find such features. To find such outlying structure, we employ the *projection depth function* in (3.2.1) over the low-dimensional space, with the goal of detecting outliers based on the outlying structure. Using the same notation as in (3.2.1), denote a normalized projection score of a point X with respect to $\mathbf{X}$ on a direction h by $s_h(X|\mathbf{X})$, i.e.

$$s_h(X|\mathbf{X}) = \frac{h^T X - \mu(F_{h^T\mathbf{X}})}{\sigma(F_{h^T\mathbf{X}})} \quad \text{such that } \|h\| = 1.$$

The $s_h(X|\mathbf{X})$ reflects the outlyingness of X over the direction h with respect to the dataset $\mathbf{X}$. Scaling a projection score $h^T X$ by the spread of $\{h^T X_j\}_{1\le j\le n}$ allows us to compare the outlyingness of X over the direction h with the other direction vectors, $h' \in \text{span}(X_1, \cdots, X_n)$. Therefore, we define the most outlying direction h^* of a data point X as a direction that yields the largest $s_h(X|\mathbf{X})$ over all directions h as follows:

Definition 3.5.1. The MOD (Most Outlying Direction) of a point $X \in \mathbb{R}^d$, denoted by $h^*(X|\mathbf{X})$, is defined as a unit vector $h \in \text{span}(X_1, \cdots, X_n)$ that maximizes $s_h(X|\mathbf{X})$, i.e.

$$h^*(X|\mathbf{X}) = \text{argsup}_h |s_h(X|\mathbf{X})| \quad \text{subject to } h \in \text{span}(X_1, \cdots, X_n) \text{ and } \|h\| = 1.$$

From the definition of a projection depth function, an MOD is the direction that achieves the supremum in a projection depth function. That is, the direction $h^*(X|\mathbf{X})$ provides the most projection outlyingness of the point X over all possible directions. In contrast to PC directions that describe overall structures relevant to the majority of data points, the most outlying direction describes individual structure of each data point that makes the point the most distinguished. Accordingly, the

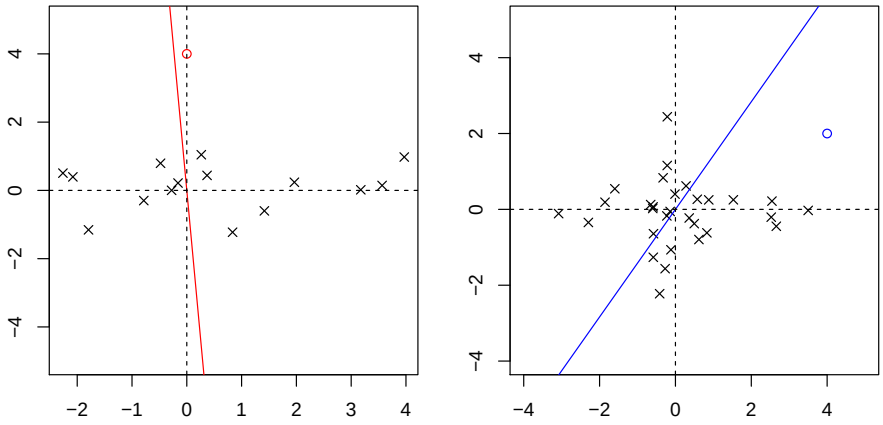


Figure 3.11: In the left (right) figure, the outlier is indicated by a red (blue) circle and the other normal points are indicated by black. The red (blue) line illustrates the MOD of the red (blue) outlier.

MODs of outliers may be used to recover hidden outlier directions, and further the visualization of the MOD often provides a useful interpretation as to why the RNA-seq outliers are abnormal as will be seen in Section 3.6. Such interpretation is informative in that users can understand the biological meaning of the abnormalities.

Figure 3.11 illustrates the MODs of two outliers with (μ, σ)=(Median, MAD). In the left figure, an outlier is indicated by a red circle with normal observations indicated by black. The MOD of the red circle is indicated by a red line. It is obviously the direction that describes the unique structure of the red outlier in the sense that the other points have little variation in the red direction. Here, the MOD is very close to the red point since the red point only goes along the vertical axis and the others rarely do. In the right panel, an outlier is indicated by a blue circle and its MOD is illustrated by a blue line. This MOD also provides good insight about the abnormality of the blue outlier compared to the other points. In this case, the MOD is not close to the blue outlier because the MOD here is balancing the projected value of the blue point with the spread of projected values of the other points. Thus, the MOD of an outlier point can be thought of as a compromise between the outlier and a direction of minimal projected variation.

The projection depth function can be rewritten based on a MOD as follows:

$$o(X|\mathbf{X}) = \sup_{\|h\|=1} \left| \frac{h^T X - \mu(F_{h^T\mathbf{X}})}{\sigma(F_{h^T\mathbf{X}})} \right| \tag{3.5.2}$$

$$= \sup_{\|h\|=1} |s_h(X|\mathbf{X})| = |s_{h^*}(X|\mathbf{X})| \tag{3.5.3}$$

where $h^* = h^*(X|\mathbf{X})$. Then, the $o(X|\mathbf{X})$ is understood as the maximum value over all possible one-dimensional outlyingness and we call this value the *projection outlyingness* (PO) or *projection outlyingness score*. As a special case, if $h^*(X_j|\mathbf{X}) = \hat{U}_{i'}$ for a data point X_j with (μ,σ)=(Mean, SD), then $o(X_j|\mathbf{X})$ becomes the absolute value of the normalized PC score $|\hat{v}_{i'j}|$ in (3.5.1). Hence, a PO score can be seen as a generalized version of the normalized PC score, allowing projection onto other directions (MODs) and other location and scale measures rather than the PC directions and (Mean, SD). In contrast to PCA that obtains $\min(d,n)$ orthogonal PC directions by sequentially maximizing the sample variances, Scissor obtains n MODs each of which maximizes the projection outlyingness of each data point. Note that the n MODs do not need to be orthogonal to each other and besides, if several outliers are involved in similar outlying structure, then the corresponding MODs tend to be close to each other, constructing fairly small angles.

The n MODs yield the n PO scores, allowing appropriate comparison of the depth of outlyingness across data points. For example, if a point has a higher PO score, then it is more likely to be an outlier. Based on a cutoff value, Scissor declares a data point as an outlier when its PO score is so large that it cannot be justified by randomness as will be seen later.

Although any univariate location and scale measures can be used for $\mu(\cdot)$ and $\sigma(\cdot)$, the Med and MAD have been widely employed due to their robustness and simplicity. Scissor also uses (μ,σ)=(Med, MAD) by default. However, for the purpose of exploring theoretical properties of the MODs and PO scores, the projection depth function involved with (Med, MAD) is very tricky to deal with because it not differentiable with respect to h. On the other hand, the projection depth function based on (Mean, SD) may not be a reliable measure of outlyingness because the mean and SD easily break down at extreme points. Nevertheless, the function with (Mean, SD) is much easier

to deal with, making the theoretical investigation much simpler. Therefore, we first investigate some interesting properties of the MODs and PO scores based on (Mean, SD) and then propose the main procedure based on (Med, MAD).

3.5.1.2 Projection depth function with (μ, σ)=(Mean, SD)

In this subsection, we consider a projection depth function based on (μ, σ)=(Mean, SD), which provides a good insight into understanding how the method works. We first describe the procedure of obtaining the MOD and PO score with a data point X_l. We will see that they are reduced to a simple and intuitive formula based on the PC directions and PC scores. In this subsection, we assume that the data vectors $X_1, \cdots, X_n$ come from a d-variate normal distribution with mean zero and some covariance matrix.

Let $\hat{\mathbf{U}}_q = [\hat{U}_1, \cdots, \hat{U}_q]$ be the first q sample eigenvectors of $\hat{\mathbf{\Sigma}}$. Then a direction vector h in S_1^q can be represented by a linear combination of the eigenvectors, $\hat{U}_1, \cdots, \hat{U}_q$, i.e.

$$h = \hat{\mathbf{U}}_q \alpha = \alpha_1 \hat{U}_1 + \cdots + \alpha_q \hat{U}_q \tag{3.5.4}$$

where α is a q-dimensional vector with $\alpha = (\alpha_1, \cdots, \alpha_q)^T$ and $\|\alpha\| = 1$. Then, the optimization problem in (3.5.2) with (μ, σ)=(Mean, SD) can be equivalently written as

$$o^2(X_l|\mathbf{X}) = \sup_{\|\alpha\|=1} \left| \frac{\alpha^T \hat{\mathbf{U}}_q^T X_l - \text{Mean}(\alpha^T \hat{\mathbf{U}}_q^T \mathbf{X})}{SD(\alpha^T \hat{\mathbf{U}}_q^T \mathbf{X})} \right|^2 \tag{3.5.5}$$

$$= \sup_{\|\alpha\|=1} \left| \frac{\alpha^T \hat{\mathbf{y}}_{q,l} - \text{Mean}(\alpha^T \hat{\mathbf{Y}}_q)}{SD(\alpha^T \hat{\mathbf{Y}}_q)} \right|^2 \tag{3.5.6}$$

where $\hat{\mathbf{y}}_{q,j} = \hat{\mathbf{U}}_q^T X_j$ for $1 \le j \le n$ are vectors of PC scores of the j-th data point X_j and $\hat{\mathbf{Y}}_q = [\hat{\mathbf{y}}_{q,1}, \cdots, \hat{\mathbf{y}}_{q,n}]$. As the mean vector and covariance matrix of $\hat{\mathbf{Y}}_q$ can be approximated by $\mathbf{0}_q$ and $\hat{\mathbf{\Lambda}}_q = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_q)$, (3.5.5) can be written as

$$o^2(X_l|\mathbf{X}) = \sup_{\|\alpha\|=1} \frac{\alpha^T \hat{\mathbf{y}}_{q,l} \hat{\mathbf{y}}_{q,l}^T \alpha}{\alpha^T \hat{\mathbf{\Lambda}}_q \alpha}. \tag{3.5.7}$$

This can be solved by a generalized eigenvalue problem

$$\mathbf{A}\alpha = \lambda \mathbf{B}\alpha$$

where $\mathbf{A} = \hat{\mathbf{y}}_{q,l}\hat{\mathbf{y}}_{q,l}^T$ and $\mathbf{B} = \hat{\mathbf{\Lambda}}_q$ with the solution λ^* being a projection outlyingness $o^2(X_l|\mathbf{X})$. Then, we get the solutions as follows:

$$\alpha^* = \frac{\hat{\mathbf{\Lambda}}_q^{-1}\hat{\mathbf{y}}_{q,l}}{\|\hat{\mathbf{\Lambda}}_q^{-1}\hat{\mathbf{y}}_{q,l}\|} \tag{3.5.8}$$

$$\lambda^* = \hat{\mathbf{y}}_{q,l}^T\hat{\mathbf{\Lambda}}_q^{-1}\hat{\mathbf{y}}_{q,l}. \tag{3.5.9}$$

It follows by (3.5.4) and (3.5.8) that the MOD of the X_l and the square of the corresponding projection outlyingness can be obtained by

$$\begin{aligned} h^*(X_l|\mathbf{X}) &= \hat{\mathbf{U}}_q\alpha^* = \hat{\mathbf{U}}_q\frac{\hat{\mathbf{\Lambda}}_q^{-1}\hat{\mathbf{y}}_{q,l}}{\|\hat{\mathbf{\Lambda}}_q^{-1}\hat{\mathbf{y}}_{q,l}\|} \\ o^2(X_l|\mathbf{X}) &= \hat{\mathbf{y}}_{q,l}^T\hat{\mathbf{\Lambda}}_q^{-1}\hat{\mathbf{y}}_{q,l} = \sum_{i=1}^{q}\left(\frac{\hat{y}_{il}}{\sqrt{\hat{\lambda}_i}}\right)^2. \end{aligned} \tag{3.5.10}$$

Therefore, the MOD and PO scores using Mean and SD have a simple analytical solution and a close relationship with PCA. The MOD is a linear combination of the PC directions with the coefficients of the properly scaled PC scores and the PO score is a sum of the squares of normalized PC scores corresponding to X_l. Note that here we consider the first q PC directions. It is easy to notice that as q increases, the projection outlyingness becomes large and dominated by the accumulated noise.

More explicitely, under a normality assumption, a PC score $\hat{y}_{il}$ follows a normal distribution with mean zero and variance $\hat{\lambda}_i$ and therefore $o^2(X_l|\mathbf{X})$ is a sum of the squares of q independent standard normal random variables. Then, it follows that $o^2(X_l|\mathbf{X})$ is distributed as the chi-squared distribution with q degrees of freedom with the mean q and variance $2q$, indicating that as more PC directions are involved, the distribution of $o^2(X_l|\mathbf{X})$ tends to be a normal distribution with increased

mean and variance. This suggests that every data point in high dimensions has their own directions where their projection outlyingness is pretty unusual just by the overwhelming noise, which implies the possibility of swamping (a nonoutlier is detected as an outlier). Therefore, searching for outliers in high dimensions based on projection outlyingness may easily fail because real outliers are less distinguishable. This property is closely connected to the geometrical representation of high-dimensional data investigated in Hall et al. (2005) and Section 2.4 of this thesis in that every point is approximately orthogonal to each other and thus each data point can be an outlying direction of itself.

It should be noted that $o^2(X_l|\mathbf{X})$ is equivalent to Hawkins' statistic H_{1l}^2 mentioned in Section 3.2 except that Hawkins' statistic is on the basis of the last few PCs. This observation provides a new interpretation of Hawkins' statistic in the sense that it is the projection outlyingness using the mean and SD over the subspace using the last few PC directions. Therefore, the outliers that are detectable using Hawkins' statistic are those which inflate the (Mean, SD)-based projection depth among the last few PCs. As mentioned in Jolliffe (2002), for the low-dimensional case, the last few PCs can reveal observations that violate the correlation structure imposed by the bulk of the data. By contrast, examining the last few PCs is not really helpful for detecting outliers in high dimensions because they mostly correspond to noise.

Thus, in order to get reliable projection outlyingness against swamping in a high dimensional space, it is important to carefully choose the subset of the PC directions that will be included. There are many existing methods to estimate the number of spike signals and here, we assume that the right number of principal components is selected, which is K. We will discuss how to choose K in Chapter 3.6.2. Then, the signal-related low rank space can be identified by the subspace spanned by the first K principal components, $\hat{U}_1, \cdots, \hat{U}_K$, for $K < \min(n,d)$. Letting S_j^k denote the subspace spanned by $\hat{U}_j, \cdots, \hat{U}_k$, i.e., $S_j^k = \text{span}(\hat{U}_j, \cdots, \hat{U}_k)$, our low dimensional representation, denoted by $\tilde{\mathbf{X}}$, can be obtained by projecting $\mathbf{X}$ onto S_1^K. Then $\tilde{\mathbf{X}}$ provides sufficient information that is involved in strong shape changes.

As mentioned earlier, the global shape change detection procedure aims to identify the strong outliers whose underlying abnormal structure are captured by the first few PCs, which are the K PCs here. Following the same process as in (3.5.4) - (3.5.10) based on the $\hat{\mathbf{U}}_K = [\hat{U}_1, \cdots, \hat{U}_K]$, we get the square of a (Mean, SD)-projection outlyingness of X_l,

$$o^2(\tilde{X}_l|\tilde{\mathbf{X}}) \quad = \quad \sum_{i=1}^{K} \Big(\frac{\hat{y}_{il}}{\sqrt{\hat{\lambda}_i}} \Big)^2. \tag{3.5.11}$$

Then, the statistic $o^2(\tilde{X}_l|\tilde{\mathbf{X}})$ follows the chi-squared distribution with K degrees of freedom under a normality assumption. Even beyond the normality assumption, the distribution $o^2(\tilde{X}_l|\tilde{\mathbf{X}})$ can still be approximated by the chi-squared distribution since the PC score $\hat{y}_{il}$, which is the mean of d random variables, roughly follows a normal distribution by the CLT.

3.5.1.3 Projection depth function with (μ, σ)=(Med, MAD)

As mentioned before, the projection depth function based on (Mean, SD) is sensitive to outliers and thus can *mask* real outliers. Masking occurs when estimates are so highly affected by outlying points that some good outliers cannot be detected. The projection depth function as an outlier identifier must be robust against masking. A well-known robust projection depth function is based on (μ, σ)=(Med, MAD) where $\text{MAD}(\mathbf{x}) = \text{Med}_j(|x_j - \text{Med}_i(x_i)|)/\Phi^{-1}(0.75)$ with the standard normal cdf Φ. Then, the normalized projection score becomes $s_h(X|\mathbf{X}) = \frac{h^T X - \text{Med}(h^T\mathbf{X})}{\text{MAD}(h^T\mathbf{X})}$ and the chosen projection depth function is

$$o(X|\mathbf{X}) = \sup_{\|h\|=1} |s_h(X|\mathbf{X})| = \sup_{\|h\|=1} \Big| \frac{h^T X - \text{Med}(h^T\mathbf{X})}{\text{MAD}(h^T\mathbf{X})} \Big|. \tag{3.5.12}$$

There are two major challenges to directly dealing with this projection depth function for our full dimensional data. One is that since the optimization is computationally so intensive, most solvers to this problem require that the sample size is larger than the dimension. Second, even if a solver can provide a solution for full dimensional data with high dimensionality, the solution

might not be meaningful because of the swamping effect from high dimensions as discussed in the projection depth function with Mean and SD. For these reasons, we first approximate the data matrix using the first few PCs and apply the projection depth function. Again, this is possible because we aim to identify outliers associated with strong shape changes in this global shape change detection step.

Similar to the procedure as in (3.5.4) - (3.5.5), the projection depth function based on the low rank approximations $\tilde{\mathbf{X}}$ can be reduced to a $K(\ll n)$ dimensional problem with sample size n in the rotated new coordinate system as follows

$$o(\tilde{X}|\tilde{\mathbf{X}}) \quad = \quad \sup_{\|\alpha\|=1} \left| \frac{\alpha^T \hat{\mathbf{U}}_K^T \tilde{X} - \text{Med}(\alpha^T \hat{\mathbf{U}}_K^T \tilde{\mathbf{X}})}{\text{MAD}(\alpha^T \hat{\mathbf{U}}_K^T \tilde{\mathbf{X}})} \right|$$

where $\alpha = (\alpha_1, \cdots, \alpha_K)^T \in \mathbb{R}^K$ with $\|\alpha\| = 1$. Again, $\hat{\mathbf{U}}_K^T \tilde{\mathbf{X}}$ is a PC scores matrix corresponding to $\hat{U}_1, \cdots, \hat{U}_K$ and it follows that

$$o(\tilde{X}|\tilde{\mathbf{X}}) = \sup_{\|\alpha\|=1} |s_\alpha(\hat{Y}|\hat{\mathbf{Y}}_K)| = \sup_{\|\alpha\|=1} \left| \frac{\alpha^T \hat{Y} - \text{Med}(\alpha^T \hat{\mathbf{Y}}_K)}{\text{MAD}(\alpha^T \hat{\mathbf{Y}}_K)} \right|$$

where $\hat{Y} = \hat{\mathbf{U}}_K^T \tilde{X}$ and $\hat{\mathbf{Y}}_K = \hat{\mathbf{U}}_K^T \tilde{\mathbf{X}}$. Now the optimization involves n data points with a smaller dimension $K \ll n$, which is usually more tractable.

Since the optimization may have many local maxima even if the global maximum is much larger than these (Stahel, 1981), ordinary nonlinear programming cannot solve the problem. Instead, an approximate algorithm proposed by Stahel (1981) has been widely used to compute a practicable solution. The algorithm compares the values $\{s_v(\hat{Y}|\hat{\mathbf{Y}}_K) : v \in \mathbf{v}\}$ over a finite set of well constructed K-dimensional directions $\mathbf{v} = \{v_1, \cdots, v_s\}$ with the direction giving the largest value being a solution, α^*. A popular procedure to generate a finite set of directions $\mathbf{v} = \{v_1, \cdots, v_s\}$ is to choose at random K indices $\{i_k\}_{1 \le k \le K}$ from $\{1, \cdots, n\}$ and find a direction v perpendicular to the hyperplane passing through $\hat{Y}_{i_1}, \cdots, \hat{Y}_{i_K}$. Repeat this procedure s times and get s directions $v_1, \cdots, v_s$.

From the solution α^* by this optimization algorithm, the MOD and the PO score are obtained as

$$\begin{aligned} h^*(\tilde{X}|\tilde{\mathbf{X}}) &= \hat{\mathbf{U}}_K\alpha^* \\ o(\tilde{X}|\tilde{\mathbf{X}}) &= s_{\alpha^*}(\hat{Y}|\hat{\mathbf{Y}}_K). \end{aligned} \tag{3.5.13}$$

3.5.1.4 Global shape change detection algorithm

We now present the proposed method for identifying strong shape changes. To effectively extract the latent outlier structure based on the projected outlyingness, an important condition is the normality of the projected points corresponding to nonoutliers. This condition is often satisfied because PC scores approximately follow the normal distribution by the CLT. However, this is not the case for some genes. Even when we filtered out the on/off genes in the pre-processing step, there still remained genes with a group of off- or rarely expressed samples, which heavily skewed the distribution of the projected points. The serious skewness often causes nonoutliers to be classified as outliers. Also, some genes have multiple transcript variants and isoforms, resulting in a bimodal distribution or a strongly skewed distribution.

Such departure from normality can produce many false discoveries and conceal biologically important outliers, and therefore some attention is needed. Here, we propose a modified projection depth approach taking into account departure from normality. To do this, we add a normality constraint to the projection depth function so that we only search the directions in which the given constraint is satisfied. Let ϕ be a function measuring the normality. Then, the projection outlyingness will be

$$o_1(\tilde{X}|\tilde{\mathbf{X}}) = \sup \left| \frac{\alpha^T\hat{Y} - \mathrm{Med}(\alpha^T\hat{\mathbf{Y}}_K)}{\mathrm{MAD}(\alpha^T\hat{\mathbf{Y}}_K)} \right| \quad \text{s.t. } \|\alpha\| = 1,\ \phi(\alpha^T\hat{\mathbf{Y}}_K) \le \rho, \tag{3.5.14}$$

where ρ is a cutoff value. While any measure of normality can be used, here we employ the winsorized A-D statistic (Feng et al., 2016) with a special emphasis on the skewness. This statistic

is robust against outliers, allowing us to keep the useful directions which indicates good outliers. From (3.5.14), we can obtain the MOD by choosing the direction giving the largest value among $\mathbf{v}\backslash\mathbf{v}'$ where $\mathbf{v}'$ is a set of directions with $\phi(v_k^T\hat{\mathbf{Y}}_K) > \rho$ for $v_k \in \mathbf{v}$. Accordingly, the obtained MOD is a more reliable direction for detecting outliers.

Based on the modified projection depth function, the global shape change algorithm is proposed as follows:

1. Apply PCA and find a low (K-) dimensional subspace and get a low dimensional approximation by projecting X_j onto it:

$$\tilde{X}_j = \hat{y}_{1j}\hat{U}_1 + \cdots + \hat{y}_{Kj}\hat{U}_K, \quad j = 1, \cdots, n.$$

2. For a data point $\tilde{X}_j$ ($j = 1, \cdots, n$), obtain the PO score $o_1(\tilde{X}_j|\tilde{\mathbf{X}})$ based on (3.5.14). Here, $\tilde{\mathbf{X}}$ is a data matrix whose columns are low-dimensional data vectors and does not need to contain $\tilde{X}_j$.

3. Collecting the PO scores for $j = 1, \cdots, n$, declare a set of data points as outliers if $o_1^2(\tilde{X}_j|\tilde{\mathbf{X}}) \geq \chi^2_{1-\alpha}(K)$ with a pre-determined level α.

3.5.1.5 Analysis with toy example

We revisit the toy example in Section 2.6 to illustrate how the proposed procedure can be used to approximate underlying outlier directions. In this toy data set, as a reminder, there are 4 outliers, denoted by $X_1, \cdots, X_4$, among $n = 200$ independent data vectors in $d = 2000$. The e_{10} was used as an underlying outlier direction to generate these outliers. This example showed the PCA subspace consistency, highlighting the situation where the outlier direction is approximately captured by the first few PC directions but none of the PC directions are individually representative of the outlier direction.

Table 3.3 shows the squares of the first 12 entries (in rows) of the MODs of the six data points, $X_1, \cdots, X_6$, denoted by $h_1^*, \cdots, h_6^*$ (in columns). The MODs were obtained by the algorithm proposed

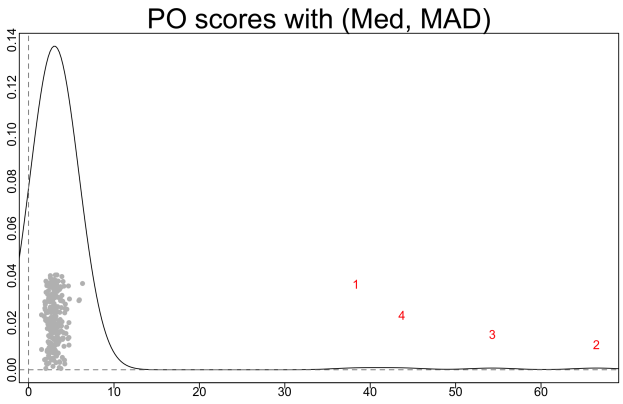


Figure 3.12: Distribution of PO scores of 200 data points. The four outliers highlighted by a red font are clearly distinguished from the other data points.

in the previous subsection using the true number of the spike components, $K = 10$. The largest value in each column is indicated using red, which shows that the MODs $h_1^*, \cdots, h_4^*$ explain approximately 90% of the underlying outlier direction e_{10}. This indicates that the MODs well approximate the underlying outlier direction and thus provide more informative directions than the individual PC directions for identifying outliers.

	h_1^*	h_2^*	h_3^*	h_4^*	h_5^*	h_6^*
1	0.000	0.000	0.000	0.000	0.006	0.010
2	0.000	0.000	0.000	0.000	0.005	0.010
3	0.000	0.002	0.000	0.002	0.182	0.004
4	0.000	0.001	0.000	0.001	0.030	0.240
5	0.001	0.001	0.001	0.001	0.028	0.163
6	0.001	0.000	0.002	0.000	0.216	0.035
7	0.003	0.000	0.000	0.000	0.228	0.217
8	0.001	0.003	0.000	0.003	0.103	0.161
9	0.001	0.000	0.000	0.000	0.025	0.002
10	0.868	0.871	0.876	0.871	0.011	0.000
11	0.000	0.000	0.000	0.000	0.000	0.000
12	0.000	0.000	0.000	0.000	0.000	0.000
	⋮	⋮	⋮	⋮	⋮	⋮
angle	21.3	21.0	20.6	21.0	84.0	89.1

Table 3.3: This table shows the estimated MODs of $X_1, \cdots, X_6$, denoted by $h_1^*, \cdots, h_6^*$, with squared entries. The first 4 columns $h_1^*, \cdots, h_4^*$ correspond to the outliers and the last two columns h_5^*, h_6^* correspond to two of the non-outliers. The largest value in each MOD is colored using a red font and the row corresponding to the outlier signal, e_{10}, are highlighted using a light blue background. This row indicates how much each h_i^* explains the outlier signal. The last row illustrates the angles between h_i^* and the true outlier direction, e_{10}. The first four MODs corresponding to the outliers are close to the true outlier direction.

A collection of the PO scores of the 200 data points are shown in Figure 3.12 with a kernel density estimate. Each point represents each data point and the four outliers, $X_1, \cdots, X_4$, are indicated by red indices. The four outliers have significantly large PO scores and are clearly separated from the other data points with the small PO scores. This example shows that the proposed procedure effectively extracts the latent outlier direction as well as isolates strong outliers from the other data points.

3.5.2 Local shape change detection

We now propose a second step procedure to deal with more challenging situations when outlying features are not distinguishable using a low rank representation. As mentioned earlier, the features whose signals are not strong enough to dominate increasing dimensions are the ones that may not be captured by the first few PCs. Then, these features remain in the residuals, which still suffer from high dimensionality, and thus simply applying the projection depth idea or other conventional outlier detection algorithm to the residuals may not be appropriate. To address this issue, this second step proposes to benefit from available knowledge about the outlying structures to extract critical information from many irrelevant features in an overwhelming number of dimensions.

As we stated earlier, shape variations missing from the low rank approximation often exhibit changes in a limited region of coverage. This is roughly because the required level of the signal for a latent local abnormality to be distinctively expressed need not be that intensive as required for a global abnormality. Such local shape changes are characterized by loss or gain in a whole/part of an exon or intron. Because widespread noise in the whole domain challenges the local variations to be found, it is advantageous to separately deal with the relevant local regions. An intuitive and simple way of doing this is to slide a window along the domain and hook potential outliers within each window area as follows.

We first define a *window direction* as a unit vector whose entries corresponding to a given window area are all equal to a constant and the rest of the entries are all zero. Then, a group of window directions can be a useful source for the projection depth function to measure local

abnormalities. The sparsity of a window direction helps to reduce the impact of noise and thus to separate meaningful outliers from inliers. It may be important to choose an appropriate size of window because too small windows can be sensitive to too fine variations and too large windows can be vulnerable to noise accumulation. We propose to use $50 \sim 200$ for a window size because most meaningful local variations often appear as such lengths of changes in expression. In addition to a specific length of windows whose union spans the transcript of a given gene, we include windows each of which corresponds to a whole exon or a whole intron. The window directions with these windows are useful to more accurately capture the whole exonic changes or whole intronic changes. We denote a collection of these window directions by $\mathbf{w}$.

The basic idea of the second step is the same as the first step in the sense that we find a direction that maximizes the one-dimensional outlyingness for each data point. However, the directions involved in the projection depth function are given some structure, which helps to accent local abnormality. This means that we take the set $\mathbf{v}$ (a collection of direction vectors that the depth function will examine) to be $\mathbf{w}$. Let I_2 be a set of the remaining sample indices after excluding the first step outliers and also let $\check{\mathbf{X}}$ denote the residual matrix whose columns ($\check{X}_j$) correspond to each of the remaining samples. Then, the second step PO scores can be written as

$$o_2(\check{X}_j|\check{\mathbf{X}}) = \sup \left| \frac{h^T \check{X}_j - \text{Med}(h^T \check{\mathbf{X}})}{\text{MAD}(h^T \check{\mathbf{X}})} \right| \quad \text{s.t. } h \in \mathbf{v},\ \phi(h^T \check{\mathbf{X}}) \leq \rho, \tag{3.5.15}$$

where ϕ and ρ for the normality condition as introduced in (3.5.14). We can obtain the MOD by choosing the direction giving the largest value among $\mathbf{v}/\mathbf{v}'$ where $\mathbf{v}'$ is a set of directions with $\phi(h_k^T \check{X}_j) > \rho$ for $h_k \in \mathbf{v}$. The MODs maximizing (3.5.15) can be used for interpreting the local abnormality. For a given outlier, the corresponding MOD informs the specific region where shape aberration occurs as well as the type (loss/gain) of the aberration involved.

In contrast to the PO scores o_1 from the global shape change detection procedure, the distribution of the statistic o_2 is unknown so that a cutoff value to determine outliers should be carefully chosen. For detecting outliers in the data from an unknown distribution, it is common to use a

box-plot and follow the rule that a point beyond an upper outer fence ($Q_3 + 1.5 \times IQR$) is considered as an outlier. However, this rule does not reflect the potential skewness of the distribution because the MAD does not account for the asymmetry. This may result in false discoveries on one side of the distribution or mask actual outliers on the other side. Rousseeuw et al. (2016) proposed an alternative to single robust scale measures to account for the potential asymmetry of a distribution. They suggested to separately apply a robust scale estimator to each of two subsamples that are above and below the median. We employ this idea to compute the second step PO scores o_2 which often show a right-skewed distribution.

Obtain the two scale estimates of $\{o_2(\check{X}_j|\check{\mathbf{X}})\}$, denoted by s_L and s_R for the left and right side, respectively. Then the proposed cutoff can be chosen by the following rule:

$$c_2 = Med(o_2) + s_R \times \Phi^{-1}(1-\alpha).$$

We only look at the right side of the distribution because we are interested in detecting a data point involved with abnormally large outlyingness.

To sum up, the local shape change algorithm is proposed as follows:

1. Collect a set of window directions.

2. Get the residual vectors $\check{X}_j$ for $j \in I_2$. For each $\check{X}_j$, obtain the second step PO score $o_2(\check{X}_j|\check{\mathbf{X}})$ based on (3.5.15) with $\mathbf{v}$ being a collection of the window directions considered. Here, $\check{\mathbf{X}}$ is a matrix whose columns are $\check{X}_j$ for $j \in I_2$, but it does not need to contain the jth sample for the jth sample inspection.

3. Collecting the PO scores, declare a set of data points as outliers if $o_2(\check{X}_j|\check{\mathbf{X}}) \geq c_2$ with a pre-determined level α.

3.6 Results

Let us now revisit the HNSCC RNA-seq data. We originally collected the 522 tumor samples for 20275 genes, and the filtering steps of Scissor discussed in Sections 3.4.1 and 3.4.2 identified 70 degraded samples and 5802 on/off genes, which were excluded from the downstream analysis. In this chapter, we apply the other parts of Scissor to the remaining 452 samples and 14473 genes. We first report the results at some important genes in Section 3.6.1 and then present the results from the genome-wide level analysis in Section 3.6.2.

3.6.1 Per-gene analysis

3.6.1.1 TP53

The TP53 gene is known as one of the most commonly mutated genes and mutations in this gene are associated with a wide range of human cancers. This gene encodes a tumor suppressor protein, called p53, which regulates cell division by keeping the cell from dividing too quickly, and thus the inactivation of this gene causes cancer (Olivier et al., 2009; Network et al., 2014). It is also known that alternative splicing of this gene results in multiple transcript variants and isoforms.

We applied Scissor to the gene TP53 with the first 12 PCs (chosen by the method that will be introduced in Chapter 3.6.2) for the global shape change detection method and the window size of 50 for the local method. The left and right panels of Figure 3.13 show the PO scores obtained from the global and local shape change detection procedures, respectively. Scissor detected 47 shape changes in total based on the cutoff values with the significance level $\alpha = 10^{-5}$, indicated by the black vertical lines in each panel. From the global shape change detection procedure (left panel), we identified 32 abnormal cases, shown as blue points, with the other grey non-outliers. The local shape change detection (right panel) identified 15 outliers, indicated as orange points.

An outlier (TCGA-BA-6871) whose outlyingness is the most extreme among the identified global shape changes is studied on the left column of Figure 3.14 with its MOD and the projected values on this MOD. The MOD is characterized by exceptional richness at the 3rd intronic region,

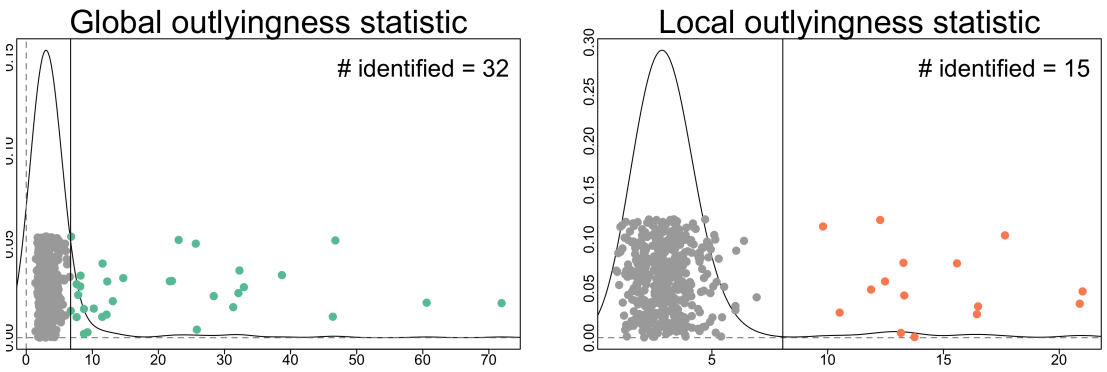


Figure 3.13: The left and right panels show the PO scores from the global and local shape change detection procedure, respectively, with kernel density estimates. The colored points indicate the identified outliers from each step based on the cutoff values indicated by the black vertical lines. These procedures identified 32 global and 15 local shape changes, respectively.

which helps to interpret the type of aberration that the red outlier is involved in. The distribution of the scaled projected values using $\frac{x-Med(\mathbf{x})}{MAD(\mathbf{x})}$ (robust Z-scores) obtained by projecting all samples onto this direction is represented with a kernel density estimate. In this plot, each point represents each sample, and there are 6 samples in different colors including the red one, which have abnormally large scores. The samples with such large scores are expected to be the ones that strongly go in this MOD. The coverage overlays for those samples are shown in Group 2 in Figure 3.16 and in Group 4 in Figure 3.18. Indeed, they are all associated with the 3rd intron retention as illustrated in the MOD. In particular, four of them (Group 2 in Figure 3.16) have the splice site mutations at the same location near the observed shape changes. This shows that the MOD not only helps to understand the abnormality of outliers but also can be used to group the identified outliers that share similar abnormalities.

An example (TCGA-CV-6940) of the outliers from the local shape change detection procedure is shown on the right column of Figure 3.14. The MOD that gives the largest outlyingness to this example among the considered window directions is shown in the middle of the column, indicating that this sample may be associated with an abnormal loss of the 7th exon. The zoom-in figure around the window area of this panel highlights the deleted coverage. Notably, we found that there is a frameshift mutation, indicated by the black vertical line, near the observed shape change. The

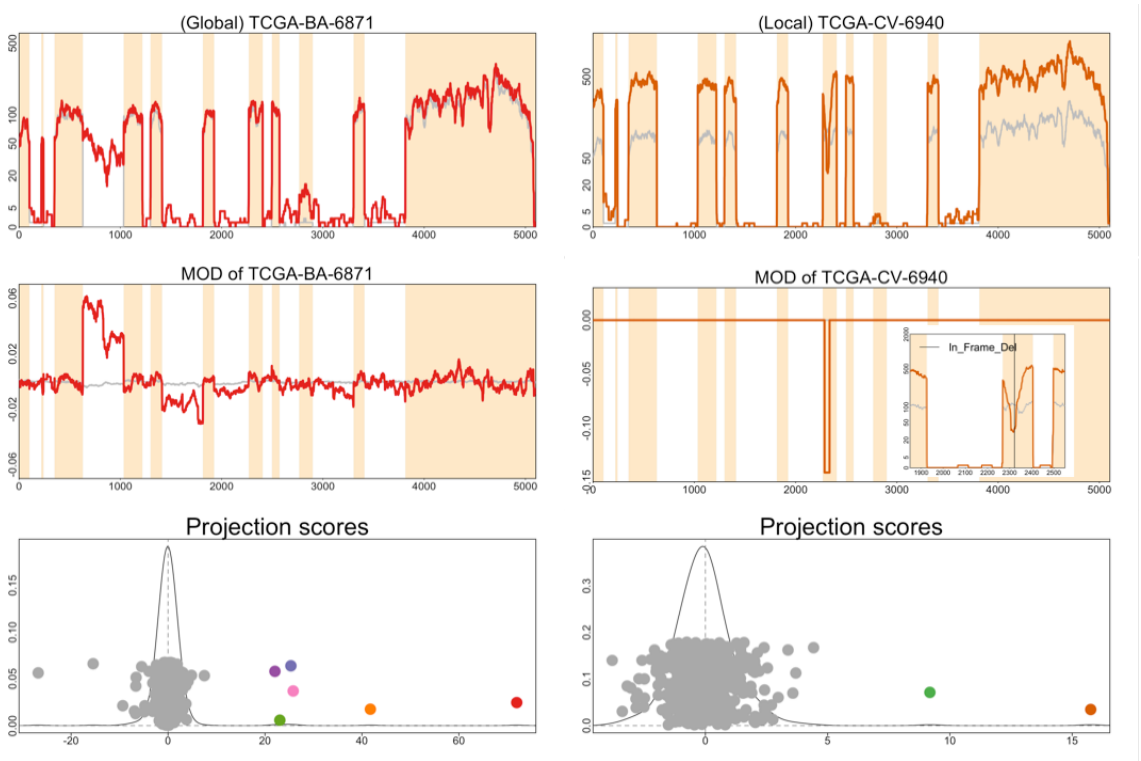


Figure 3.14: (Left) The top figure displays an example (Sample TCGA-BA-6871) of identified global shape changes. The middle panel shows the corresponding MOD, illustrating the abnormal high expression in the 3rd intronic region. The projected values onto this MOD are shown at the bottom with several colored points that have significantly high values. The most distinguished red point corresponds to this outlier. (Right) Another identified local shape change (Sample TCGA-CV-6940) is displayed in the top panel. And the corresponding MOD is shown below with a zoom-in plot to highlight the abnormally deleted area. The projected scores onto this MOD are shown at the bottom. The extreme values are colored and the most extreme one corresponds to this sample.

robustly normalized projection values obtained by projecting all cases (except for the identified ones from the first step) onto this MOD are presented in the bottom figure with a kernel density estimate. Two points with unusually large scores are colored and the most extreme one corresponds to this particular sample (TCGA-CV-6940). The other green sample (TCGA-CV-7178) is shown using the same color in Group 6 in Figure 3.17 and shows a similar shape change pattern. It turned out that this green sample also has a frameshift mutation in the middle of the area where the abnormal deletions are observed, strongly indicating that these mutations are responsible for the resulting shape changes.

As we already mentioned with some of the above examples, a large fraction of the detected shape changes are associated with mutations. Table 3.4 summarizes how many mutations were associated with shape changes for each mutation. More than half of the splice site mutations were identified and most of them were from the global method, which potentially implies that splice site mutations are more associated with global shape changes in the gene TP53. On the other hand, most of the identified frameshift mutations were from the local method, which implies that shape changes from frameshift mutations tend to be smaller. Although there were two samples with silent mutations that have shape changes, those samples have non-silent mutations as well that may result in shape changes. So this does not support the notion that silent mutation is associated with shape changes. Among the identified 47 shape changes, we found that 34 of them are associated with the called mutations.

Among the identified outliers, the shape changes associated with a splice site mutation are shown in Figure 3.16. The groups here are based on the location where splice mutations occur, indicated by the red vertical lines, and so the outliers in each group share similar shape change patterns. The samples in Figure 3.17 are all the examples where frameshift mutation was associated with shape changes, and we grouped them by the mutation position. The mutation of each sample is indicated by color.

It should be also pointed out that the 17 splice site mutations that were not identified by Scissor (Table 3.4) have no significant shape changes even though splice site mutations were called. One example is shown in the top row of Figure 3.15. The red vertical line indicates the position of the called splice site mutation which is expected to retain the adjacent intronic reads. However, no clear

Mutation (452)	# of outliers (Step 1, Step 2)
Splice site (39)	22 (19, 3)
Nonsense (58)	4 (2, 2)
Silent (5)	2 (1, 1)
Frame shift (67)	9 (2, 7)
Other (196)	9 (6, 3)

Table 3.4: Table for the number of coincidences of each mutation and significant shape changes.

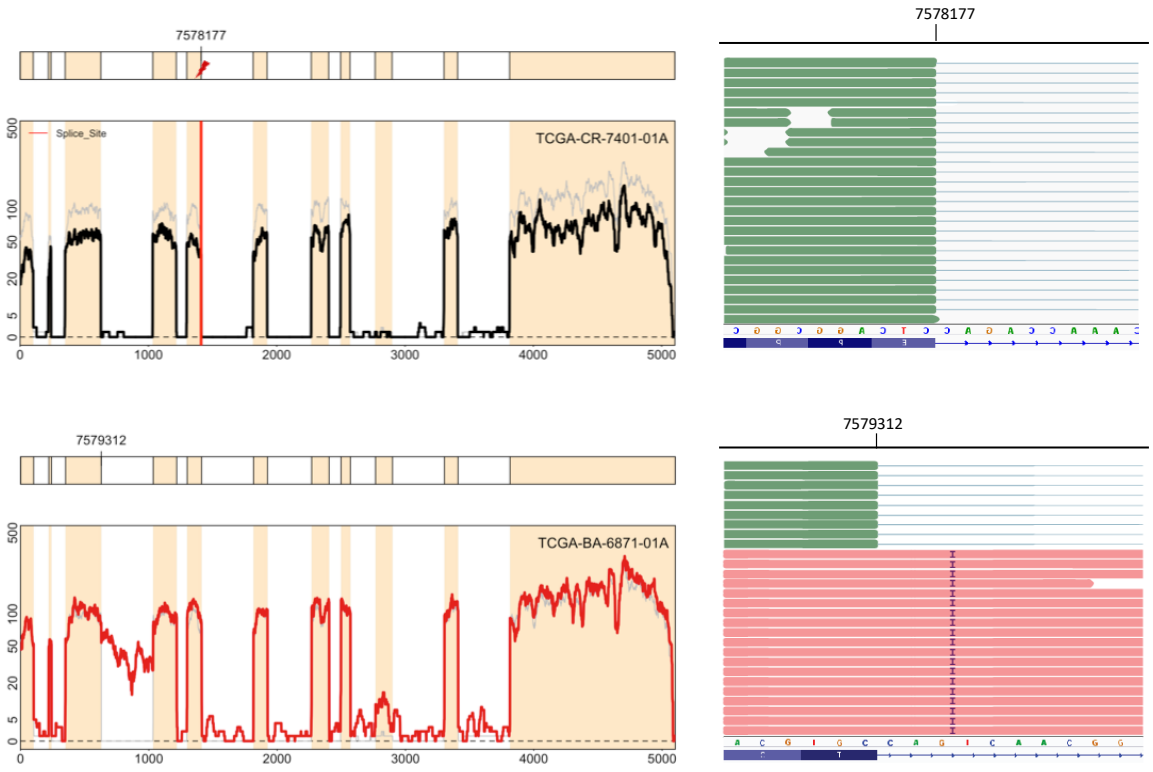


Figure 3.15: Top: an example with an inactive splice site mutation. The red vertical line indicates where the mutation is called. On the right, the IGV screenshot shows the reads mapping to the exon-intron boundary with the splice mutation called. All reads are green indicating normal splicing, despite the mutation called. Bottom: an identified shape change in the absence of any mutation called. The IGV screenshot shows the reads aligned to the exon-intron boundary near the observed shape change, with pink indicating abnormally spliced reads. We observed many abnormal reads that span the exon-intron junction and involves insertion, despite no mutation called.

shape change was observed near the position. For a more detailed view, we present the Integrative Genomics Viewer (IGV) screenshot which shows the reads aligned near the mutation position for this sample. It clearly shows that the reads in this area are normally split and the flanking intron is not covered. Although the screenshot exhibits only a subset of the reads, we confirmed that there was no retained read spanning the exon-intron junction. This observation demonstrates that this sample does not involve abnormal splicing, potentially indicating that the splice mutation was inactive or was incorrectly called.

Other interesting examples are shape changes identified from Scissor in the absence of mutations called. We identified 13 such shape variants and one of them is displayed in the bottom row of

Figure 3.15. This sample (TCGA-BA-6871) is the one with the most extreme outlyingness that we previously saw in Figure 3.15, and it is characterized by the retained 3rd intron. The IGV screenshot taken for a subset of the reads mapping to the boundary of the exon flanking the $5'$ end of the 3rd intron is displayed to right of the sample overlay. Notably, it shows that there are many abnormally retained reads (pink) that span the exon-intron junction. Moreover, these reads all have insertions at the location 4bp away from the exon-intron junction, which implies that the insertion is an important variant responsible for the intron retention, but was missed by the variant caller.

All the 13 shape variants with an absence of called mutations are displayed in Figure 3.18. The samples are grouped by their distinct shape changes. The three samples on the top row show several consecutive exons skipped, which is often observed under the intergenic deletion. The shape changes in Group 4 are characterized by their abnormalities at the 2nd or 3rd intronic regions. Group 5 shows abnormally high read coverage at the 6-9 intronic regions. Lastly, the two cases in Group 6 do not really exhibit apparent shape abnormality, but it should be noted that these samples are expressed at high levels with ambiguously retained reads at a wide range of introns.

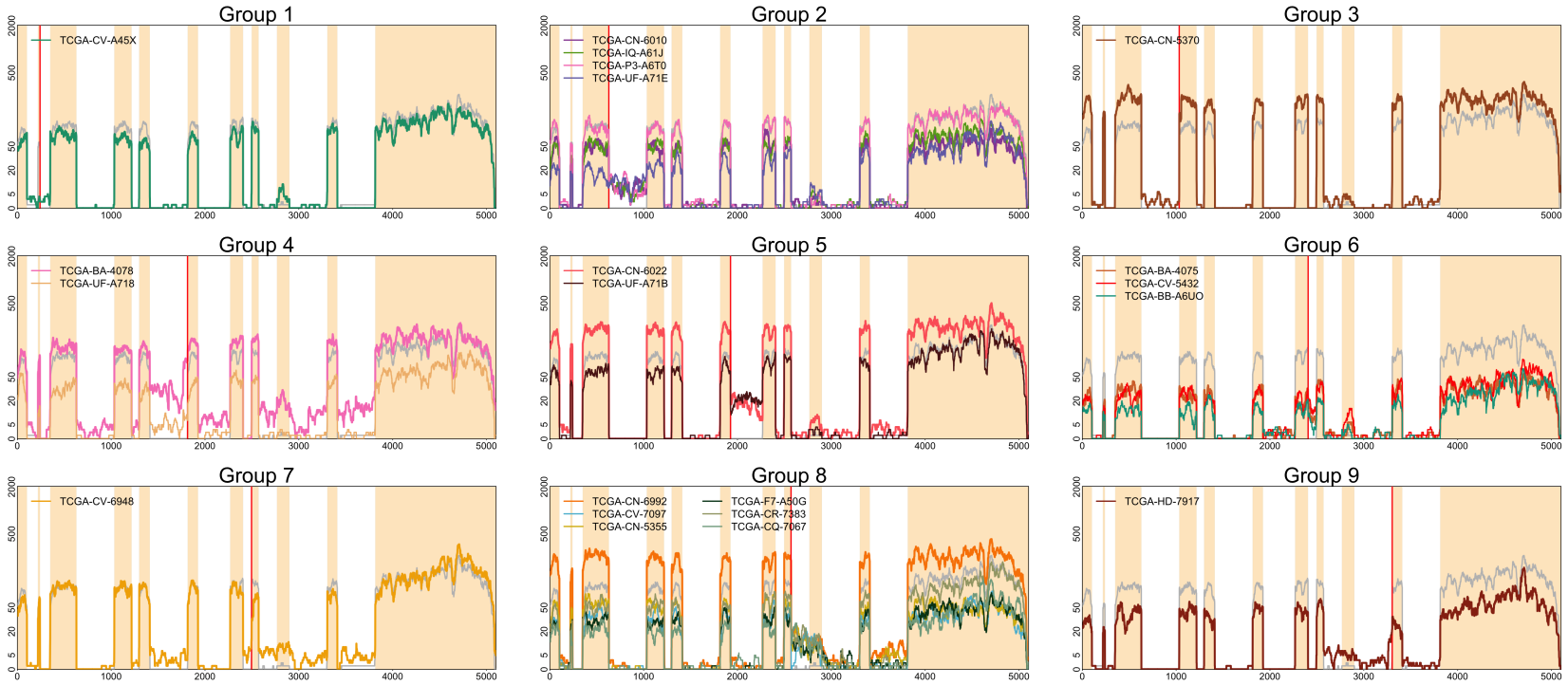


Figure 3.16: The 21 outliers involved in distinct splice site mutations among the 47 identified outliers. The outliers are grouped by the sites of the mutations and the outliers in each group exhibit similar types of aberration.

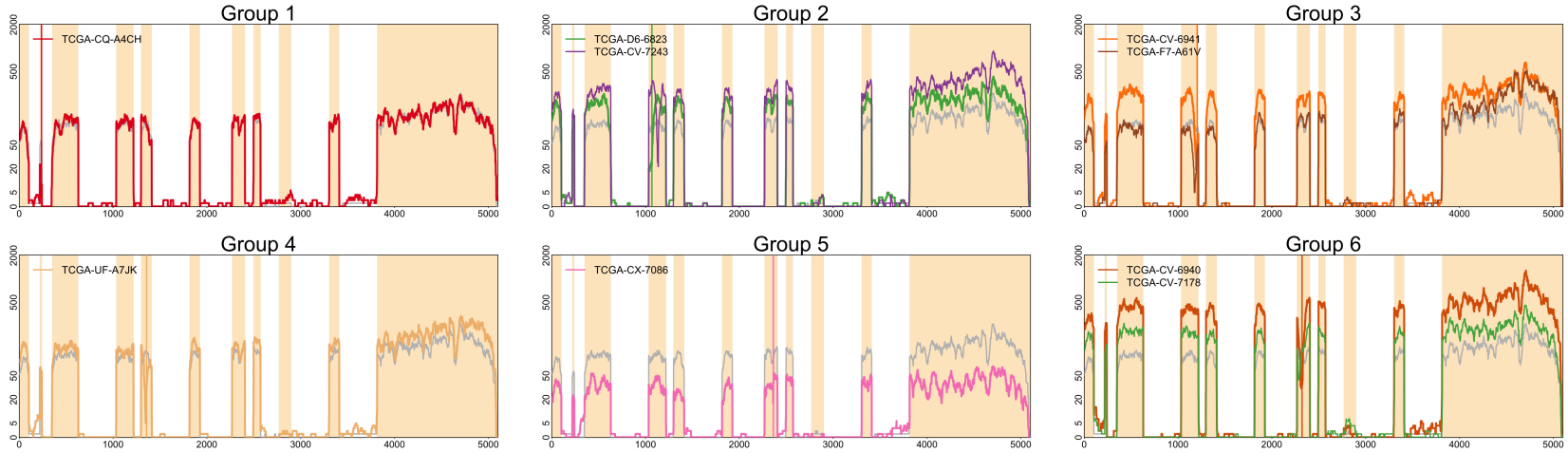


Figure 3.17: The 9 shape change cases involved with frameshift mutations. These outliers are grouped based on where the shape changes occur.

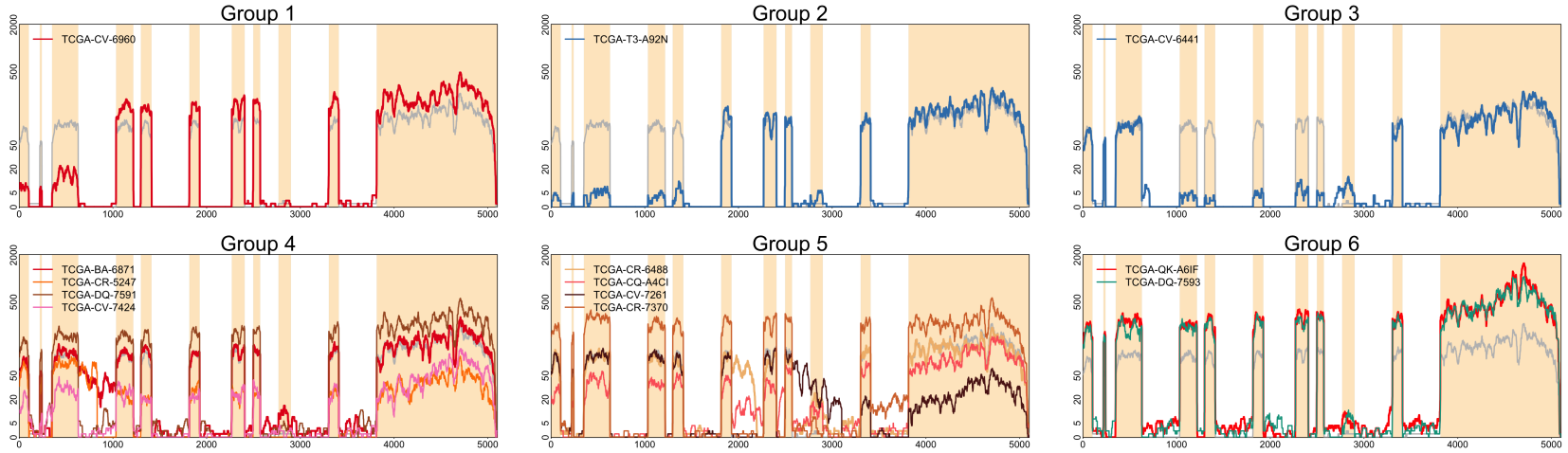


Figure 3.18: The 13 RNA-seq outliers that are not associated with any mutations are displayed. The outliers are grouped by the types of aberration. The groups 1-3 exhibit intragenic deletions at different regions. Group 4 retains the 2nd intron or the 3rd intron. Group 5 retains several different introns. The samples in group 6 are highly expressed and exhibit moderate levels of intron retention at almost all intronic regions.

3.6.1.2 CDKN2A

The CDKN2A gene is known as a frequently altered tumor suppressor gene in a wide variety of tumors. In HNSCC, particularly, most of the mutations in this gene are somatic mutations, which can produce no functional p16 protein (Stransky et al., 2011; Loyo et al., 2013; Lim et al., 2014; Mountzios et al., 2014). This gene is also known to encode distinct proteins as a result of alternatively spliced variants and the identification of the associated clusters have been studied in Kimes et al. (2014).

We applied Scissor to the gene CDKN2A. The left and right panels of Figure 3.19 show the PO scores obtained from the global and local shape change detection procedures, respectively. Scissor detected 41 shape changes in total based on the cutoff values with the significance level $\alpha = 10^{-5}$, indicated by the black vertical lines in each panel. The 39 global shape changes were identified, indicated by the blue points, shown with the other grey non-outliers. The local shape change detection identified 2 outliers, indicated as orange points.

Comparison of the identified shape changes and the called mutations is summarized in Table 3.5. Notably, Scissor identified all 9 splice site mutations with strong outlyingness by the global method. The corresponding 9 samples can be divided into two groups based on mutation position and the resulting shape changes are shown in Figure 3.20. The samples in the left panel have splice

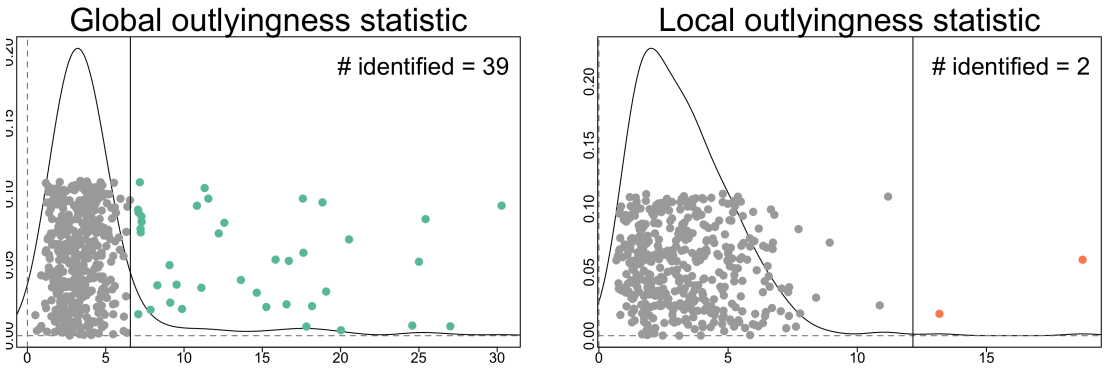


Figure 3.19: The left and right panels show the kernel density plots of the projection outlyingness $\{\text{PO}_j\}_{1\leq j\leq 452}$ from the global and local method, respectively. The detected outliers from each step are colored and the cutoff values are indicated by the vertical lines.

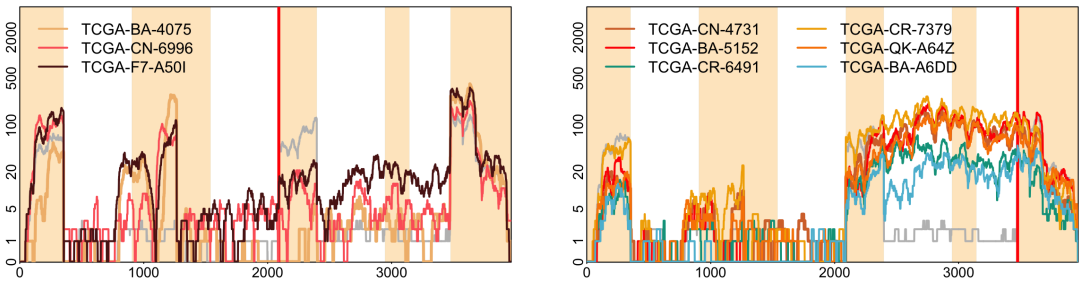


Figure 3.20: Identified splice mutations in the gene CDKN2A. The samples in the left panel have splice mutations at the right end of the 2nd intron, indicated by the red vertical line, which led to the skipped exons. The samples in the right panel have splice mutations at the right end of the last intron, which cause the adjacent intron retention.

site mutations at the $3'$ end of the 2nd intron, indicated by the red vertical line, which are responsible for skipping the flanking exon. The right panel shows the other 6 samples with splice site mutations that occur at the $3'$ end of the last intron. As a result, those samples all retain the adjacent intron.

The other shape variants that do not have splice site mutations are displayed in Figure 3.21. For example, the samples in Group 1 skipped the 1st exon and part of the 3rd exon as well as abnormally retained read coverage from the 3rd intron to the 4th intron. An interesting group is Group 7 whose samples are characterized by different exon usage of the 4th exon. While the other samples do not use this exon, only these two samples have expression of this exon.

Mutation (452)	# of outliers (Global, Local)
Splice site (9)	9 (9, 0)
Nonsense (53)	2(2, 0)
Silent (0)	0 (0, 0)
Frame shift (18)	0 (0, 0)
Other (20)	2 (2, 0)

Table 3.5: Table for the number of coincidences of each mutation and significant shape changes.

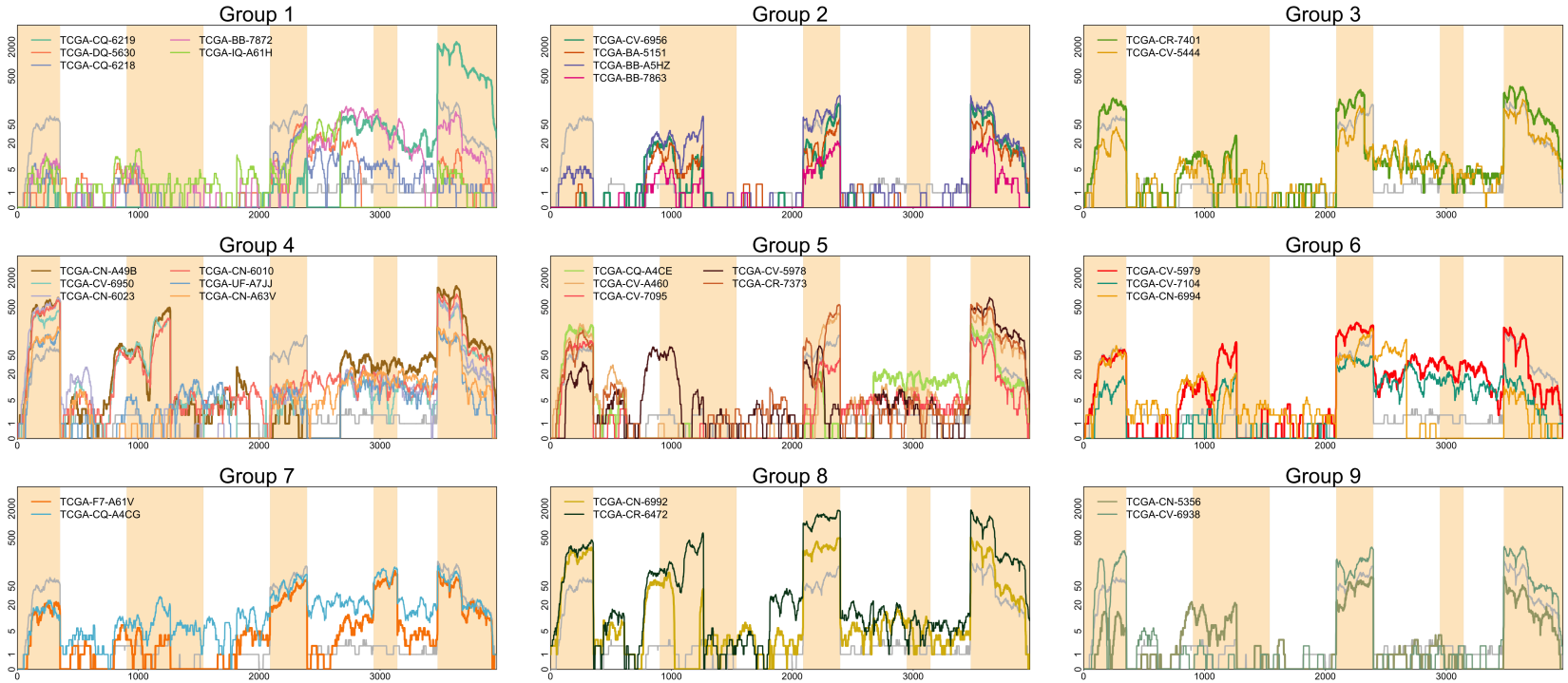
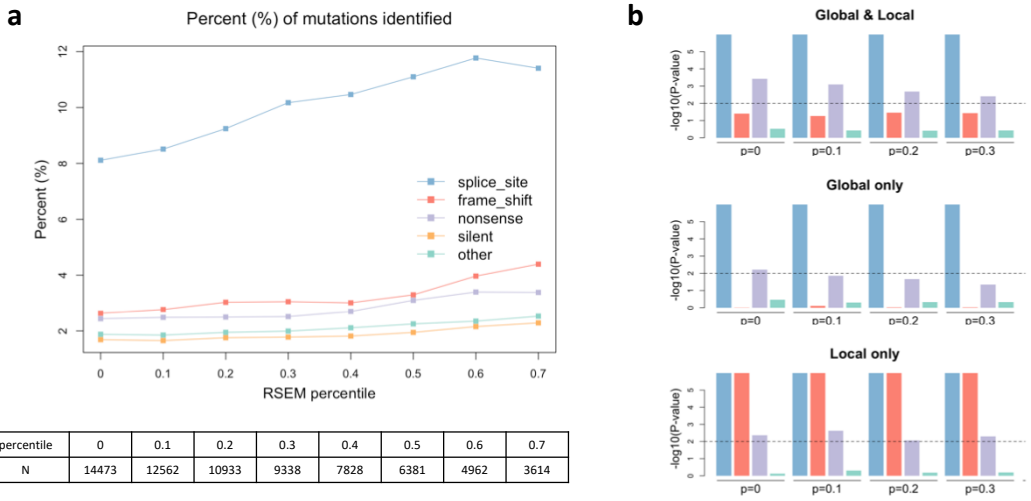


Figure 3.21: Identified shape changes in the absence of splice site mutations in the gene CDKN2A. The outliers are grouped by distinct shape abnormality patterns.

3.6.2 Genome-wide analysis

We applied Scissor to 14473 genes with a significance level $\alpha = 10^{-5}$ for the global and local shape change detection procedures. On average, four shape changes were identified at each gene, but the number of outliers show a very skewed distribution. A group of genes (e.g. TP53, CDKN2A, FAT1) have a lot more shape changes than expected ($\sim 10\%$ of the samples) whereas the mode of the distribution is zero, indicating that a substantial number of genes have no shape changes.

Figure 3.22 illustrates association between each mutation type and shape change. Each line of the left panel (a) shows percentage of each mutation identified across the cohort with different sets of genes filtered by RSEM (RNA-Seq by Expectation Maximization) (Li and Dewey, 2011). The RSEM software quantifies gene and isoform abundances by computing maximum likelihood



percentile	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
N	14473	12562	10933	9338	7828	6381	4962	3614

Figure 3.22: Mutational analysis with the genome-wide results from Scissor. Panel (a) presents the percentage of identified mutations in different sets of genes filtered by RSEM. Each line represents each mutation type. As we consider more highly expressed genes, the overall percentage goes up. Panel (b) reports the output from Fisher's exact test to see if there is any significance difference in the number of identified mutations with silent mutations as a base line. The top plot shows the results with the global and local outliers together, and splice site mutations and nonsense mutations are found to be significant. The middle plot shows the results with the global outliers only, and splice site mutations are strongly significant but the other mutation types are weak. The bottom plot reports the results from the local outliers only, showing that splice site, frameshift, and nonsense mutations are significantly higher than silent mutations.

estimates based on the EM algorithm. Since a highly expressed gene often shows high RSEM values with large variability, we computed median and MAD of RSEM values for each gene and gradually filtered out low-expressed genes based on the percentiles ($p = 0.1, 0.2, \cdots, 0.7$) of the computed median and MAD. The horizontal axis of panel (a) indicates this percentile and the table below shows the number of genes considered at each percentile threshold. Overall, the percentage of identified mutations goes up as we consider more highly expressed genes, and this is because mutations in low coverage are often inactive. Also, Scissor identified a much higher fraction of splice site mutations (blue line) than the other mutation types, indicating that splice site mutations are more responsible for shape changes.

A silent mutation makes a change in a DNA codon that has no corresponding change in the amino acid translation. Thus, a silent mutation is not expected to have shape changes, also as supported by the low percentage (yellow) in the Figure 3.22. An interesting question is whether the numbers of identified mutations are significantly higher than identified silent mutations which can be considered as a base line. To see if there is any significant difference, we performed Fisher's exact test with a silent mutation as the reference. The bar plots in the panel (b) show the results from the exact test for both global and local outliers, global outliers only, and local outliers only. Each bar represents each mutation type and the y-axis indicates $-\log_{10}$(p-value). Overall, splice

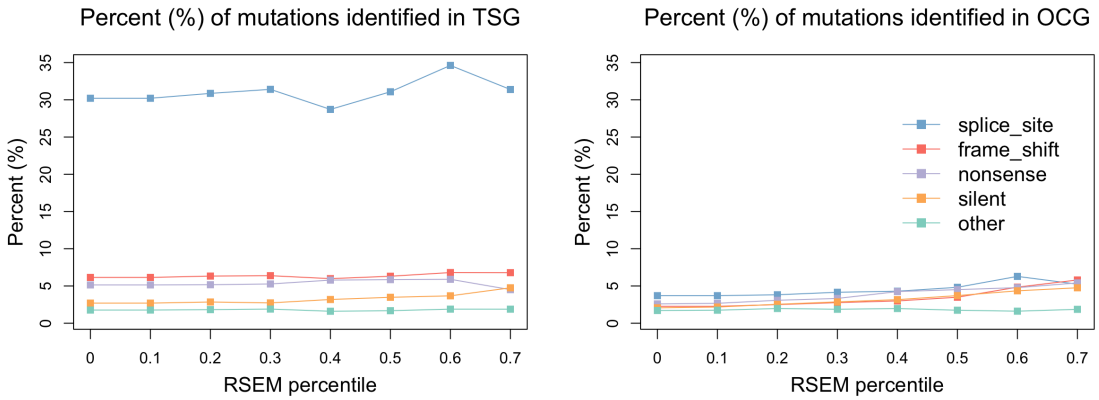


Figure 3.23: (Left) Percentage of identified mutations in a TSG gene set. (Right) Percentage of identified mutations in an oncogene set. Overall, the percentages in TSG are higher than the percentages in OCG. In particular, significantly high fraction of splice site mutations are associated with shape changes in TSG.

site mutations and nonsense mutations are strongly associated with both outlier types in our shape change analysis as shown in the top of Panel A. And as expected, the global shape changes (middle) are more associated with splice site mutations whereas the local shape changes (bottom) show strong association with more mutation types, in particular, splice site, frameshift, and nonsense mutations. This is partially because a splice site mutation tends to result in a larger shape change because it often skips a whole (or several) exon(s) or a whole (or several) intron(s). On the other hand, frameshift and nonsense mutations often lead to smaller shape changes.

As we mentioned earlier, a tumor suppressor gene (TSG) regulates cell division by keeping the cell from dividing too quickly, and thus the inactivation of a TSG can lead to cancer. On the other hand, an oncogene is a gene that is not supposed to be activated, and when it is turned on, cells grow or divide out of control, which can cause cancer. Figure 3.23 exhibits percentage of identified mutations in TSG (left) and in oncogenes (right). Percentage of identified mutations in TSG are mostly higher than the ones in Figure 3.22 (a), and remarkably, more than 30% of splice site mutations in TSG were identified, which is substantially higher than before. Therefore, this supports the notion that shape changes are highly associated with tumor suppressor genes. On the other hand, we observed much smaller proportions in oncogenes compared to TSG. In particular, identified splice mutations are no longer significantly higher than the other mutation types.

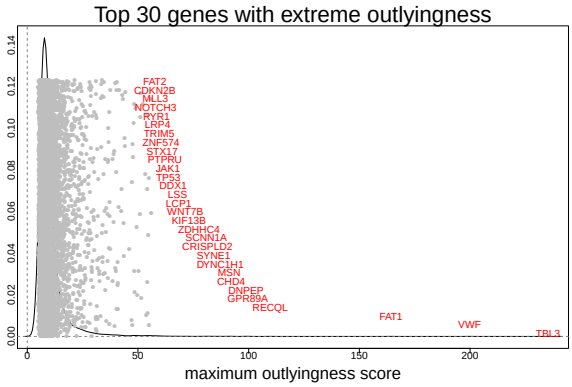


Figure 3.24: Distribution of maximum scores with a kernel density estimate. The x-axis indicates the scores and the y-axis indicates the height of the kernel density estimate. Each point represents one gene and the height of the points are random jitter except for the top 30 genes. Those top 30 genes with the most extreme maximum outlyingness scores are indicated using red text and include some important TSG (FAT1, TP53, JAK1).

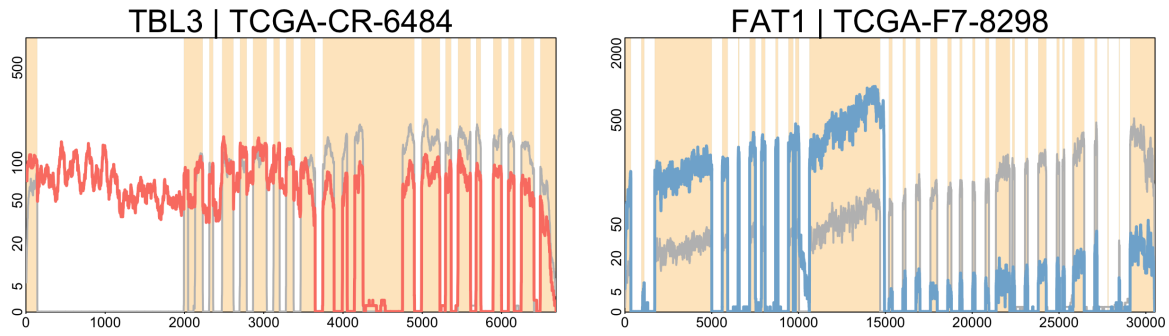


Figure 3.25: The most outlying sample in each of the two genes, TBL3 and FAT1, is shown in each panel. These two genes are chosen due to their extreme scores as illustrated by Figure 3.24. The sample (TCGA-CR-6484) from the gene TBL3 is distinguished by its unusually retained introns. The sample (TCGA-F7-8298) from the gene FAT1 abnormally retains a part of the intron near base position 15000 and skips all following exons.

An interesting observation of the genome-wide analysis of Scissor is that many of the biologically important genes have extreme outlyingness scores (approximately greater than 20). For example, the genes TP53 and CDKN2A have their maximum outlyingness scores of 72 and 30, respectively, as shown in Figure 3.13 and Figure 3.19. Such large scores indicate that the corresponding shape variations have strong underlying signals, which are more likely to be novel. Therefore, the maximum outlyingness score from each gene can be used to discover some key genes. Figure 3.24 shows a distribution of the largest scores from genes with a kernel density estimate. Each point represents a gene and the x-axis indicates the maximum scores. Here, the y-axis indicates the height of the kernel density estimate, and the heights of most points are random. The top 30 genes with the most extreme scores are shown in red text and some important tumor suppressor genes, FAT1, TP53, and JAK1, are included in the list. Here, we only highlight the top 30 genes where the maximum scores are greater than 50, but we observed that scores around 20 tend to be extreme as well. In this way, Scissor provides some key genes that are potentially important and need further investigation.

The gene TBL3 is the gene whose maximum outlyingness score is the most extreme across genes as indicated by Figure 3.24. The most outlying sample (TCGA-CR-6484) from this gene is displayed in the left panel of Figure 3.25. This sample is distinguished by its abnormal retention of several introns while the other samples have normal splicing as indicated by the mean curve (grey). Another extreme sample from the gene FAT1 (the 3rd gene in Figure 3.24) is illustrated in the right

panel of Figure 3.25. This sample retains the part of the intron near base position 15000 and then skips all following exons. While this sample has no mutations called, we have found many insertions and deletions in the reads aligned to the exon-intron junction where intron retention occurs.

CHAPTER 4. DETERMINING THE NUMBER OF SPIKES IN PCA

4.1 Motivation

Although the sample covariance matrix is not a consistent estimator for the population counterpart for a large d, the sample covariance matrix analyzed by PCA often reveals important underlying structure of high dimensional data. As in mentioned in Chapter , Johnstone (2001) introduced the spike covariance model where all but finitely many eigenvalues of the population covariance matrix are the same. Here, we denote the large eigenvalues in (1.1.2) by $\alpha_1, \cdots, \alpha_K$ for convenience, i.e.

$$\mathbf{U}^T \mathbf{\Sigma} \mathbf{U} = \mathbf{\Lambda} = \text{diag}(\alpha_1, \cdots, \alpha_K, \sigma^2, \cdots, \sigma^2) \tag{4.1.1}$$

where $\mathbf{\Sigma}$ is the population covariance matrix and has the spectral decomposition $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ with $\alpha_1 \geq \cdots \geq \alpha_K > \sigma^2 > 0$. The eigenvalues $\alpha_1, \cdots, \alpha_K$ correspond to spikes and they are mostly assumed to be much larger than the non-spike eigenvalues. The directions corresponding to spikes can be considered as important underlying structure of data whereas the other directions corresponding to non-spikes can be considered as noise. Under the spike covariance model, many interesting asymptotic properties have been investigated in, for example, Kritchman and Nadler (2008, 2009), Johnstone and Lu (2009), Jung and Marron (2009), Benaych-Georges and Nadakuditi (2011), and Shen et al. (2012). In particular, Jung and Marron (2009), Shen et al. (2012),and Theorems 2.5.1 and 2.5.2 studied in Chapter 1.3 showed PCA subspace consistency. These studies show that PCA is a useful tool for dimension reduction and data visualization for high dimensional data.

One challenging problem in PCA is how to determine the number of spikes K. The scree plot is one of the popular ways that have been proposed. Based on the plot of ordered eigenvalues, one

looks for an elbow that distinguish the eigenvalues that are remarkably large from relatively small eigenvalues. Although this method is very simple, looking for an elbow is often subjective and it may be unclear that the elbow really gives the meaningful separation. Moreover, when one has a large number of datasets to be analyzed, it is hard to look at all scree plots corresponding to each dataset. For example, when applying PCA to RNA-seq data for many separate genes as in Kimes et al. (2014), visual inspection would entail looking at more than 20,000 scree plots and deciding on elbows. This is obviously intractable.

There has been much research on how to determine the number of components. Most of the previous works assumed that the population eigenvalues corresponding to non-spikes are all equal as in (4.1.1). When non-spike eigenvalues are all the same, the empirical distribution of non-spike sample eigenvalues except for a few first large ones should be close to the standard Marcenko-Pastur distribution. However, we have observed datasets whose non-spike sample eigenvalues do not follow the standard Marcenko-Pastur distribution. They rather show even more heavy tails and right skewness. This observation motivates us to generalize the underlying distribution of the population eigenvalues for determining the number of principal components.

In this chapter, we review some important results in random matrix theory and introduce an idea to choose the number of spikes. The remainder is organized as follows. In Section 4.2, we review related work. Some important known results in random matrix theory are reviewed in Section 4.3. In Section 4.4, based on the theoretical results, the proposed method is described. We conclude this chapter with real data examples in Section 4.5.

4.2 Related work

Methods for determining the number of spikes in PCA have been developed in different fields such as the statistics (Besse and de Falguerolles, 1993; Krzanowski and Kline, 1995; Choi et al., 2014), signal processing (Wax and Kailath, 1985; Kritchman and Nadler, 2008, 2009), and econometrics (Harding et al., 2007; Passemier and Yao, 2012, 2014), with various approaches. For excellent background, see Chapter 6 of Jolliffe (2002). As a similar problem, a low rank matrix

reconstruction problem based on SVD has been studied in Wongsawat et al. (2005); Shabalin and Nobel (2013); Nadakuditi (2014).

To our best knowledge, it dates back to the works of Bartlett (1954) and Lawley (1956) which are based on likelihood ratio tests to check for equality of the smallest eigenvalue. The problem of testing the hypothesis of multiplicity of the smallest eigenvalue was also considered in Wax and Kailath (1985) and Zhao et al. (1986) based on various information theoretic criteria for model selection such as Akaike's AIC criterion and Schwart-Rissanen's minimum description (MDL) criterion. These methods are based on the large sample asymptotic and may not work well for high dimensional data with a limited number of observations.

Remarkable developments in random matrix theory allow us to understand asymptotic behaviors of the sample eigenvalues in high dimensional space. It follows that many methods for determining the number of signals have been introduced based on these asymptotic results. Kritchman and Nadler (2008) developed an algorithm for rank determination based on the asymptotic distribution of the largest noise eigenvalue. Their algorithm performs sequential hypothesis tests on whether the largest eigenvalue at each step arises from a signal rather than from noise. The statistical procedure at each step involves estimating noise variance and setting a threshold based on the Tracy-Widom distribution where the largest noise eigenvalue follows (Johnstone, 2001). Their algorithm has been considered as a good benchmark for judging performance of other methods for determining the number of components in many papers.

Passemier and Yao (2012, 2014) proposed a method estimating the number of spikes under the case where there are possibly equal spikes. Based on the different asymptotic behaviors of spike and non-spike sample eigenvalues, they determine a threshold for the successive spacings of the ordered sample eigenvalues. The larger spacings are expected for spike sample eigenvalues than for noise eigenvalues so that one may separate spikes and non-spike eigenvalues based on an appropriately determined threshold for the spacing. This method is very intuitive in the sense that the proposed procedure is somehow similar to the naive procedure based on scree plots with a more reasonable separation based on random matrix theory. To avoid false determination due to ties of

spike eigenvalues, they also proposed a more robust estimator by using consecutive two or more spacings that should be larger than a threshold at the same time to be considered as spikes.

Choi et al. (2014) introduced exact tests for rank determination based on the Kac-Rice test derived by Taylor et al. (2013). By formulating the problem in a regularized regression problem with a nuclear norm penalty, they obtained a conditional survival function of the kth singular value given all the other singular values as a test statistic for a hypothesis test on whether the rank of the unknown signal matrix is less than k or not. The test statistic is uniformly distributed under the null hypothesis and plays a role of a p-value. Their method may be more accurate than other methods based on asymptotics of sample eigenvalues because the procedure depends on the exact test rather than the asymptotic test.

4.3 Known results on the generalized spike covariance model

In this section, we review some important results of random matrix theory. Let $\{w_{ij}\}_{1\le i\le d, 1\le j\le n}$ be i.i.d. random variables satisfying

$$E(w_{11}) = 0, \;\; E(|w_{11}|^2) = 1, \;\; E(|w_{11}|^4) < \infty$$

and let (T_d) be a sequence of $d \times d$ nonnegative definite Hermitian matrices. Write $\mathbf{Z}_n = (\mathbf{u}_1, \cdots, \mathbf{u}_n)$ where $\mathbf{u}_j = (w_{1j}, \cdots, w_{dj})^T$ and $\mathbf{X}_n = T_d^{\frac{1}{2}} \mathbf{Z}_n$. Then, the sample covariance matrix of $\mathbf{X}_n$ can be expressed as $\mathbf{S}_n = \frac{1}{n}\mathbf{X}_n\mathbf{X}_n^T = \frac{1}{n}T_d^{\frac{1}{2}}\mathbf{Z}_n\mathbf{Z}_n^T T_d^{\frac{1}{2}}$. Let (H_n) be a sequence of the empirical spectral distributions of (T_d) with the dimension to sample size ratio, $y_n = \frac{d}{n}$, and assume that H_n weakly converges to a nonrandom probability distribution H on $[0, \infty)$ as n tends to infinity satisfying $y_n \to y$ where y is a positive constant. The limit distribution H is referred to as the population spectral distribution (PSD). For the simplest case where $T_d = I_d$, the ESDs of (T_d) are $H_n(t) = \delta_1(t)$ for all n and thus the PSD $H(t)$ can be obtained as $H_n(t) \to H(t) = \delta_1(t)$. The classical Marcenko-Pastur law says that the ESD of $\mathbf{S}_n$ converges to a nonrandom limiting spectral distribution (LSD) G under this case.

A generalized version of the classical Marcenko-Pastur law has been developed (Silverstein, 1995) when the PSD $H(t)$ is an arbitrary probability measure. The PSD H and the LSD G are linked by the following inverse map of the companion Stieltjes transform $\underline{s}(z)$ of the LSD G,

$$z = g_{y,H}(\underline{s}) = -\frac{1}{\underline{s}} + y\int \frac{t}{1+t\underline{s}} dH(t), \quad z \in \mathbb{C}^+,$$

which is called the Silverstein equation. Let $F_{y,H}$ be the distribution whose Stieltjes transform is $m_{y,H} = g_{y,H}^{-1}$. Throughout the proposal, we call $F_{y,H}$ the Marcenko-Pastur (MP) distribution with indexes (y,H). A lot of theory and applications in the spectral analysis have been established from this equation. One crucial result is the Lemma 4.3.1 which indicates the analytical relationship between the two supports of the PSD H and the MP distribution $F_{y,H}$ (Silverstein and Choi, 1995). Define

$$\psi_{y,H}(\alpha) = g_{y,H}(-1/\alpha) = \alpha + y\alpha\int \frac{t}{\alpha - t} dH(t) \tag{4.3.1}$$

for $\alpha \notin \Gamma_H$ and $\alpha \neq 0$.

Lemma 4.3.1. *(Silverstein and Choi, 1995) If* $\lambda \notin \Gamma_{F_{y,H}}$, *then* $m_{y,H}(\lambda) \neq 0$ *and* $\alpha = -1/m_{y,H}(\lambda)$ *satisfies*

(a) $\alpha \notin \Gamma_H$ *and* $\alpha \neq 0$ *(so that* $\psi_{y,H}(\alpha)$ *is well-defined);*

(b) $\psi'_{y,H}(\alpha) > 0$.

Conversely, if α *satisfies (a)-(b), then* $\lambda = \psi_{y,H}(\alpha) \notin \Gamma_{F_{y,H}}$.

Roughly speaking, what the lemma says is that given the PSD H one can characterize the support of the MP distribution $F_{y,H}$ as illustrated in Figure 4.26. In Figure 4.26, the curve indicates the $\psi_{y,H}(\alpha)$ with $y = 2$ and the PSD $H = \delta_1$ and the regions indicated by yellow are $\{\alpha\}$ satisfying (a) and (b) in Lemma 4.3.1 and the corresponding $\psi_{y,H}(\alpha)$. According to Lemma 4.3.1, the support of the MP distribution $F_{y,H}$ is indicated as blue on the y-axis.

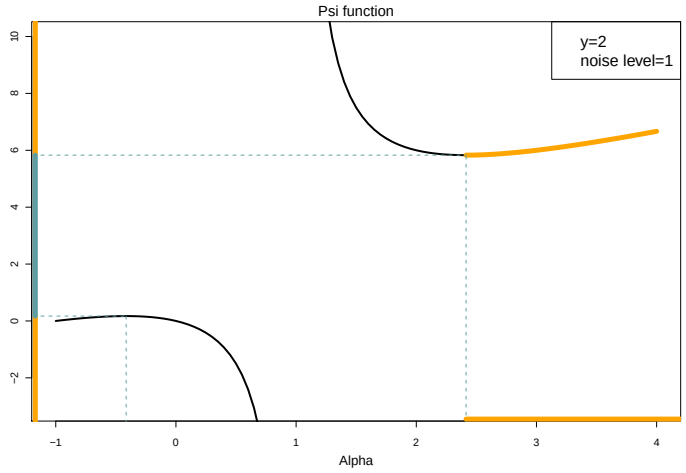


Figure 4.26: The curve illustrates $\psi_{y,H}$ with $y = 2$ and the PSD $H = \delta_1$. The yellow regions correspond to the α satisfying (a) and (b) in Lemma 4.3.1 and the blue region indicates the support of the MP distribution $F_{y,H}$.

Baik and Silverstein (2006) investigated the asymptotic behaviors of the sample eigenvalues corresponding to the spikes under Johnstone's spike model. They showed that sample spike eigenvalues converge almost surely to some functions of the corresponding population spike eigenvalues and the different functions are considered for the different types of spikes. Bai and Yao (2012) extended their results to a generalized spike model where the spectrum of a base set follows an arbitrary distribution. A generalized spike model has a population covariance matrix T_d that is of the form

$$T_d = \begin{pmatrix} \mathbf{\Lambda} & 0 \\ 0 & V_d \end{pmatrix}$$

where $\mathbf{\Lambda}$ is an $M \times M$ matrix whose eigenvalues are the population spikes, $\alpha_1 > \cdots > \alpha_K$ of respective multiplicity (n_k) with $\sum_{k=1}^{K} n_k = M$, and V_d is a $d' \times d'$ matrix whose eigenvalues are $\beta_{n,1} \geq \cdots \geq \beta_{n,d'}$. Under certain conditions on the true eigenvalues, they showed the almost sure convergence of a sample spike eigenvalue to some function of the corresponding population

eigenvalue as n and d both increase such that $\frac{d}{n} \to y$. Our method is motivated by a generalized spike model, which allows flexibility on non-spike population eigenvalues.

4.4 Methodology

In this section, we propose an algorithm to choose the number of spikes for a generalized spike model where the PSD is not constrained to be the point mass at σ^2, i.e. $H(t) = \delta_{\sigma^2}(t)$. The algorithm will be first described under the scenario where the PSD is known, which is unrealistic in most cases but easy to understand the basic idea. Next, we will introduce the main algorithm which is applicable when the PSD is unknown. Also, we propose a graphical tool to assess the distribution assumption for the PSD. In this section, we denote the PSD by $H(t;\boldsymbol{\theta})$ to emphasize the parameters that identify the $H(t)$ and the eigenvalues of S_n in decreasing order by $\hat{\lambda}_1, \cdots, \hat{\lambda}_{d \wedge n}$.

4.4.1 Estimation when the PSD is known

Let us first consider the case when the PSD $H(t;\boldsymbol{\theta})$ is known. Since we know the support of the corresponding LSD $F_{y,H}$ according to Lemma 4.3.1, a straightforward way to estimate the number of underlying spikes may be counting the number of eigenvalues above the upper boundary of the support. However, this procedure may slightly overestimate the true number of spikes because it ignores the variation in the largest noise eigenvalue. In the point mass PSD, for example, the resulting upper boundary for the LSD is $b_y = \sigma^2(1+\sqrt{y})^2$ whereas the largest eigenvalue is known to converge in distribution to the Tracy-Widom law whose support contains b_y, giving a non-ignorable probability of the largest eigenvalue being greater than b_y (Johnstone, 2001; Ma et al., 2013). To reflect such variation of the largest noise eigenvalue, in this particular example, the approximate Tracy-Widom quantile can replace the upper boundary b_y and this provides a more precise estimation (Kritchman and Nadler, 2008).

However, except for this simplest case ($H(t;\boldsymbol{\theta}) = \delta_{\sigma^2}$), we do not know the distribution of the largest noise eigenvalue. Thus, we approximate the distribution by simulation under a sufficiently large number of independent replications and then we take a certain quantile (e.g. 99th percentile)

as a threshold above which sample eigenvalues are considered as spikes. The simulation procedure to get a level α threshold, s_α, is described in Algorithm 1.

Algorithm 1.

1. For $b = 1, 2, \cdots, B$ with a sufficiently large number $B \in \mathbb{N}$,
   (a) Generate d random variables $\{\beta_1^{(b)}, \cdots, \beta_d^{(b)}\}$ from $H(t;\boldsymbol{\theta})$ and get a $d \times d$ diagonal matrix $T_{(b)}$ by taking $\text{diag}(T_{(b)}) = \{\beta_1^{(b)}, \cdots, \beta_d^{(b)}\}$;
   (b) Generate a $d \times n$ matrix $Z_{(b)}$ whose entries are independent variables from $N(0,1)$;
   (c) Get the largest eigenvalue $\hat{\lambda}_1^{(b)}$ of $\frac{1}{n} T_{(b)}^{1/2} Z_{(b)} Z_{(b)}^T T_{(b)}^{1/2}$.
2. Obtain $(1-\alpha)$-quantile s_α based on the set $\{\hat{\lambda}_1^{(1)}, \cdots, \hat{\lambda}_1^{(B)}\}$.

4.4.2 Estimation when the PSD is unknown

In practice, the PSD $H(t;\boldsymbol{\theta})$ is unknown and should be estimated as well. Here we consider the scenario that the type of distribution is known but with unknown parameters, e.g. the case that the PSD is δ_{σ^2} with an unknown parameter σ^2. The main algorithm is based on a sequence of nested hypothesis tests:

$$H_0^{(m)} : K \leq m-1 \quad vs. \quad H_1^{(m)} : K \geq m$$

where K is the true number of spikes and $m = 1, \cdots, d \wedge n$. At the m-th stage, we estimate the PSD parameters assuming $\hat{\lambda}_m, \cdots, \hat{\lambda}_{d\wedge n}$ to be non-spikes and test whether or not the $\hat{\lambda}_m$ is from a spike based on the approximate distribution of the largest noise eigenvalue obtained by Algorithm 1 with the estimated parameters. If the null is rejected, i.e. there are at least m spikes, then we proceed to the $(m+1)$-th hypothesis test after excluding $\hat{\lambda}_m$. Otherwise, we stop the procedure and conclude that there are at most $m-1$ spikes. Note that when we consider a point mass PSD and a theoretically obtained threshold from the Tracy-Widom law instead of the simulated one, this procedure plays

the same role as the method proposed by Kritchman and Nadler (Kritchman and Nadler, 2008) with a carefully estimated noise variance.

Estimation of unknown parameters of the PSD. An interesting topic in the random matrix theory is the estimation of the PSD parameters. (Li et al., 2013; Bai et al., 2010). Although we employ the Bai's method of moments in this paper because it is intuitive and is to implement, other PSD parameter estimation methods can be used as well. Bai et al. (2010) proposed a method to estimate the unknown parameters based on the relationship between the moments of the PSD $H(t;\boldsymbol{\theta})$ and the moments of the MP distribution $F_{y,H}$. The following lemma (Nica and Speicher, 2006) describes the relationship.

Lemma 4.4.1. *(Nica & Speicher, 2006) The moments* $\alpha_j = \int x^j dF_{y,H}(x),\ j \geq 1$ *of the LSD* $F_{y,H}$ *are linked to the moments* $\beta_j = \int t^j dH(t)$ *of the PSD* H *by*

$$\alpha_j = y^{-1} \sum y^{i_1+i_2+\cdots+i_j} (\beta_1)^{i_1} (\beta_2)^{i_2} \cdots (\beta_j)^{i_j} \phi^{(j)}_{i_1,i_2,\cdots,i_j} \tag{4.4.1}$$

where the sum runs over the following partitions of j*:*

$$(i_1,\cdots,i_j) : j = i_1 + 2i_2 + \cdots + ji_j, \quad i_l \in \mathbb{N},$$

and $\phi^{(j)}_{i_1,i_2,\cdots,i_j}$ *is the multinomial coefficient*

$$\phi^{(j)}_{i_1,i_2,\cdots,i_j} = \frac{j!}{i_1!i_2!\cdots i_j!(j+1-(i_1+i_2+\cdots+i_j))!}.$$

We denote the estimate of $\boldsymbol{\theta}$ obtained from the method of moments by $\hat{\boldsymbol{\theta}}$. Note that the proposed moment estimator has been proved to be strongly consistent and asymptotically normal (Bai et al., 2010).

We employ this method of moments to estimate the PSD parameters $\boldsymbol{\theta}$ and the main algorithm to estimate the number of spikes K is described as follows:

Main algorithm.

From $m = 1$, iterate the following procedure until it stops.

1. Based on $\{\hat{\lambda}_m, \hat{\lambda}_{m+1}, \cdots, \hat{\lambda}_{d\wedge n}\}$, obtain the PSD parameters $\hat{\boldsymbol{\theta}}^{(m)}$.

2. Applying Algorithm 1 with the estimated $\hat{\boldsymbol{\theta}}^{(m)}$, obtain the $(1-\alpha)$-quantile $s_\alpha^{(m)}$ with a predetermined level α.

3. If $\hat{\lambda}_m > s_\alpha^{(m)}$, reject $H_0^{(m)}$ and go to the step 1 replacing m by $m+1$. Otherwise, we do not reject $H_0^{(m)}$, the estimate of the number of spikes is $\hat{K} = m-1$, and stop here.

4.4.3 PSD diagnostics

The proposed method for determining the number of spikes depends on the correct specification of the underlying PSD. Under a misspecified PSD, the estimated number would be unreliable. Here, we develop a graphical tool for the diagnostic to check if the assumed PSD is correctly specified.

Let $\underline{s}$ be the true companion Stieltjes transform as defined in Section 4.2 and let the sample companion Stieltjes transform, denoted by $\underline{s}_n$, be

$$\underline{s}_n(u) = -\frac{1-d/n}{u} + \frac{1}{n}\sum_{i=1}^{d}\frac{1}{\hat{\lambda}_i - u}$$

for $u \in \Gamma_{F_n}^c$ where F_n denotes the ESD from the data and Γ_F denotes the support of the distribution F. Then, Li et al. (2013) shows that, under certain conditions, (a) $\underline{s}_n(u)$ converges to $\underline{s}(u)$ for any $u \in \text{int}(\liminf_{n\to\infty}\Gamma_{F_n}^c \backslash 0)$; (b) the $\psi_{y,H}(\alpha)$ defined in (4.3.1) uniquely determines the PSD H. Combining (a) and (b) gives us an important fact that if the PSD of the data is truely H, then the $\hat{\psi}_n(\alpha) = \underline{s}_n^{-1}(-1/\alpha)$ will be close to the true $\psi_{y,H}(\alpha)$ for $\alpha \in A$ where $A = \{\alpha : \alpha \notin \Gamma_H, \alpha \neq 0, \text{ and } \psi'_{y,H}(\alpha) > 0\}$.

Based on this fact, we propose *psi envelopes* which carefully study the difference between $\psi_{y,H}$ and $\hat{\psi}_n$ by constructing simulated envelopes, $\hat{\psi}_n^q$ for $q = 1, \cdots, Q$. The basic idea is the same for Q-Q envelopes (Hannig et al., 2001). The envelope $\hat{\psi}_n^q$ is from Q independent replications obtained

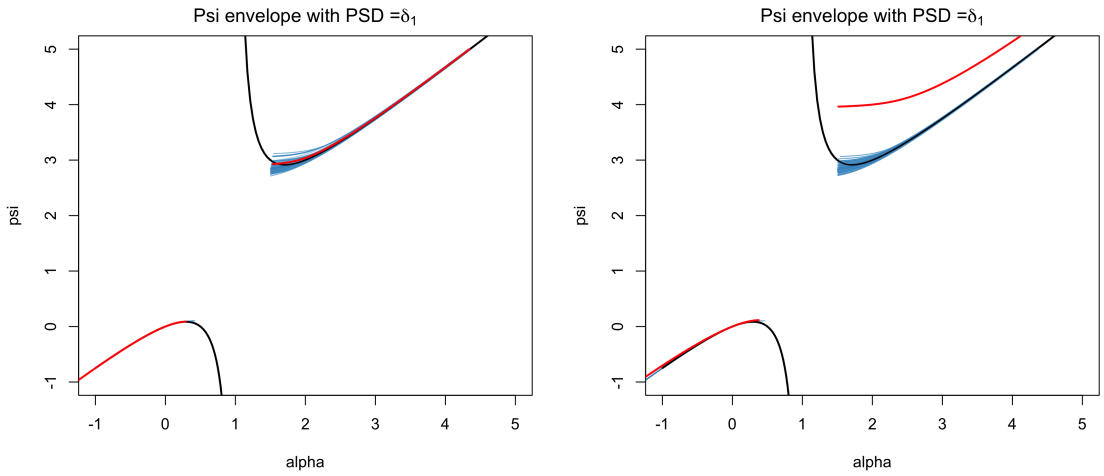


Figure 4.27: Examples of the psi envelope for assessment of the point mass PSD $H = \delta_1$. The left panel shows the case where the sample eigenvalues are truly from the assumed H whereas the right panel shows the case where the sample eigenvalues are from the different point mass PSD $H^{'} = \delta_{1.2}$. The figure demonstrates the sensitivity of the psi envelope diagnostic.

as in Algorithm 1. Then, the psi envelope enables assessment of the PSD assumption by checking whether or not the $\hat{\psi}_n$ is covered by the envelope. Two examples of the psi envelope are shown in Figure 4.27, which are checking whether or not two different sets of eigenvalues are from the PSD $H = \delta_1$. For the left plot, the sample eigenvalues truly from the $H = \delta_1$ are considered, and the sample eigenvalues from a different PSD $H^{'} = \delta_{1.2}$ for the right plot. In each plot, the function ψ_{y,δ_1} is shown in black with 100 blue envelope and the estimated function $\hat{\psi}_n$ in red. The $\hat{\psi}_n$ which is based on the eigenvalues truly from the H is well covered by the envelope whereas the $\hat{\psi}_n$ from $H^{'}$ shows clear deviation from the envelope. As this example shows, the psi envelope provides a useful graphical tool for the PSD diagnostic.

4.5 Real data analysis

In this section, we apply the proposed methods to our main example, RNA-seq gene expression data. Let us first briefly describe the data structure. A collection of 522 head and neck squamous carcinoma RNA-seq observations were obtained from the TCGA Research Network as described earlier. In this chapter, we study the base-level gene expression for each of several important cancer related genes. For each gene, the RNA-seq read-depths are measured along the length of the

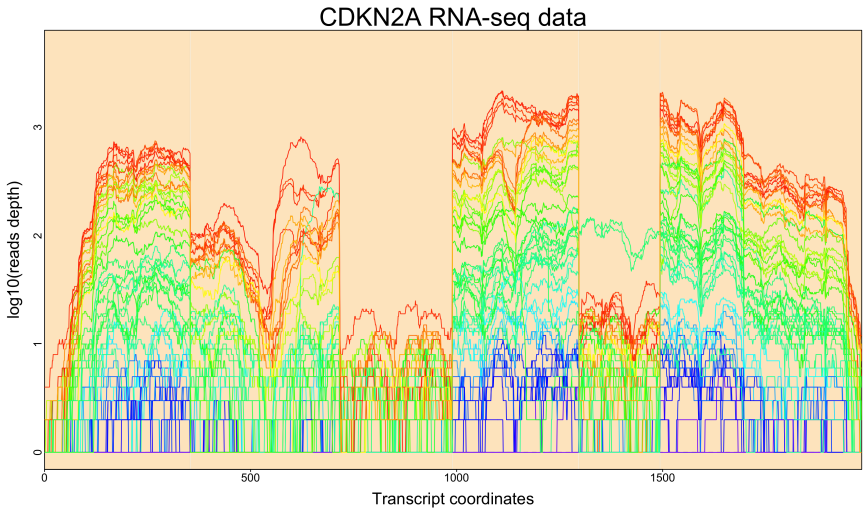


Figure 4.28: A set of RNA-seq observations for the gene CDKN2A are plotted on the log scale.

transcript. The resulting data structure is an expression count matrix $\boldsymbol{R} = \{r_{ij}\}_{1\le i\le d, 1\le j\le n}$ where r_{ij} is the read count at the ith position for the jth patient, d is the length of the transcript at the gene being studied, and n is the sample size. Since RNA-seq counts data show unstable variations within and between observations, the shifted logarithm transformation was taken to stabilize such heterogeneity, i.e. $\mathbf{X} = \{x_{ij}\}$ where $x_{ij} = \log_{10}(r_{ij}+1)$. For each gene, our analysis is based on the resulting matrix X and the example of the gene CDKN2A is described in Figure 4.28.

Since RNA-seq data typically have 1,000-20,000 dimensions while there are only hundreds of samples available, PCA is a very useful tool to reduce a huge dimension size and visualize the underlying relationship between samples or variables. In many cases, RNA-seq data are analyzed for hundreds or thousands of genes for various purposes such as discovery of key genes, detection of interesting genetic events, or identification of novel clusters, so that an automatic choice of the number of PCs at each gene is useful. To our best knowledge, however, there is no existing method to select the number of PCs that is applicable to RNA-seq data. In Section 4.5.1, we show why the existing methods are not appropriate for the RNA-seq data and suggest a new PSD model which is more suitable. And we compare our results with some existing methods for several important genes in Section 4.5.2.

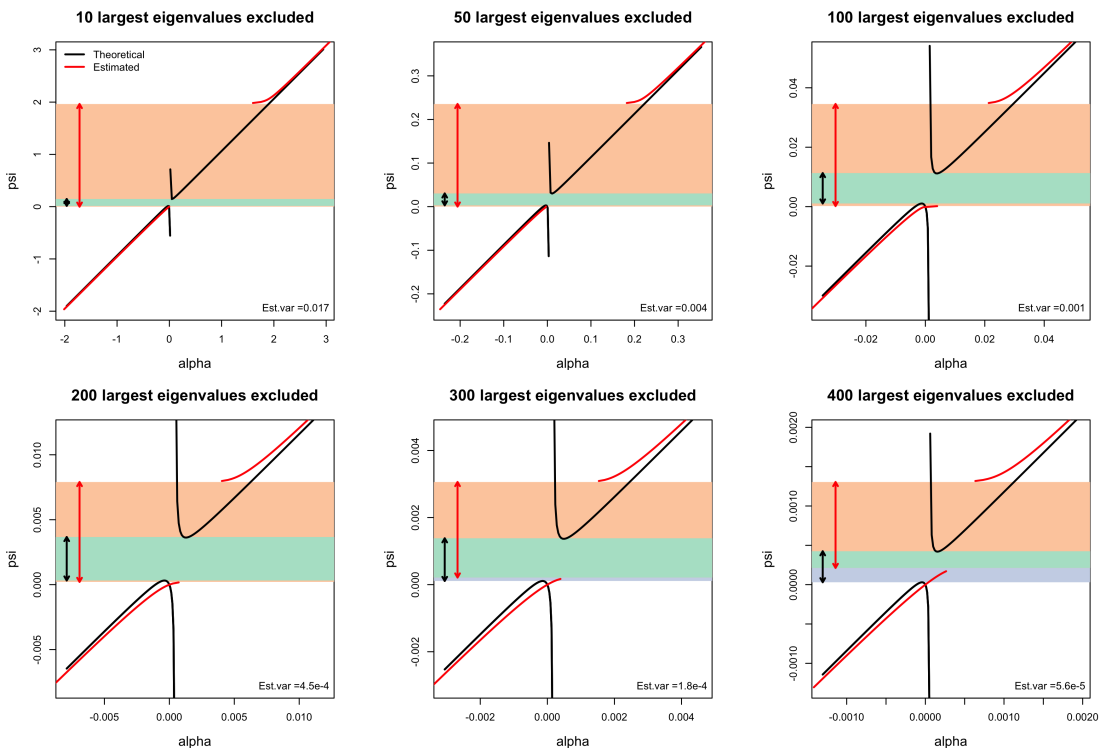


Figure 4.29: Comparison between the theoretical functions $\psi_{y_n,\delta_{\hat{\sigma}^2}}$ (black) and the estimated functions $\hat{\psi}_n$ (red) for the RNA-seq data at the gene CDKN2A. In each plot, the estimates $\hat{\sigma}^2$ and $\hat{\psi}_n$ are obtained based on the sample eigenvalues except for the 10, 50, 100, 200, 300, 400 largest eigenvalues. The blue rectangular area with the black arrows indicate the theoretically expected supports of the LSDs under the point mass PSD. The orange rectangular area with the red arrows indicate the supports of the ESDs from the data. The green rectangular area indicates the intersection of the blue and orange regions. The estimated noise variances are provided at the bottom of each plot. The Figure shows a point mass PSD provides a very poor fit to the data.

4.5.1 The proposed noise model

Figure 4.29 shows graphics which demonstrate that the point mass PSD for a noise distribution of RNA-seq data is clearly not appropriate. Here, the gene CDKN2A is considered with $d = 1978$, $n = 522$. In each plot, the theoretical function $\psi_{y_n,\delta_{\hat{\sigma}^2}}$ (black) with the estimated noise level $\hat{\sigma}^2$ and $y_n = \frac{d}{n}$ is compared with the estimated function $\hat{\psi}_n$ from data (red) after excluding the 10, 50, 100, 200, 300, 400 largest eigenvalues sequentially. The red arrow represents the support of the empirical spectral distribution from the data, that is, from the smallest to the largest sample eigenvalue. Correspondingly, the support of the theoretical M-P distribution extended to the 99th

percentile of the distribution of the largest eigenvalue, known as the Tracy-Widom law, is represented by the blue arrow (Ma et al., 2012). This enables more accurate comparisons of the two distributions by taking into account the variation in the maximum eigenvalue. Although the theoretical and data-driven functions, $\psi_{y_n,\delta_{\hat{\sigma}^2}}$ and $\hat{\psi}_n$, as well as the corresponding supports become comparable as more eigenvalues are kicked out, it is clear that they strongly disagree even when almost all eigenvalues are elliminated. This demonstrates that the noise eigenvalues from the data do not follow the classical M-P distribution, which motivates our improved PSD models.

Under the questionable white-noise assumption, we easily expect that too many PCs would be determined to be significant because of the extreme skewness of the ESD. The severe positive skewness makes the kth sample eigenvalue, the maximum of the remaining eigenvalues at the kth stage, be much greater than the theoretically possible maximum eigenvalue, resulting in the rejection of the hypothesis that the kth eigenvalue is from noise. In almost all cases, we observed that this was indeed true as described in Table 4.6. From the perspective of the dimension reduction, however, such high numbers of PCs may not be helpful especially for downstream statistical analyses that mostly require a much smaller dimension size.

Accordingly, the PSD that has a point mass at the noise variance, i.e. $H(t) = \delta_{\sigma^2}$, is not a suitable distribution for RNA-seq data whose eigenvalues are extremely right-skewed. To capture the extreme positive skewness, we suggest a right-skewed PSD, particularly the Gamma distribution truncated at some upper quantile. Other right-skewed distributions may be used but the Gamma distribution fit the best from our experience. Also, we can control the degree of positive skewness by adjusting the truncation quantile. Our method is sensitive to this choice of the quantile and precise specification is a topic for future research. Truncation determined by the upper 0.995 quantile gave reasonable values so we use it for the rest of the paper. Figure 4.30 shows the psi envelopes introduced in Section 4.4.3 for assessing the gamma assumption after kicking out a few largest sample eigenvalues. As in Figure 4.29, Figure 4.30 shows the estimated functions $\hat{\psi}_n$ in red sequentially removing a first few eigenvalues with the red vertical arrows for the supports of the ESDs. The black curves represent the theoretical functions $\psi_{y_n,H(t;\hat{\boldsymbol{\theta}})}$ with the black arrows

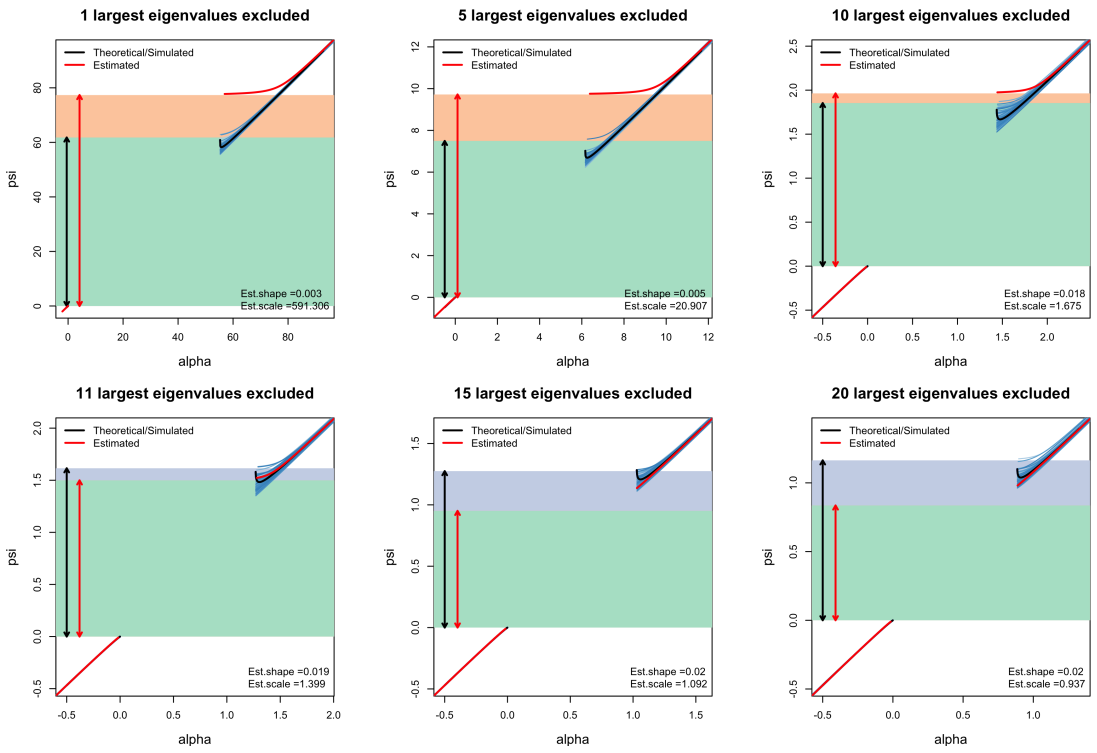


Figure 4.30: The psi envelopes for assessing the Gamma PSD for the gene CDKN2A. In each plot, the remaining eigenvalues excluding the 1, 5, 10, 11, 15, 20 largest eigenvalues are compared with the estimated Gamma PSD. The blue, orange, and green rectangular regions and red and black arrows are determined similarly as in Figure 4.29. The PSD parameters are provided at the bottom of each plot. The figure indicates 11 large non-noise eigenvalues.

indicating the approximate supports of the corresponding LSDs. As more eigenvalues are excluded up to 11, the sample psi curve gets closer to the psi envelopes, which supports that the remaining sample eigenvalues roughly follow the LSD based on the estimated PSD $H(t, \hat{\boldsymbol{\theta}})$. As we will see in Section 4.5.2, the critical value 11 is exactly the estimated number of spikes from the proposed method in Section 4.4.

4.5.2 Application to important genes

We compared our method (CM) with the methods proposed by Kritchman and Nadler (KN) and Passemier and Yao (PY) for the RNA-seq data at several important tumor-related genes: CDKN2A, TP53, FAT1, PTEN, CASP4, CHEK2, EGFR, and PIK3CA. Let us first briefly describe the two methods.

Method of Kritchman and Nadler (KN). Kritchman and Nadler (2008) developed an algorithm for rank determination based on the asymptotic distribution of the largest noise eigenvalue. Their algorithm performs sequential hypothesis tests on whether the largest eigenvalue at each step arises from a signal rather than from noise. The statistical procedure at each step involves estimating noise variance and setting a threshold based on the Tracy-Widom distribution where the largest noise eigenvalue follows (Johnstone, 2001). Their algorithm has been considered as a good benchmark for judging performance of other methods for determining the number of components in many papers.

Method of Passemier and Yao (PY). Passemier and Yao (2012, 2014) proposed a method for estimating the number of spikes under the case where there are possibly equal spikes. Based on the different asymptotic behaviors of spike and non-spike sample eigenvalues, they determine a threshold for the successive spacings of the ordered sample eigenvalues. Because larger spacings are expected for spike sample eigenvalues than for noise eigenvalues, one may separate spikes and non-spike eigenvalues based on an appropriately determined threshold for the spacing. This method is very intuitive in the sense that the proposed procedure is somehow similar to the naive procedure based on scree plots with a more reasonable separation based on random matrix theory. To avoid false determination due to ties of spike eigenvalues, they also proposed a more robust estimator by using consecutive two or more spacings that should be larger than a threshold at the same time to be considered as spikes.

The estimated number of spikes for each gene from the three methods are summarized in Table 4.6. As expected, both KN and PY methods determine a huge number of spikes and, in particular, the KN results indicate that all PCs are spikes for these eight genes. When the all PCs are declared as spikes, the noise variance cannot be estimated because there is no noise eigenvalue any more, as indicated by NA in Table 4.6. On the other hand, our proposed method provides biologically reasonable and practical number of spikes. Although we provide the results for the eight chosen genes, we have observed that this is indeed true for almost all genes. We believe that the proposed method can give valuable contribution to distinguish meaningful and important signals from noise

	d	CM ($\hat{\tau}$, $\hat{v}$)	KN ($\hat{\sigma}^2$)	PY ($\hat{\sigma}^2$)
CDKN2A	1978	11 (0.019, 1.399)	521 (NA)	197 (0.00046)
CHEK2	2595	5 (0.014, 2.029)	521 (NA)	208 (0.00034)
CASP4	2688	6 (0.011, 3.665)	521 (NA)	219 (0.0039)
PIK3CA	3686	8 (0.013, 1.310)	521 (NA)	521 (NA)
TP53	3876	12 (0.013, 1.505)	521 (NA)	521 (NA)
PTEN	5535	17 (0.011, 1.235)	521 (NA)	521 (NA)
EGFR	7965	13 (0.010, 2.434)	521 (NA)	521 (NA)
FAT1	15232	16 (0.008, 3.781)	521 (NA)	521 (NA)

Table 4.6: Estimates of the number of spikes from the proposed method (CM) and two existing methods (KN and PY). The lengths of transcripts (dimensions of the RNA-seq data) for eight genes are provided in the second column. For the CM method, the estimated shape and rate parameters ($\hat{\tau}$, $\hat{v}$) of the Gamma are provided and the estimated noise variance $\hat{\sigma}^2$ are also provided for the KN and PY methods. The table shows that the proposed CM method estimates reasonable numbers of spikes whereas the KN and PY methods provides far too large number of spikes. NA indicates that the noise variance is not available.

in RNA-seq data. Furthermore, the method can be also applied to other types of data set where a point mass PSD is not appropriate with a carefully chosen PSD.

BIBLIOGRAPHY

Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C. D., Annala, M., Aprikian, A., Armenia, J., Arora, A., et al. (2015). The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025.

Ahn, J., Lee, M. H., and Lee, J. A. (2018). Distance-based outlier detection for high dimension, low sample size data. *Journal of Applied Statistics*, pages 1–17.

Ahn, J., Marron, J. S., Muller, K. M., and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):760–766.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148.

Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H., and Marron, J. (2018). A survey of high dimension low sample size asymptotics. *Australian & New Zealand Journal of Statistics*, 60(1):4–19.

Bai, Z., Chen, J., and Yao, J. (2010). On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Australian & New Zealand Journal of Statistics*, 52(4):423–437.

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer.

Bai, Z. and Yao, J. (2012). On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177.

Bai, Z. D. (2008). Methodologies in spectral analysis of large dimensional random matrices, a review. In *Advances In Statistics*, pages 174–240. World Scientific.

Bai, Z.-D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Annals of probability*, pages 316–345.

Bai, Z. D. and Yin, Y. Q. (1988). Convergence to the semicircle law. *The Annals of Probability*, pages 863–875.

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294.

Bai, Z.-D., Yin, Y.-Q., et al. (1988). Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a wigner matrix. *The Annals of Probability*, 16(4):1729–1741.

Baik, J., Arous, G. B., Péché, S., et al. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408.

Barnett, V. and Lewis, T. (1974). *Outliers in statistical data*. Wiley.

Bartlett, M. S. (1954). A note on the multiplying factors for various χ 2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 296–298.

Benaych-Georges, F. and Nadakuditi, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521.

Besse, P. and de Falguerolles, A. (1993). Application of resampling methods to the choice of dimension in principal component analysis. In *Computer intensive methods in statistics*, pages 167–176. Springer.

Buermans, H. and Den Dunnen, J. (2014). Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1932–1941.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94.

Choi, Y., Taylor, J., and Tibshirani, R. (2014). Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *arXiv preprint arXiv:1410.8260*.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. (2016). A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13.

Dai, W. and Genton, M. G. (2016). Directional outlyingness for multivariate functional data. *arXiv preprint arXiv:1612.04615*.

Dang, X. and Serfling, R. (2010). Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Planning and Inference*, 140(1):198–213.

Davila, J. I., Fadra, N. M., Wang, X., McDonald, A. M., Nair, A. A., Barbara, R. C., Wu, X., Blommel, J. H., Jen, J., Rumilla, K. M., et al. (2016). Impact of rna degradation on fusion detection by rna-seq. *BMC genomics*, 17(1):814.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, pages 1803–1827.

Dvinge, H. and Bradley, R. (2015). Widespread intron retention diversifies most cancer transcriptomes. In *Genome Medicine*, pages 7–45.

Eswaran, J., Horvath, A., Godbole, S., Reddy, S. D., Mudvari, P., Ohshiro, K., Cyanam, D., Nair, S., Fuqua, S. A., Polyak, K., et al. (2013). Rna sequencing of cancer reveals novel splicing alterations. *Scientific reports*, 3:1689.

Fan, J. and Wang, W. (2015). Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *arXiv preprint arXiv:1502.04733*.

Feng, Q., Hannig, J., and Marron, J. (2016). A note on automatic data transformation. *Stat*, 5(1):82–87.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711.

Girshick, M. (1939). On the sampling theory of roots of determinantal equations. *The Annals of Mathematical Statistics*, 10(3):203–224.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons.

Hannig, J., Marron, J., and Riedi, R. (2001). Zooming statistics: Inference across scales. *Journal of the Korean Statistical Society*, 30(2):327–345.

Harding, M. C., Jorgenson, D., King, G., Linton, O., Lorenzoni, G., Mazur, B., Panageas, S., and Patel, K. (2007). Structural estimation of high-dimensional factor models.

Hawkins, D. M. (1974). The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, 69(346):340–344.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

Hicks, S. C. and Irizarry, R. A. (2014). When to use quantile normalization? *BioRxiv*, page 012203.

Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Jung, H., Lee, D., Lee, J., Park, D., Kim, Y. J., Park, W.-Y., Hong, D., Park, P., and Lee., E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. pages 1242–1248.

Jung, S. and Marron, J. S. (2009). Pca consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130.

Jung, S., Sen, A., and Marron, J. S. (2012). Boundary behavior in high dimension, low sample size asymptotics of pca. *Journal of Multivariate Analysis*, 109:190–203.

Kamber, M. and Han, J. (2001). *Data mining: Concepts and techniques*, volume 2. Morgan Kaufmann Publishers San Francisco.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345.

Kimes, P. K., Cabanski, C. R., Wilkerson, M. D., Zhao, N., Johnson, A. R., Perou, C. M., Makowski, L., Maher, C. A., Liu, Y., Marron, J. S., et al. (2014). Sigfuge: single gene clustering of rna-seq reveals differential isoform usage among cancer samples. *Nucleic acids research*, 42(14):e113–e113.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38.

Kritchman, S. and Nadler, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19–32.

Kritchman, S. and Nadler, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941.

Krzanowski, W. J. and Kline, P. (1995). Cross-validation for choosing the number of important components in principal component analysis. *Multivariate Behavioral Research*, 30(2):149–165.

Lawley, D. (1953). A modified method of estimation in factor analysis and some large sample results. In *Uppsala symposium on psychological factor analysis*, volume 17, pages 35–42. Taylor & Francis.

Lawley, D. (1956). Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43(1/2):128–136.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Li, W., Chen, J., Qin, Y., Bai, Z., and Yao, J. (2013). Estimation of the population spectral distribution from a large dimensional sample covariance matrix. *Journal of Statistical Planning and Inference*, 143(11):1887–1897.

Li, W. and Yao, J. (2018). On structure testing for component covariance matrices of a high dimensional mixture. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):293–318.

Lim, A. M., Do, H., Young, R. J., Wong, S. Q., Angel, C., Collins, M., Takano, E. A., Corry, J., Wiesenfeld, D., Kleid, S., et al. (2014). Differential mechanisms of cdkn2a (p16) alteration in oral tongue squamous cell carcinomas and correlation with patient outcome. *International journal of cancer*, 135(4):887–895.

Liu, R. Y. (1992). Data depth and multivariate rank tests. *L1-statistical analysis and related methods*, pages 279–294.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

Loyo, M., Li, R. J., Bettegowda, C., Pickering, C. R., Frederick, M. J., Myers, J. N., and Agrawal, N. (2013). Lessons learned from next-generation sequencing in head and neck cancer. *Head & neck*, 35(3):454–463.

Ma, Z. et al. (2012). Accuracy of the tracy–widom limits for the extreme eigenvalues in white wishart matrices. *Bernoulli*, 18(1):322–359.

Ma, Z. et al. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457.

Mountzios, G., Rampias, T., and Psyrri, A. (2014). The mutational spectrum of squamous-cell carcinoma of the head and neck: targetable genetic events and clinical impact. *Annals of oncology*, 25(10):1889–1900.

Nadakuditi, R. R. (2014). Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60(5):3002–3018.

Nadler, B. et al. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817.

Network, C. G. A. R. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525.

Network, C. G. A. R. et al. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543.

Nica, A. and Speicher, R. (2006). *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press.

Olivier, M., Hollstein, M., and Hainaut, P. (2009). Tp53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology*, page a001008.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Opitz, L., Salinas-Riester, G., Grade, M., Jung, K., Jo, P., Emons, G., Ghadimi, B. M., Beißbarth, T., and Gaedcke, J. (2010). Impact of rna degradation on gene expression profiling. *BMC medical genomics*, 3(1):36.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From rna-seq reads to differential expression results. *Genome biology*, 11(12):220.

Ozsolak, F. and Milos, P. M. (2011). Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413.

Passemier, D. and Yao, J. (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal Of Multivariate Analysis*, 127:173–183.

Passemier, D. and Yao, J.-F. (2012). On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory and Applications*, 1(01):1150002.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

Paul, D. and Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29.

Paul, D. and Johnstone, I. M. (2012). Augmented sparse principal component analysis for high dimensional data. *arXiv preprint arXiv:1202.1242*.

Reddy, R. (2015). A comparison of methods: normalizing high-throughput rna sequencing data. *bioRxiv*, page 026062.

Ro, K., Zou, C., Wang, Z., and Yin, G. (2015). Outlier detection for high-dimensional data. *Biometrika*, 102(3):589–599.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25.

Romero, I. G., Pai, A. A., Tung, J., and Gilad, Y. (2014). Rna-seq: impact of rna degradation on transcript quantification. *BMC biology*, 12(1):42.

Rousseeuw, P. J., Raymaekers, J., and Hubert, M. (2016). A measure of directional outlyingness with applications to image data and video. *arXiv preprint arXiv:1608.05012*.

Shabalin, A. A. and Nobel, A. B. (2013). Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118:67–76.

Shen, D., Shen, H., and Marron, J. (2016). A general framework for consistency of principal component analysis. *Journal of Machine Learning Research*, 17(150):1–34.

Shen, D., Shen, H., and Marron, J. S. (2012). A general framework for consistency of principal component analysis. *arXiv preprint arXiv:1211.2671*.

Shen, D., Shen, H., and Marron, J. S. (2013). Consistency of sparse pca in high dimension, low sample size contexts. *Journal of Multivariate Analysis*, 115:317–333.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339.

Silverstein, J. W. and Choi, S.-I. (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309.

Stahel, W. A. (1981). *Breakdown of covariance estimators*. Fachgruppe für Statistik, Eidgenössische Techn. Hochsch.

Stransky, N., Egloff, A. M., Tward, A. D., Kostic, A. D., Cibulskis, K., Sivachenko, A., Kryukov, G. V., Lawrence, M. S., Sougnez, C., McKenna, A., et al. (2011). The mutational landscape of head and neck squamous cell carcinoma. *Science*, 333(6046):1157–1160.

Taylor, J., Loftus, J., and Tibshirani, R. (2013). Tests in adaptive regression via the kac-rice formula. *arXiv preprint arXiv:1308.3020*.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426.

Wang, Q., Yao, J., et al. (2013). On the sphericity test with large-dimensional observations. *Electronic Journal of Statistics*, 7:2164–2192.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57.

Wax, M. and Kailath, T. (1985). Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):387–392.

Wong, J. J.-L., Au, A. Y. M., Ritchie, W., and Rasko, J. E. J. (2016). Intron retention in mrna: No longer nonsense. pages 41–49.

Wongsawat, Y., Rao, K., and Oraintara, S. (2005). Multichannel svd-based image de-noising. In *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, pages 5990–5993. IEEE.

Yao, J., Bai, Z., and Zheng, S. (2015). *Large sample covariance matrices and high-dimensional data analysis*. Number 39. Cambridge University Press.

Zhang, Q., Li, H., Jin, H., Tan, H., Zhang, J., and Sheng, S. (2014). The global landscape of intron retentions in lung adenocarcinoma. In *BMC Medical Genomics*, pages 7–15.

Zhao, L., Krishnaiah, P. R., and Bai, Z. (1986). On detection of the number of signals in presence of white noise. *Journal of multivariate analysis*, 20(1):1–25.

Zhou, Y.-H. and Marron, J. (2016). Visualization of robust l1pca. *Stat*, 5(1):173–184.

Zuo, Y. (2003). Projection-based depth functions and associated medians. *Annals of Statistics*, pages 1460–1490.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of statistics*, pages 461–482.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Górczak, K., Klamecka, K., Figlerowicz, M., and Siatkowski, I. (2015). The impact of normalization methods on rna-seq data analysis. *BioMed research international*, 2015.