STATISTICAL LEARNING OF INTEGRATIVE ANALYSIS

Meilei Jiang

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

Jan Hannig

Yufeng Liu

Perry Haaland

Nicolas Fraiman

ABSTRACT

MEILEI JIANG: Statistical Learning Of Integrative Analysis
(Under the direction of J. S. Marron and Jan Hannig)

Integrative analysis is of great interest in modern scientific research. This dissertation mainly focuses on developing new statistical methods for integrative analysis.

We first discuss a clustering analysis of a microbiome dataset in combination with phylogenetic information. Discovering disease related pneumotypes of the infected lower lung is difficult because the lower lung typically has few species of microbes and there is a low level of overlap from patient-to-patient, which makes it hard to calculate reliable distances between patients. We address this challenge by incorporating information from phylogenetic relationships, which results in improved clustering. When applied to an existing dataset, the method produces statistically distinct, easily described pneumotypes, which are better than those from standard approaches.

In the second part, we discuss an integrative analysis of disparate data blocks measured on a common set of experimental subjects. We introduce Angle-Based Joint and Individual Variation Explained (AJIVE) capturing both joint and individual variation within each data block. This is a major improvement over earlier approaches to this challenge in terms of a new conceptual understanding, much better adaption to data heterogeneity and a fast linear algebra computation. Detailed comparison between AJIVE and competitors is discussed using a particular optimization view point.

In the third part, we introduce a new perturbation framework, which estimates the angle between an arbitrary given direction and the underlying signal spaces. We also propose an efficient data-driven bootstrap procedure to compute this angle. While the Wedin bound in the AJIVE is "subspace oriented" and uniform for both row space and column space, this angle is "direction oriented" and specially adaptive to give improved inference in the row space.

To my beloved parents

ACKNOWLEDGEMENTS

This dissertation would have not been completed without the help, inspiration, and encouragement from many individuals during the last five years of my PhD study.

Foremost, I would like to express my sincere gratitude to my advisors, Professor J. S. Marron and Professor Jan Hannig, for their guidance and support. Their patience, inclusive mindset, immense knowledge and academic enthusiasm make me have a great experience of PhD life. Not only I gain many knowledge, but also I learn the way to be a great scholar as well as a good person from them. It is a great pleasure and luck for me to have them as my advisors.

Besides my advisors, I am grateful to Doctor Perry Haaland, my boss at Becton Dickinson Technologies. Part of my dissertation is accomplished when I was an intern at there. I greatly appreciate the support and guidance from him. My sincere thanks also go to Professor Yufeng Liu, Professor Shankar Bhamidi, Professor Kai Zhang and Professor Yin Xia for their encouragement and great help during my PhD study. In addition, I also want to express my thanks to the committee member Professor Nicolas Frainman for reading my dissertation and providing useful comments.

Special gratitude is extended to Rui Chen from the Economics Department of Duke University. Her encouragement and accompany help me overcome many challenges in my pursuit of a doctoral degree. Last but not the least, I want to thank my parents, Jianguo Jiang and Congju Liu, for their love on me. This dissertation is also an accomplishment for them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

In the past decade, we are experiencing the era of "big data" and "information explosion". Recently many datasets are large-scale not only in the sense of large-sample size and high dimension, but also in the sense of being multi-source. It is a ubiquitous challenge as well as opportunity for modern scientific research to integrate the information from disparate data blocks measured on a common set of experimental subjects. This thesis aims at addressing this challenge by developing new statistical learning methods for integrative analysis.

Chapter 2 discusses a clustering analysis of a microbiome dataset which incorporates phylogenetic information. Modern genomic methods have led to a dramatic increase in the ability to study a wide variety of microbiomes. In particular, this has enabled precise quantification of various bacterial species present in a range of different sample types, resulting in many statistical challenges. For example, discovering disease related pneumotypes of the infected lower lung is difficult because the lower lung typically has few species of microbes and there is a low level of overlap from patient-to-patient. This type of sparsity presents a special challenge to standard analysis methods because it is hard to calculate reliable distances between patients. We address this challenge by using information from phylogenetic relationships, which results in improved clustering. In our application to a pneumonia dataset, the method produces statistically distinct, easily described pneumotypes. It is seen that our pneumotypes are better than those from standard approaches, using a SWISS score analysis. A simulation study explores the ways in which this new approach generally performs better than standard methods.

Chapter 3 discusses a novel integrative statistical learning framework, Angle-based Joint and Individual Variation Explained (AJIVE), which can simultaneously explore the joint and individual variation within each data block. This work provides major improvements over earlier approaches to this challenge in several ways. First, there is a new conceptual understanding. Section, much better adaption to data heterogeneity leads to much sharper statistical inference. Third, there is

now a fast linear algebra computation. Important mathematical contributions are the use of score subspaces as the principal descriptors of variation structure and the use of perturbation theory as the guide for variation segmentation. This leads to an exploratory data analysis method which is insensitive to the heterogeneity among data blocks and does not require separate normalization. An application to cancer data reveals different behaviors of each type of signal in characterizing tumor subtypes. An application to a mortality data set reveals interesting historical lessons.

Chapter 4 discusses the relationship between AJIVE and existing integrative methods from an optimization point of view. It shows that some popular methods, such as Singular Value Decomposition, Partial Least Square and Canonical Correlation Analysis, are not as well suited for the AJIVE task. We also find the connections between AJIVE and Flag Mean as well as Common Orthogonal Basis Extraction, which provide geometric and optimization interpretations for AJIVE.

Chapter 5 discusses a new perturbation analysis framework for a given direction. Perturbation analysis in AJIVE relies on the estimation of the Wedin's bound, which is not only very conservative for non-square matrix, but also "subspace-oriented" rather than "direction-oriented". Based on the random matrix theory, a novel perturbation analysis framework for a given direction has been proposed for addressing this challenge. A singular value shrinkage estimator is used to recover the signal matrix from the data matrix. An efficient bootstrapping procedure has been proposed to compute the direction specific perturbation angle. Simulation study shows that this computation algorithm is quite efficient and robust under different settings. Applying this perturbation analysis on the AJIVE directions, we can compute the reliability of each AJIVE direction on each data block.

CHAPTER 2

Finding Community Subtypes On Microbiome Dataset

2.1 Introduction

Modern genomic methods have led to a dramatic increase in the ability to study a wide variety of microbiomes. An important question is whether or not knowledge of the lower lung microbiome in patients may aid in the diagnosis, treatment, or prevention of pneumonia (Koenig and Truwit, 2006; Beck et al., 2012; Yamasaki et al., 2013; Dickson et al., 2014; Segal et al., 2014). This question is approached through determining whether or not there are community subtypes or pneumotypes. We address the problem of finding pneumotypes by analyzing data from a previously published study of intensive care unit patients (Bousbia et al., 2012). Most of the subjects in this dataset were diagnosed with one of several types of pneumonia. For this microbiome dataset, we focus on *Operational Taxonomic Units* (OTUs), a general biological term including species and genera. The matrix of the proportions of bacterial OTUs found in each subject's lung is called the *Relative Abundance* (RA) matrix. Li (2014) gave an overview of high-dimensional data analysis for microbiome data. As shown in Figure 2.1, in each row of this microbiome data only a few entries are non-zero, and the distributions of RA over OTUs among different subjects are very diverse. These features of this microbiome data make it challenging to apply standard analytic methods because of the difficulty of computing meaningful distances between subjects.

A phylogenetic tree (shown in Figure 2.3) constructed using PhyloPhlAn (Segata et al., 2013) represents relationships between OTUs based on evolutionary distance. Using standard Euclidean distance on the RA row vector shown in Figure 2.1, two subjects who do not share OTUs are far apart whether or not the OTUs are close in the phylogenetic tree. The method that we propose, *Tree Weighted Abundance* (TWA), combines RA with phylogenetic distances so that subjects with microbiomes that are close in the phylogenetic tree will also be close in distance. If the organization of the phylogenetic tree is closely related to the true clusters, then we show via a simulation study

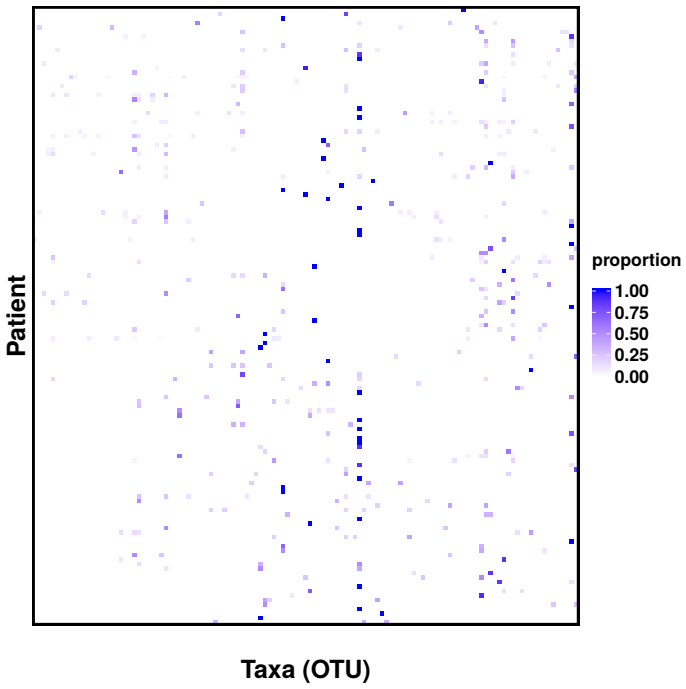


Figure 2.1: Operational taxonomic unit (OTU) relative abundances for the LRI microbiome. Rows represent subjects and columns represent OTUs. Each entry is the proportion of the observed read counts over the total observed read counts for a subject. Columns are ordered by the positions of OTUs in the phylogenetic tree. It shows many zero values, and great diversity of non-zero values, which motivates our approach.

that our approach is better than standard approaches that don't incorporate information from the phylogenetic tree. Related microbiome analyses which incorporate relative abundance with phylogenetic information include Matsen and Evans (2013); Zhao et al. (2015); Chen et al. (2016); Wu et al. (2016). A standard method that combines phylogenetic distances with RA is Principal Coordinate Analysis (PCoA) (Gower, 1966), which uses the abundance weighted Unifrac distance matrix (Lozupone and Knight, 2005).

We carefully study important drivers in a simulation study summarized in Figure 2.2 with the goal of understanding which method is better over a wide range of potential biological settings. This shows that TWA is better for a broad range of moderate tree information strength, while PCoA is the best for a limited set of conditions with very weak tree information (bottom of Figure 2.2). The black X shows the approximate location of this microbiome data, which shows the benefit of using TWA in this particular case. In a neighborhood of X, there are many settings where TWA is much better. The methods perform similarly for very strong tree information, i.e. the upper right region. A full description is in Section 2.4.

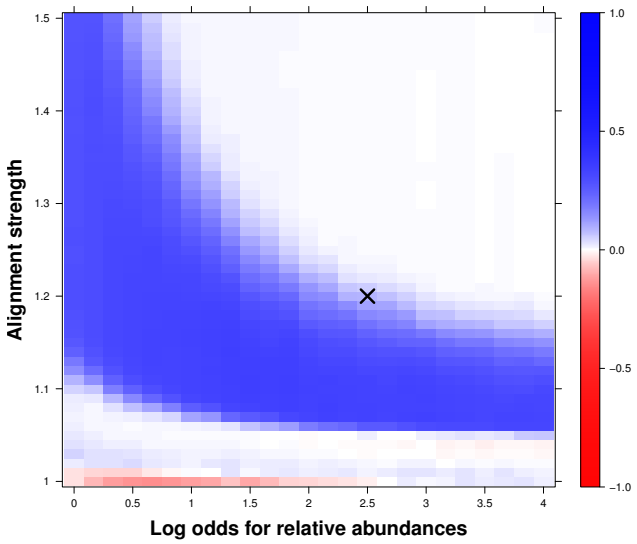


Figure 2.2: Heat map mutual information comparison of TWA and PCoA. Shades of blue show superior performance of TWA, with white used for essentially similar cases, and red when PCoA is better. The vertical axis represents the strength of alignment between the phylogenetic tree and clustering in the data. The horizontal axis represents how strongly relative abundances are concentrated in the parts of the phylogenetic tree related to the clusters. This shows a large range of biological settings where TWA is better.

The rest of the paper is organized as follows. Section 2.2 describes the motivation and methodology from the viewpoint of Object Oriented Data Analysis (Wang and Marron, 2007; Marron and

Alonso, 2014a). Section 2.2.1 describes standard statistical clustering algorithms for microbiome data. Section 2.2.2 discusses Object Oriented Data Analysis (Wang and Marron, 2007; Marron and Alonso, 2014a), which is a modern framework for approaching complex data challenges and which provides a viewpoint to address our novel approaches and other approaches to find community subtypes. Section 2.3 gives the empirical clustering analysis results on the microbiome dataset. Three distinct pneumotypes with interesting descriptions are identified by the approach we propose. Compared to standard statistical approaches, our approach gives more balanced and better statistically separated clustering results. Section 2.4 provides a simulation study to investigate whether and how the use of phylogenetic tree information improves the identification of clusters. Section 2.5 discusses the connection between the simulation data and the LRI microbiome data and concludes with final remarks.

2.2 Methods

2.2.1 Standard clustering algorithms for microbiome analysis

For microbiome analysis, two of the most commonly used algorithms are k-means clustering and mixture modeling.

k-means Clustering The k-means clustering method is a standard clustering approach for partitioning a dataset into k distinct clusters. Detailed description of its properties can be found in the book by Kaufman and Rousseeuw (1990). Choosing the best value of k is complex. For our analysis, the Calinski Harabasz Index (Calinski and Harabasz, 1974) is used to estimate the number of clusters.

Dirichlet Multinomial Mixture Modeling The Dirichlet Multinomial Mixture Model (DMM, Holmes et al. (2012)) is a probabilistic model of communities for count data. DMM is based on the multinomial distribution, having parameter vectors from a mixture of Dirichlet components as a prior. The model is fitted by Expectation Maximization, and a Laplace approximation is used to select the best model based on cross-validation. Bayes' theorem is used to calculate the posterior probability that a sample was generated from each of the classes. Ding and Schloss (2014) apply DMM modeling to analyze the Human Microbiome Project (NIH HMP Working Group, 2009) data.

In their analysis, the data objects are genera-level abundance vectors. Our data from Bousbia et al. (2012) included only RA values. DMM needs counts as input, so we made an ad hoc conversion to pseudocounts by multiplying all rows of the RA matrix by an arbitrary value, 100, and rounding to integers. If the original count data had been available, such an approach could be thought of as a normalization for sequencing depth.

2.2.2 Object Oriented Data Analysis viewpoint

Object Oriented Data Analysis provides a viewpoint for analyzing complex datasets, such as curves (Ramsay and Silverman, 2002, 2005), trees (Wang and Marron, 2007; La Rosa et al., 2012), shapes (Huckemann et al., 2010; Jung et al., 2012) and images (Sen et al., 2008; Lu et al., 2014). Instead of simply treating data objects as vectors in Euclidean space, an essential idea of Object Oriented Data Analysis is taking into account their intrinsic information. Object oriented terminology is useful for understanding these methods for finding community subtypes and how our new TWA approach relates to them.

RA Vectors As Data Objects The rows of the RA matrix can be directly represented as Euclidean vectors on a simplex. Straightforward analysis, say k-means clustering, can be applied to identify subtypes. However, for the microbiome data considered in this paper, both empirical results in Section 2.3 and simulation results in Section 2.4 show that treating the RA as vectors does not efficiently represent the subtype structure of subjects because of the high sparsity and diversity. In particular, only using RA vectors as data objects does not reflect the strong and useful relationships between the OTUs.

PCoA Scores As Data Objects PCoA scores computed from the weighted Unifrac distance matrix are data objects that also combine phylogenetic tree information with RA. The Unifrac distance matrix has been frequently used for microbiome analysis. For instance, Costello et al. (2009) applied this approach to study the bacterial community variation in human body habitats. The unique fraction metric, or Unifrac (Lozupone and Knight, 2005), measures the phylogenetic distance between two subjects (communities of OTUs) based on a phylogenetic tree. The distance is calculated as the fraction of total branch length that is unshared by the taxa in the two communities. In microbiome analysis, the branch length is typically weighted by the RA of taxa, which is called

weighted Unifrac distance. After computing distances for all pairs of subjects, we get a weighted Unifrac distance matrix. PCoA then uses multidimensional scaling (Borg and Groenen, 2005) to produce a coordinate matrix. Community subtypes are next identified by applying, for example, k-means clustering to the coordinate matrix produced by PCoA.

TWA Vectors As Data Objects We propose TWA as a novel data object that combines RA and phylogenetic tree information. TWA is the abundance weighted by the cophenetic distance matrix (Sokal and Rohlf, 1962) of OTUs in the phylogenetic tree, where cophenetic distance is the sum of the branch lengths traversed between two leaves in the phylogenetic tree. Denote the cophenetic distance matrix and RA matrix as $\mathbf{D}$ and $\mathbf{X}$ respectively. The matrix $\mathbf{X}$ has subjects as rows and OTUs as columns. Then, the Tree Weighted Abundance is calculated as

$$\text{TWA} = \mathbf{X} \cdot \mathbf{D}.$$

The rows of TWA can serve as the data objects for various clustering approaches, such as k-means clustering. This new data object, which incorporates phylogenetic tree information with RA, is in the spirit of weighted UniFrac distances. Weighted UniFrac, however, measures distances between subjects, whereas TWA can be best thought of as a transformation of RA. Because of the weighting scheme, subjects with OTUs that are phylogenetically related will be close in TWA space even when there is no overlap in OTUs, which is essentially important for data of the type shown in Figure 2.1.

2.3 Analysis of the real microbiome data

In this section we describe the application of TWA on the real microbiome data.

2.3.1 Data description and processing

The data analyzed were downloaded from the supplemental materials of Bousbia et al. (2012). In this study bronchoalveolar lavage (BAL) samples were collected from ICU patients who had mechanical ventilators. 185 subjects had been diagnosed with pneumonia. An additional 25 subjects without pneumonia served as controls. The data was obtained via amplification of 16S rRNA genes followed by cloning and sequencing. A specialized set of PCR primers were designed for this

study that allowed an unusually high level of resolution. Most of the OTUs, consequently, could be identified at the species-level using BLAST against GenBank. Relative abundance estimates were provided by the authors.

The initial data included 157 OTUs. We selected for analysis only taxonomic units with sufficient representation in GenBank to include in the calculation of a phylogenetic tree. In a few cases, we aggregated OTUs at the genus level in order to have sufficient reference genomes for accurate calculation of the tree. In total, 121 OTUs, mostly at the species level, had enough reference genomes available for phylogenetic analysis.

We restricted our analysis to subjects that had bacteria identified in their sample. Final data included 124 pneumonia subjects and 13 controls. We verified that the filtered data had about the same level of sparsity as the original. The mean number of OTU's per subject before filtering was 3.6 (s.e.=0.24) and after filtering was 3.2 (s.e.=0.20). We calculated the Bray-Curtis diversity (Bray and Curtis, 1957), a measure of the dissimilarity between a pair of subjects, of all pairs of subjects before and after filtering. For each subject we calculated the median value all pairs. The mean value across subjects was 1.0 (s.e.=0) for both the original and filtered datasets.

2.3.2 Computing the phylogenetic tree

Phylogenetic trees constructed for 16S rRNA sequencing data are often created using GreenGenes (DeSantis et al., 2006). This is the case for data downloaded from HMP. We based our analysis, however, on a phylogenetic tree constructed using PhyloPhlAn (Segata et al., 2013), because that also uses protein information. PhyloPhlAn assigns phylogeny based on the consideration of more than 400 proteins selected from nearly 4000 genomes. The desire to use a protein based phylogeny that relies on many genes instead of just one, led us to the filtering of the data described above. The PhyloPhlAn tree generated from the filtered data is shown in Figure 2.3. The construction of the tree is particularly important because the distances between tips of the tree are used in the subsequent analysis.

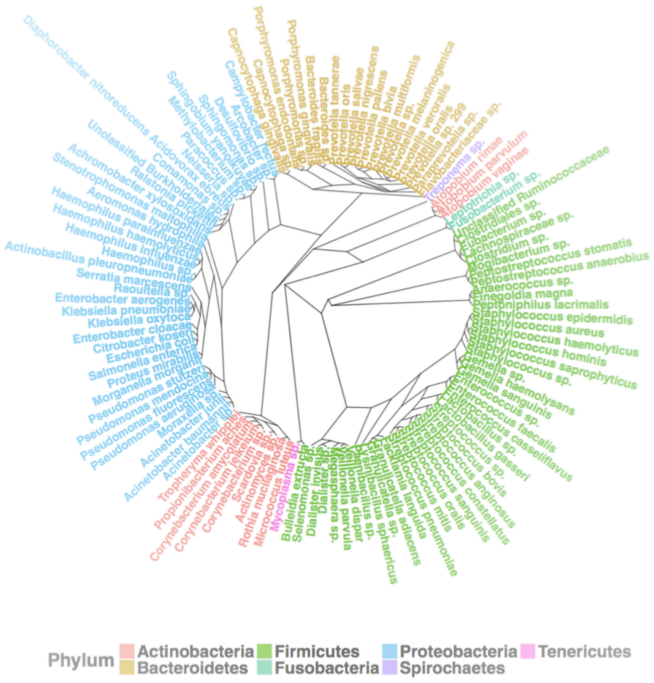


Figure 2.3: The unrooted phylogenetic tree computed by PhyloPhlAn. Tips are labeled with the OTU names. Tips are colored based on the phylum identification from NCBI. In general, this tree is consistent with the NCBI taxonomy.

2.3.3 k-means clustering on TWA

For our analysis, we used the R-package *fpc* (Hennig, 2015). This implements a standard k-means algorithm which initializes the means several times by using seeds that are randomly selected data points. Applying k-means clustering on TWA for values of k ranging from 2 to 10, as seen in Figure 2.4, the Calinski Harabasz index (Calinski and Harabasz, 1974) was highest for three clusters so all subsequent analysis will be based on 3-means clustering.

In the TWA space two subjects are close together when their OTUs are close in phylogenetic space even if they have no OTUs in common. As a consequence, we expect subjects with OTUs from similar phyla to be grouped together. This can be seen in the left panel of Figure 2.5, which uses color bars to show, for each of the three subtypes, the phylum level RA (the number shown) and the size of each subtype. Phyla with very low RA ($< 1.0\%$) are grouped into the *Other* category. The bars with RA value less than 5.0% are not labeled with an RA value. Subtype 1 is dominated by OTUs belonging to *Proteobacteria* (yellow). Subtype 2 is dominated by OTUs from the phylum *Firmicutes* (pink). Subtype 3, which is the most diverse, has *Bacteriodetes* (brown)

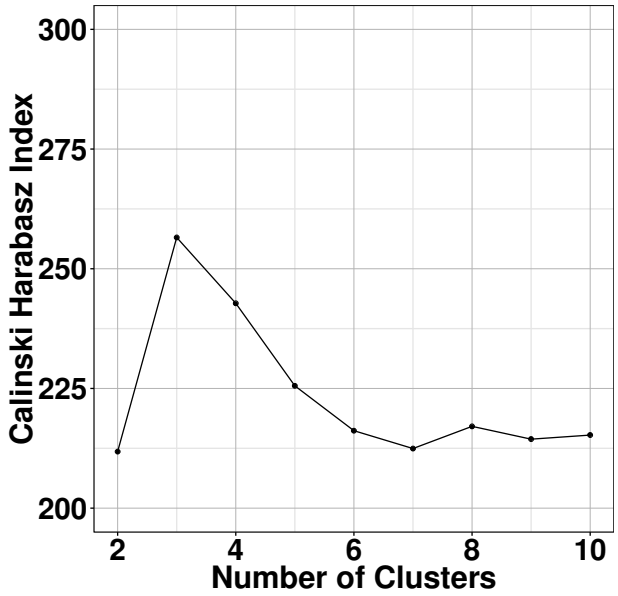


Figure 2.4: Calinski Harabasz index values for different numbers of clusters for k-means on TWA. Larger value of Calinski Harabasz index indicates a better separated clustering results.The optimal clustering should be given by the first local maximum of the values. For this microbiome data, 3-means clustering gives the optimal result while 4-means clustering is reasonably good as well.

appearing most prominently. These three subtypes represent different types of infections which make biological sense and are reasonably balanced. The right panel shows direct comparison of the clustering by PCoA, and will be discussed in Section 2.3.6.

2.3.4 Description of TWA subgroups based on diagnosis

The distribution of diagnosis groups within subtypes is shown in Figure 2.6. The data were normalized to account for differences in group sizes. The diagnosis groups are not clearly driven by the subtypes. All diagnoses except for Aspiration Pneumonia (AP) are equally represented in Subtype 1. Aspiration Pneumonia is over represented in Subtype 2 but the Control group is under represented. The Control group is over represented in Subtype 3. The diagnoses of Community Acquired Pneumonia (CAP), Non-ventilator Associated Hospital Acquired Pneumonia (NVHAP) and Ventilator-Associated Pneumonia (VAP) are all most likely to be assigned to Subtype 1 then Subtype 2 then Subtype 3.

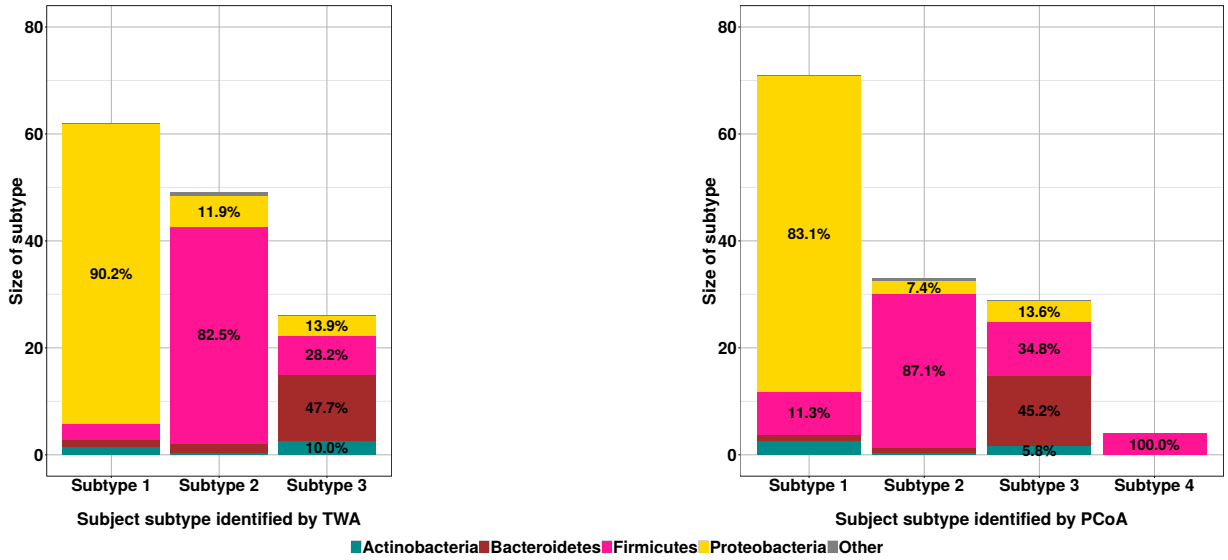


Figure 2.5: Phylum level relative abundances by subtype from TWA and PCoA. The left panel shows the barplot of the subject subtype identified by TWA, and the right panel shows the corresponding barplot for PCoA. For each subject, relative abundances are aggregated at the phylum level. For each subtype, the phylum percentages are the means of RA for all OTUs in that phylum across the subjects in that subtype. Bar heights show number of subjects in that subtype. The TWA clustering appears to do a better job of separating subjects with dominant *Proteobacteria* infections from those with dominant *Firmicutes* infections.

2.3.5 Principal components visualization

In order to visualize the clustering results in the TWA space, PCA is applied to the TWA. The first three principal components explain a large amount of the variation, about 97.4%, and the first PC explains 73%. The corresponding scree plot is provided in Figure 2.7.

igure 2.8 shows a standard PCA scatter plot. On the diagonal of Figure 2.8, 1-d projections, i.e., scores, are shown on the horizontal axes, and the vertical axes are randomly jittered for the purposes of easier visualization. The black curves on the diagonal plots are smooth histograms, i.e., kernel density estimates. The off diagonals show pairwise scatter plots. The three subtypes from Figure 2.5 are shown in different colors. They are most visually distinct in the plot of PC1 versus PC2. From visual inspection we conclude that PC1 separates TWA Subtype 1 (red) from Subtype 2 (green). PC2 separates Subtype 3 (blue) from the others. PC3 is less related to subtype structure. This is not surprising since the main subtype structure lies in the 2-d subspace determined by the 3 cluster means. The PC3 scores distribution weakly suggests bimodality, which may indicate additional structure in the data. Similar plots for RA and PCoA are shown and discussed in Section 2.3.6. These show that TWA gives a generally better visual interpretation.

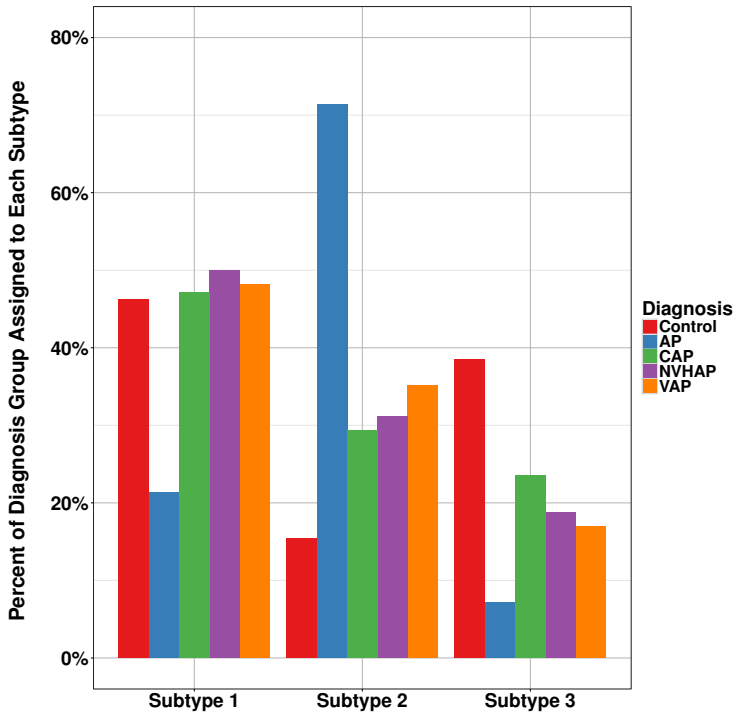


Figure 2.6: Distribution of diagnosis groups in each subtype. The diagnosis groups are not of equal sizes so in order to compare the distributions of the diagnoses across subtypes, we calculated the percent of each diagnosis that is associated with each subtype. Bars of the same color add up to 100%. All diagnoses except for Aspiration Pneumonia are equally represented in Subtype 1. Aspiration Pneumonia is over represented in Subtype 2 where as the Control group is under represented. The Control group is over represented in Subtype 3.

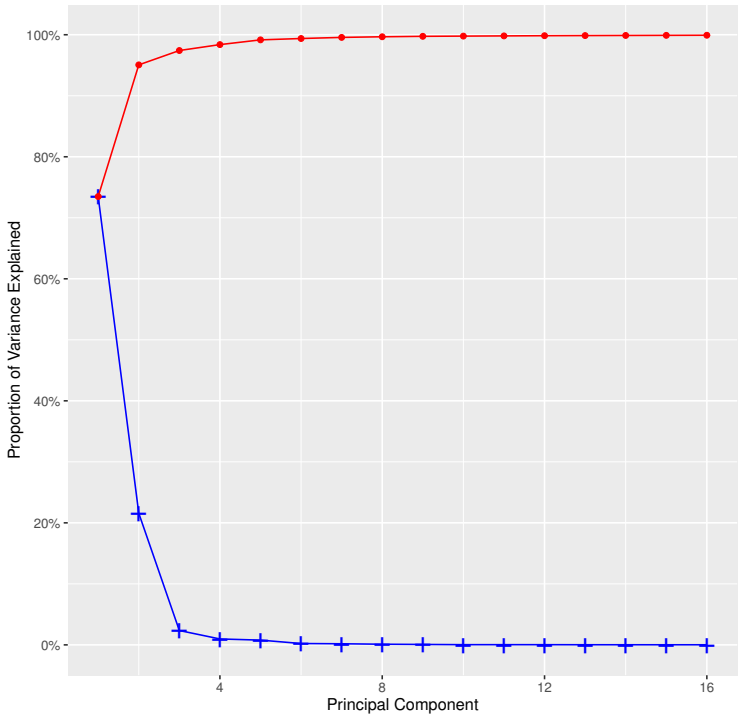


Figure 2.7: The PCA scree plot of TWA matrix. The scree plot of first sixteen PCs shows the proportion of total variance attributed to each principal component (blue line) and the cumulative variance explained (red line). The first three principal components explain about 96% percent of the variance in the data. The first seven PCs explain 99.3% of the total variation. The first sixteen PCs explain 99.9% of the total variation.

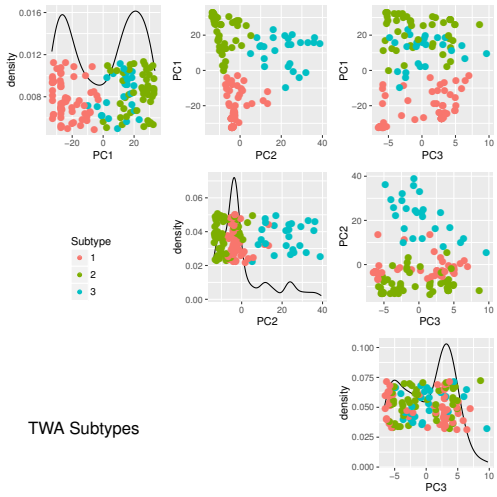


Figure 2.8: The PCA scatter plot matrix for TWA. Visualization of the first three principal components. The points are colored based on their subtype membership from the 3-means clustering on TWA shown in Figure 2.5. The plots on the diagonal show the univariate distribution for each of the three PCs. Off diagonal plots are the pairwise scatter plots. The three clusters are clearly differentiated in the plot of PC1 versus PC2.

2.3.6 Comparison of results from other approaches

Clustering results based on the three approaches are shown in Figure 2.9. Applying the DMM modeling to the taxa level RA matrix does not identify any subtypes so these results are not shown in the figure. Applying k-means clustering to the RA matrix identifies two subtypes based on the Calinski-Harabasz index. Subjects in the smaller of the two clusters have a single dominant OTU (shown as red in the left column of the colored bar, and as a short vertical line segment in the RA matrix). Subjects in the other cluster are simply all the rest. As shown in the right two columns of the colored bar, clustering based on the PCoA score matrix gives results that are fairly similar to TWA. Subtype 1 in TWA remains in the first subtype of PCoA. In contrast, PCoA distributes members of TWA Subtype 2 among its other subtypes. An interesting question is which phyla have left the TWA Subtype 2 to join the others. This is answered by the right panel of Figure 2.5, which shows the phylum level RA by subtype from the PCoA approach. Note that both *Proteobacteria* and *Firmicutes* have been moved to PCoA Subtype 1. Furthermore, *Firmicutes* have also been moved to PCoA Subtype 3 and to the new PCoA Subtype 4. Finally notice the overall subtype sizes are much less balanced for PCoA.

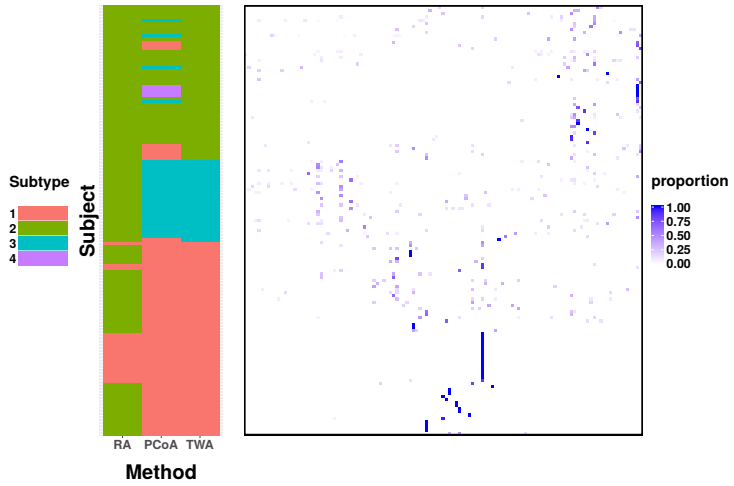


Figure 2.9: Subject cluster labels generated through different approaches and the corresponding relative abundance matrix. The left side of the figure shows three columns. Each column corresponds to an approach; RA, PCoA and TWA. The rows (subjects) are ordered by cluster labels of TWA results. The entries are color coded by cluster membership. The plot on the right shows the RA matrix, where the rows correspond to the colored bar. The column order is based on the phylogenetic tree. This shows TWA gives the most useful clustering.

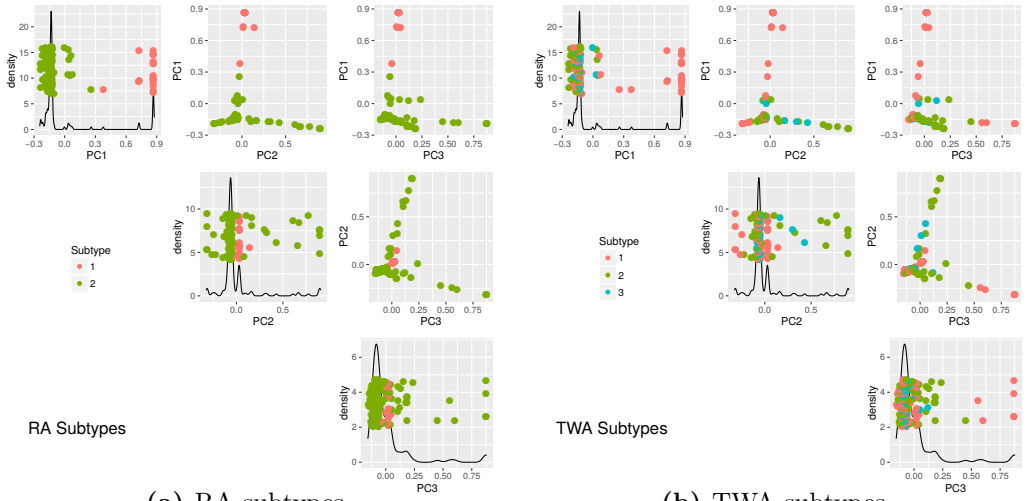
Since k-means clustering gives different numbers of clusters on different data objects, in order to make a complete comparison we evaluate 2-means, 3-means and 4-means clustering on all three data objects. SWISS (Cabanski et al., 2010) scores and Calinski-Harabasz index values, two nonparametric criteria for clustering evaluation, are provided in Table 2.1 for all three methods. Large values of the Calinski-Harabasz index and small values of the SWISS score indicate better separated clusters. As shown in Table 2.1, 3-means clustering based on TWA gives the best results among all approaches. PCoA is inferior to TWA, and overall is somewhat better than clustering directly on RA values.

Table 2.1: Calinski-Harabasz index values and SWISS scores from k-means clustering on different data objects. This shows much better clustering performance for TWA.

Criterion	Calinski-Harabasz index			SWISS score		
Clustering	2-means	3-means	4-means	2-means	3-means	4-means
RA	33.32	25.55	22.56	0.80	0.59	0.48
PCoA	54.50	51.38	55.45	0.71	0.65	0.56
TWA	211.81	256.54	242.79	0.39	0.299	0.301

Principal components visualization of RA and PCoA clustering results Visualization of RA and PCoA clustering results, analagous to Figure 4 based on TWA, are shown here. For the RA matrix, the first three Principal Component scatter plots colored by RA and TWA clustering results are shown in Figure 2.10(a) and Figure 2.10(b) respectively. Note that the only difference between Figure 2.11(a) and Figure 2.11(b) is the coloring of the points. PC1 shows a bimodal structure, which separates RA Subtype 1 (red) and RA Subtype 2 (green). This shows how k-means found those clusters. PC2 and PC3 don't show a clear subtype view. TWA subtypes do not follow the data structure in these scatter plots.

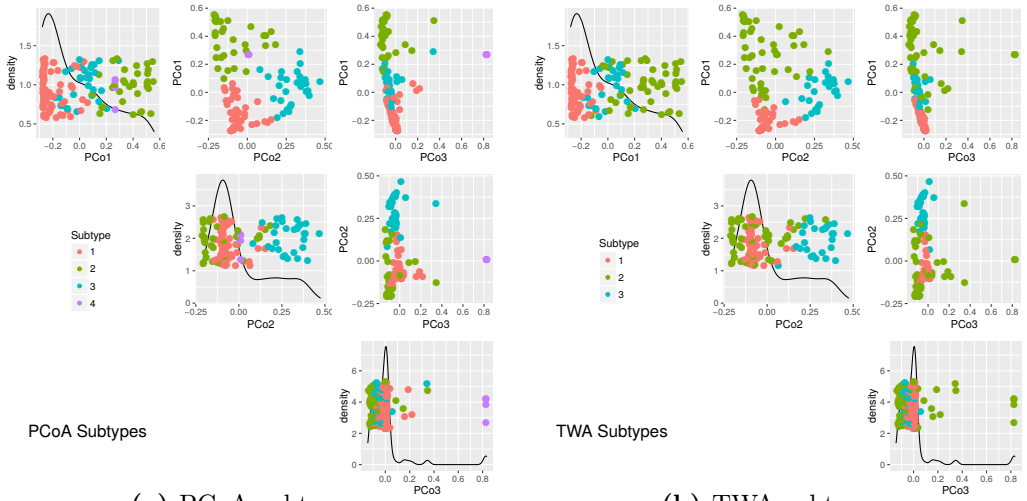
For the PCoA score matrix, the first three Principal Coordinates scatter plots are shown in Figure 2.11. The points in Figure 2.11(a) and Figure 2.11(b) are colored by PCoA and TWA clustering results, where again only colors differ. As seen in Figure 2.11(a), PCo1 separates PCoA Subtype 1 (red) from PCoA Subtype 2 (green) as well as distinguishes TWA Subtype 1 (red) with TWA Subtype 2 (green), Figure 2.11(b). PCo2 separates PCoA Subtype 3 (blue) with other PCoA subtypes and also separates TWA Subtype 3 (blue)with others. Thus there is substantially overlap



(a) RA subtypes. **(b)** TWA subtypes.

Figure 2.10: The PCA scatter plots based on the RA matrix. Visualization of the first three principal components of RA. The plots on the diagonal show the univariate distribution for each of the three PCs. Off diagonal plots are the pairwise scatter plots. The points in Figure 2.10(a) and Figure 2.10(b) are colored based on RA clustering and TWA clustering subtypes respectively. The TWA subtypes are not very related to this view.

between PCoA and TWA clusterings. A new small subtype, PCoA Subtype 4 (purple), appears in PCo3.



(a) PCoA subtypes. **(b)** TWA subtypes.

Figure 2.11: The scatter plots based on the PCoA scores matrix. Visualization of the first three principal coordinates of PCoA scores matrix. Format is similar to Figure 2.10. The points in Figure 2.11(a) and Figure 2.11(b) are colored based on PCoA clustering and TWA clustering subtypes respectively. PCo1 and PCo2 separates three PCoA subtypes and TWA subtypes. PCoA Subtype 4 can be seen as the small set of purple points in the PCo3 dimension in Figure 2.11(a).

We conclude for this example, that incorporating information from the phylogenetic tree leads to much better results than direct analysis of RA values. TWA incorporates the phylogenetic

information more effectively than PCoA. The three subtypes of subjects generated by TWA are statistically separated, well balanced, and are also easy to explain in terms of phyla.

2.4 Simulation study

The analysis in Section 2.3 indicates that TWA is a preferable approach to clustering analysis on this microbiome data. The empirical results in Section 2.3 also indicate that in each subtype of subjects the RAs concentrate on some specific OTUs, which are close in the phylogenetic tree. We use a simulation study to study potential biological contexts beyond this case. We generate OTU RA matrices that have the similar sparse features of the lung microbiome data and vary the connections between the subtype structure and the phylogenetic tree structure. The four clustering approaches in Section 2.2 are compared. Our study makes clear situations in which each analysis strategy is the best choice.

In our simulation study, the four analysis methods in Section 2.2 are compared. Applying **DMM modeling on RA** (DMM) and ***k*-means clustering on PCoA scores of Unifrac** (PCoA) are popular approaches to microbiome analysis in the literature. Applying ***k*-means clustering on RA** (RA) is a straightforward clustering approach. Applying ***k*-means clustering on TWA** (TWA) is the novel approach we propose. These methods are summarized in Table 2.2.

Table 2.2: Four methods of microbiome clustering analysis.

Method	Clustering Algorithm	Data Object	Uses Phylogenetic Tree
PCoA	k-means	PCoA-UniFrac	Yes
RA	k-means	RA	No
TWA	k-means	TWA	Yes
DMM	DMM modeling	RA	No

2.4.1 Simulation settings

The simulation model, which is designed to reflect the type of sparsity observed in the original dataset, consists of four groups/subtypes of subjects with sizes 50, 40, 30, and 20, for a total of 140 subjects. A successful analysis should find that there are four subtypes and that the membership in these subtypes agrees closely with the known group memberships. The groups are defined in terms

of their probability distribution over 100 OTUs, which are divided into five sets of sizes 10, 20, 20, 30, and 20 respectively. In this design, each group of subjects is assumed to be strongly associated with a specific set of OTUs. For simplicity of notation, a subject in Group 1 is most likely to have OTUs observed from Set 1, a subject from Group 2 is most likely to have OTUs observed from Set 2, and so on. We call Set 1 the *home* set of Group 1, etc. Set 5 is not associated with any groups of subjects and is included to make the cluster problem more challenging.

We define the *home set abundance odds ratio*, r, as the odds ratio of observing an OTU from the home set versus observing an OTU from any of the other four sets. For this simulation, r ranges from 1 to 16. The value $r = 1$ is the minimum value by definition of home set. The value $r = 16$ is a case where all methods perform well.

Two of the analysis methods require a phylogenetic tree. We generate simple phylogenetic trees where the cophenetic distances between any two OTUs from the same set are fixed at 1 and the distances between any two OTUs from different sets are fixed at d. We call d the *tree distance ratio*. The case $d < 1$ is not considered because then the phylogenetic tree would put the OTUs in the same set far from each other. The minimum considered value $d = 1$ is the case where the phylogenetic tree has no information about home sets. The maximum value $d = 1.5$ is a case where all methods perform well.

In summary, the home set abundance odds ratio, r, represents the strength of the connections between the subtype structure of subjects and the home set taxa RA values. The tree distance ratio, d, represents the strength of the alignment between subtype structure of subjects and the phylogenetic tree. Values of these two parameters, which cover a broad range of association strengths, are intended to generate conditions that will help us differentiate among the analysis methods. For each value of r, 200 datasets are generated. For each dataset, a phylogenetic tree is generated for each value of d.

2.4.2 Data generating model

In this section we introduce the data generating model in the simulation study. We intend to generate RA datasets that resemble the sparsity in the real microbiome data. In particular

- Sparsity in RAs is imposed by limiting the number of OTUs observed per subject.

- Diversity among subjects is imposed by randomly selecting observed OTUs from a larger set of equally likely OTUs in a way that has low probability of observing the same OTUs from two members of the same group.

More precisely, the simulation model creates datasets by three steps.

Step 1: Determine how many OTUs will be observed for each subject.

- The number of observed OTUs for each subject i in Group m is generated as

$$N_i = \text{Poisson}(\lambda_m) + 1,$$
$$\text{where } \lambda_1 = 2, \lambda_2 = 3, \lambda_3 = 5, \text{ and } \lambda_4 = 6.$$

Step 2: Select which OTUs are observed for each subject.

- The vector $\mathbf{p}_m$ indicates the probabilities of each OTU observed in subjects from the Group m. For example, for Group 1,

$$\mathbf{p}_1 = \frac{(\overbrace{9r, ..., 9r}^{10\times}, \overbrace{1, ..., 1}^{20\times}, \overbrace{1, ..., 1}^{20\times}, \overbrace{1, ..., 1}^{30\times}, \overbrace{1, ..., 1}^{20\times})}{90 \times r + 90}$$

  where r is the parameter representing the *home set abundance odds ratio*.

- For Subject i in Group m, N_i observed OTUs are selected through a non-replacement sampling among all OTUs with bias probability $\mathbf{p}_m$. In particular, a binary vector t_i will record the selected OTUs. If $t_{i,k} = 1$, the RA value of kth OTU will be generated as in the Step 3; otherwise, its RA value will just be 0.

Step 3: Generate RA values for the observed OTUs for each subject.

- x_i is the ith row of the RA matrix X, i.e. the RA values of each OTU for subject i.
- For a subject i, supposing it belongs to Group m, RA values are generated as

$$x_i = \text{Multinomial}(1000, \mathbf{q}_i)/1000$$

  where

$$q_{i,k} = \frac{t_{i,k} \times p_{m,k}}{\sum_{k'} t_{i,k'}}$$

2.4.3 Results of the simulation study

We evaluate the performance of the four methods in Table 2.2, using Normalized Mutual Information (NMI) (Ana and Jain, 2003). NMI is an information theory based measure that evaluates the correspondence between the cluster results and the group memberships. NMI is bounded in [0,1]. NMI = 1 when the cluster results and the group memberships are identical, and NMI = 0 when they are independent.

The left subfigure of Figure 2.12 shows results for representative d values and plots NMI values for all three methods against all values of r. For each fixed value of d, the performances of all four methods improve as r increases. Note that the curves for RA (purple) and DMM (black) are the same for all values of d (i.e. in each panel) since these methods ignore tree information. RA has better performance than DMM when r is small and DMM dominates RA in other cases. The tree-based methods, TWA (green) and PCoA (red), generally outperform the nontree-based methods, RA and DMM. Typically TWA is superior to RA, although they give almost the same results when $d = 1$. DMM is advantageous over TWA and PCoA only if d is very small and r is large. These are seen in the top left and top middle panels of the left subfigure of Figure 2.12. It is not surprising since $d \approx 1$ means that the strength of the phylogenetic relationships is extremely weak, in which case tree-based methods give very little benefit. When r is large, it is unlikely that taxa will be observed outside the home set for a subtype, which is the most favorable case for DMM. Note that in this case PCoA dominates TWA for small values of r, but TWA outperforms PCoA for larger values of r.

The top right panel and bottom left panel in the left subfigure of Figure 2.12 indicate that for moderate values of d, TWA outperforms PCoA. TWA and PCoA have similar performances when the tree information is very strong and the home set abundance odds ratio r is very large, as can be seen in the bottom middle panel and bottom right panel of left side of Figure 2.12. The blue dashed line in the bottom left panel shows the relative location of the LRI microbiome data in the context of simulation study, where TWA is a favorable approach and it is consist with empirical analysis in Section 2.3. Further discussion of this topic can be found in Section 2.5. The top left panel also shows that when tree strength is extremely weak. PCoA dominates TWA for small values of r, but TWA outperforms PCoA for larger values of r.

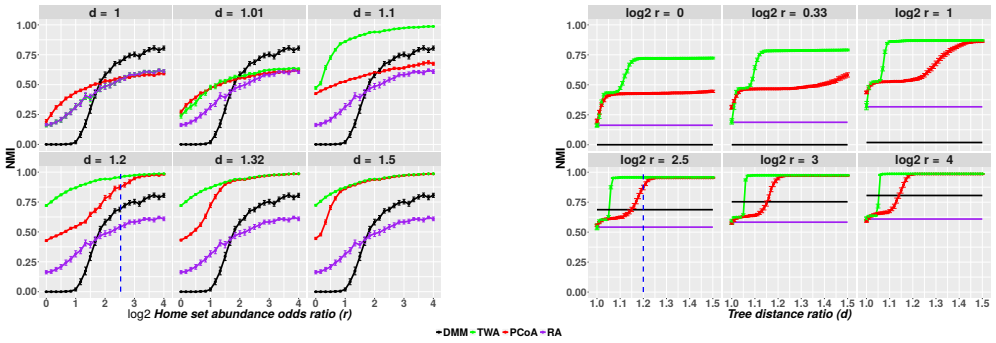


Figure 2.12: Comparison between TWA and other methods by home set abundance odds ratio (r) and tree distance ratio (d). The left subfigure shows mean NMI values for representative d values and for all values of r. The right subfigure shows mean NMI values for representative r values and for all values of d. Error bars represent two times the standard error. From the left subfigure, TWA is superior to RA except the case $d = 1$ where their performance are almost the same; TWA dominates PCoA in most cases. PCoA has the best performance when both d and r are small. In the right subfigure, for each value of r, NMI of TWA and PCoA converges to the same limit as d increases. But TWA converges faster than PCoA. The blue dashed lines show the approximate location of the microbiome data in the simulation study.

The right subfigure of Figure 2.12 provides an alternative viewpoint from which to compare these four methods. Namely, a representative set of r values is shown and NMI is plotted against all values of d. Once again, DMM (black) and RA (purple) do not use the phylogenetic tree, so their performances are constant in each panel. Particularly, the performance of RA is almost the lower bound of the performance of TWA (green) and PCoA (red). Moreover, the right subfigure of Figure 2.12 shows that for any fixed r, the performance of tree-based methods, TWA and PCoA, improves as d increases and converges to a limit, which is the upper bound on their performance. In particular, their upper bounds are almost the same and go to 1 as r increases, which indicates a perfect clustering. This implies that TWA and PCoA are equivalent and efficient when d is large enough. For any fixed r, however, TWA is more sensitive to tree information and its performance converges to the limit with respect to d much faster than PCoA.

In summary, TWA has the best performance in most of the cases considered in the simulation study, and TWA is generally helpful for identifying subtypes when subtype structure is somehow related to phylogenetic structure. PCoA, however, outperforms the other methods when d and r are extremely small. DMM dominates the other methods when d is extremely small and r is relatively large. RA is an inferior approach in all the cases.

2.5 Discussion

2.5.1 Comparison between the LRI microbiome data and the simulation study

In Section 2.3, TWA was applied to the microbiome data, and three easily described pneumotypes were identified. The simulation study in Section 2.4 also shows that TWA is an efficient method for utilizing the phylogenetic tree information to identify pneumotypes under a wide range of different biological contexts. An interesting question is how to relate the real microbiome data to the simulation model. Based on the subtypes identified by TWA, the home set abundance odds ratio and the tree distance ratio have the values $r = 5.77$ and $d = 1.20$. The details of this post hoc calculation procedure based on TWA clustering results are provided in the next section. Compared to the simulation study, the home set abundance odds ratio of $r = 5.77$ (2.53 on the log2 scale) and tree distance ratio of $d = 1.20$ matches most closely to the bottom left panel in the left subfigure of Figure 2.12 (the position of the blue dashed line). We see that TWA and PCoA outperform the RA approaches around this region of parameters. For that reason, in Figure 2.2, we focus on TWA and PCoA and study the difference in performance between the two methods based on NMI. The home set abundance odds ratio and tree distance ratio for this microbiome are shown by a black x on the figure. TWA has better performance in most regions (blue) except for the case when d is extremely small (red). In the region where both d and r are relatively large and the region where d is relatively small, the performance of TWA is nearly equivalent to PCoA (white). The real microbiome dataset falls in the area where TWA has better performance. This is consistent with the results of Section 2.3. Hence, TWA is not likely to perform worse than PCoA and may perform much better for a wide variety of sparse microbiome contexts.

2.5.2 Calculation of ratio

In this section, we describe the calculation of the home set abundance odds ratio r and tree distance ratio d based on the 3-means clustering results on TWA for this microbiome data.

First, we discuss the identification of the home set OTUs in each subtype of subjects. In each subtype, the odds-ratio of each OTU for each subtype is calculated by taking the mean RA values across all subjects and dividing by the sum of the mean relative abundances in the other two

subtypes. As in the simulation model, an OTU is assigned to a home set of the subtype if the odds ratio was 1 or greater in the subtype. This assignment is unique because the odds ratio can only be larger than 1 for at most one subtype. All OTUs could be unambiguously assigned to one of the three sets based on this criterion. The sets had 38, 41, and 42 OTUs assigned, respectively. In the context of the simulation model, each cluster is a group of subjects and the set of OTUs assigned to it can be considered the *home* set of OTUs for the group. The pneumotypes, home sets, and the phylogenetic tree can now be used to calculate the home set odds ratio and tree distance ratio.

Next we describe the calculation of the home set abundance odds ratio. For each home set, sum up the RA values across all OTUs in the home set per patient. Next, for each home set, calculate the average relative abundance per pneumotype. Finally, the home set odds ratio is the relative abundance in the associated pneumotype divided by the sum of RA values in the other two subtypes. The overall odds ratio is the average of the three home set abundance odds ratios. For this microbiome dataset, the home set abundance odds ratio is 5.77.

Last we describe the calculation of the tree distance ratio. All pairwise cophenetic distances were calculated from the phylogenetic tree. The pairwise distances were divided into two sets. The first set consists of all distances for which the pairs of OTUs are in the same home set. The second set corresponds to pairs of OTUs from different home sets. The tree distance ratio is the average cophenetic distance in the second set divided by the average distance in the first set. For this microbiome dataset, the tree distance ratio is 1.20.

2.5.3 Remarks

In our analysis we only focus on two of the most common clustering algorithms used in microbiome analysis: k-means clustering and mixture modeling. Other clustering algorithms, such as hierarchical clustering, can be also easily applied to TWA. Comparing to the choice of clustering algorithms, the definition of dissimilarity between two subjects plays a more fundamental role. In fact, if the subjects fall into very distinct subgroups, every algorithm will work well. In the opposite case, every algorithm will perform poorly. From the viewpoint of Object Oriented Data Analysis, correctly defining the data object is central to complex data set analysis. As we can

see in Table 2.1, applying various clustering algorithms on TWA always lead to statistically better separated subtypes than other approaches.

Our study shows, for microbiome data with sparsity and diversity, how the strength of tree information makes an impact on the efficiency of clustering approaches that incorporate information from the phylogenetic tree. Another interesting problem we have not investigated yet is how the sparsity and diversity differentiate the performance of TWA and PCoA.

CHAPTER 3

Angle-Based Joint And Individual Variation Explained

3.1 Introduction

A major challenge in modern data analysis is data integration, combining diverse information from disparate data sets measured on a common set of experimental subjects. Simultaneous variation decomposition has been useful in many practical applications. For example, Kühnle (2011), Lock and Dunson (2013), and Mo et al. (2013) performed integrative clustering on multiple sources to reveal novel and consistent cancer subtypes based on understanding of joint and individual variation. The Cancer Genome Atlas (TCGA) (Network et al., 2012) provides a prototypical example for this problem. TCGA contains disparate data types generated from high-throughput technologies. Integration of these is fundamental for studying cancer on a molecular level. Other types of application include analysis of multi-source metabolomic data (Kuligowski et al., 2015), extraction of commuting patterns in railway networks (Jere et al., 2014), recognition of brain-computer interface (Zhang et al., 2015), etc.

A unified and insightful understanding of the set of data blocks is expected from simultaneously exploring the joint variation representing the inter-block associations and the individual variation specific to each block. Lock et al. (2013) formulated this challenge into a matrix decomposition problem. Each data block is decomposed into three matrices modeling different types of variation, including a low-rank approximation of the joint variation across the blocks, low-rank approximations of the individual variation for each data block, and residual noise. Definitions and constraints were proposed for the joint and individual variation together with a method named JIVE; see `https://genome.unc.edu/jive/` and O'Connell and Lock (2016) for Matlab and R implementations of JIVE, respectively. Details of JIVE method are introduced in Section 3.2.

JIVE was a promising framework for studying multiple data matrices. However, Lock et al. (2013) algorithm and its implementation was iterative (thus slow) and performed rank selection

based on a permutation test. It had no guarantee of achieving a solution that satisfied the definitions of JIVE, especially in the case of some correlation between individual components. The example in Figure 3.4 in 3.2.2 shows that this can be a serious issue. An important related algorithm named COBE was developed by Zhou et al. (2016). COBE considers a JIVE-type decomposition as a quadratic optimization problem with restrictions to ensure identifiability. While COBE removed many of the shortcomings of the original JIVE, it was still iterative and often required longer computation time than the Lock et al. (2013) algorithm. Neither Zhou et al. (2016) nor Lock et al. (2013) provided any theoretical basis for selection of a thresholding parameter used for separation of the joint and individual components.

A novel solution, *Angle-based Joint and Individual Variation Explained (AJIVE)*, is proposed here for addressing this matrix decomposition problem. It provides an efficient *angle-based algorithm* ensuring an identifiable decomposition and also an insightful new interpretation of extracted variation structure. The key insight is the use of row spaces, i.e., a focus on scores, as the principal descriptor of the joint and individual variation, assuming columns are the n data objects, e.g., vectors of measurements on patients. This focuses the methodology on variation patterns across data objects, which gives straightforward definitions of the components and thus provides identifiability. These variation patterns are captured by the *score subspaces* of $\mathbb{R}^n$. Segmentation of joint and individual variation is based on studying the relationship between these score subspaces and using perturbation theory to quantify noise effects (Stewart and Sun, 1990).

The main idea of AJIVE is illustrated in the flowchart of Figure 3.1. AJIVE works in three steps. First we find a low-rank approximation of each data block (shown as the far left color blocks in the flowchart) using SVD. This is depicted (using blocks with colored dashed line boundaries) on the left side of Figure 3.1 with the black arrows signifying thresholded SVD. Next, in the middle of the figure, SVD of the concatenated bases of row spaces from the first step (the gray blocks with colored boundaries) gives a joint row space (the gray box next to the circle), using a mathematically rigorous threshold derived using perturbation theory in Section 3.3.3. This SVD is a natural extension of Principal Angle Analysis, which is also closely related to the multi-block extension of Canonical Correlation Analysis (Nielsen, 2002) as well as to the flag means of the row spaces (Draper et al., 2014); see Section 4.5 of Chapter 4 for details. Finally, the joint and individual space approximations are found using projection of the joint row space and its orthogonal

complements on the data blocks as shown as colored boundary gray squares on the right with the three joint components at the top and the individual components at the bottom.

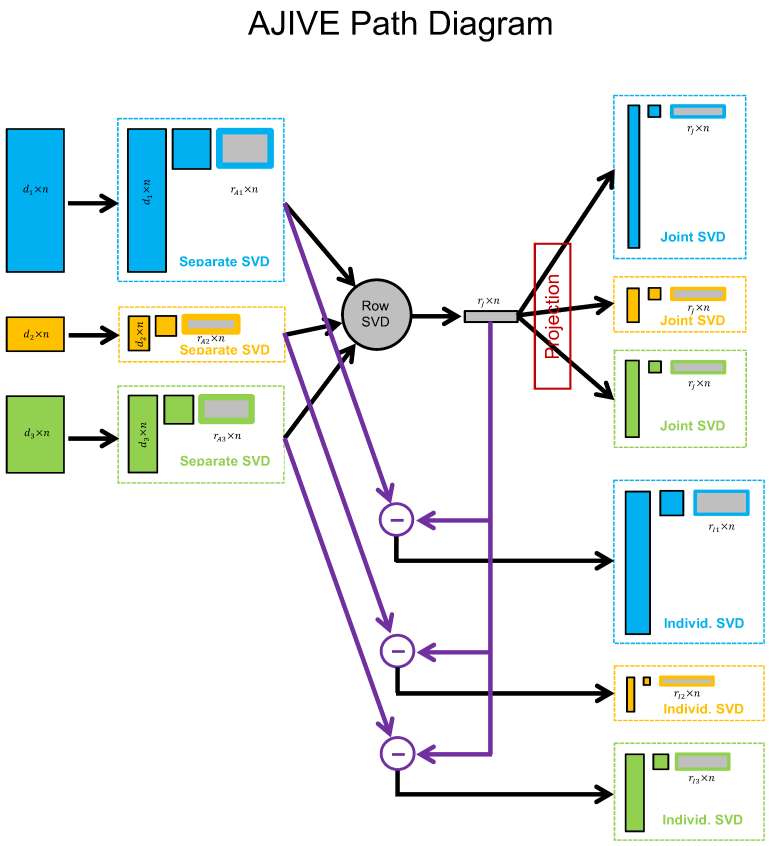


Figure 3.1: Flow chart demonstrating the main steps of AJIVE. First low rank approximation of each data block is obtained on the right. Then in the middle joint structure between the low rank approximations is extracted using SVD of the stacked row basis matrices. Finally, on the right, the joint components (upper) are obtained by projection of each data block onto the joint basis (middle) and the individual components (lower) come from orthonormal basis subtraction.

Using score subspaces to describe variation contained in a matrix not only empowers the interpretation of analysis but also improves understanding of the problem and the efficiency of the algorithm. An identifiable decomposition can now be obtained with all definitions and constraints satisfied even in situations when individual spaces are somewhat correlated. Moreover, the need to select a tuning parameter used to distinguish joint and individual variation is eliminated based on

theoretical justification using perturbation theory. A consequence is an algorithm which uses a fast built in singular value decomposition to replace lengthy iterative algorithms. For the example in Section 3.1.1, implemented in Matlab, the computational time of AJIVE (10.8 seconds) is about 11 times faster than the old JIVE (121 seconds) and 39 times faster than COBE (422 seconds). The computational advantages of AJIVE get even more pronounced on data sets with higher dimensionality and more complex heterogeneity such as the TCGA data analyzed in Section 3.4.1. For a very successful application of AJIVE on integrating fMRI imaging and behavioral data see Yu et al. (2017).

Other methods that aim to study joint variation patterns and/or individual variation patterns have also been developed. Westerhuis et al. (1998) discusses two types of methods. One main type extends traditional Principal Component Analysis (PCA), including Consensus PCA and Hierarchical PCA first introduced by Wold et al. (1987, 1996). An overview of extended PCA methods is discussed in Smilde et al. (2003). Abdi et al. (2013) discuss a multiple block extension of PCA called multiple factor analysis. This type of method computes the block scores, block loadings, global loadings and global scores.

The other main type of method is extensions of Partial Least Squares (PLS) (Wold, 1985) or Canonical Correlation Analysis (CCA) (Hotelling, 1936) that seek associated patterns between the two data blocks by maximizing covariance/correlation. For example, Wold et al. (1996) introduced multi-block PLS and hierarchical PLS (HPLS) and Trygg and Wold (2003) proposed *O2-PLS* to better reconstruct joint signals by removing structured individual variation. A multi-block extension can be found in Löfstedt et al. (2013).

Yang and Michailidis (2015) provide a very nice integrative joint and individual component analysis based on non-negative matrix factorization. Ray et al. (2014) do integrative analysis using factorial models in the Bayesian setting. Schouteden et al. (2013, 2014) propose a method called DISCO-SCA that is a low-rank approximation with rotation to sparsity of the concatenated data matrices.

A connection between extended PCA and extended PLS methods is discussed in Hanafi et al. (2011). Both types of methods provide an integrative analysis by taking the inter-block associations into account. These papers recommend use of normalization to address potential scale heterogeneity, including normalizing by the Frobenius norm, or the largest singular value of each data block

etc. However, there are no consistent criteria for normalization and some of these methods have convergence problems. An important point is that none of these approaches provide simultaneous decomposition highlighting joint and individual modes of variation with the goal of contrasting these to reveal new insights.

3.1.1 Toy Example

We give a toy example to provide a clear view of multiple challenges brought by potentially very disparate data blocks. This toy example has two data blocks, $\mathbf{X}$ (100×100) and $\mathbf{Y}$ (10000×100), with patterns corresponding to joint and individual structures. Such data set sizes are reasonable in modern genetic studies, as seen in Section 3.4.1. Figure 3.2 shows colormap views of these matrices, with the value of each matrix entry colored according to the color bar at the bottom of each subplot. The data have been simulated so expected row means are 0. Therefore mean centering is not necessary in this case. A careful look at the color bar scalings shows the values are almost four orders of magnitude larger for the top matrices. Each column of these matrices is regarded as a common data object and each row is considered as one feature. The number of features is also very different as labeled in the y-axis. Each of the two raw data matrices, $\mathbf{X}$ and $\mathbf{Y}$ in the left panel of Figure 3.2, is the sum of joint, individual and noise components shown in the other panels.

The joint variation for both blocks, second column of panels, presents a contrast between the left and right halves of the data matrix, thus having the same rank-1 score subspace. If for example the left half columns were male and right half were female, this joint variation component could be interpreted as a contrast of gender groups which exists in both data blocks for those features where color appears.

The $\mathbf{X}$ individual variation, third column of panels, partitions the columns into two other groups of size 50 that are arranged so the row space is orthogonal to that of the joint score subspace. The individual signal for $\mathbf{Y}$ contains two variation components, each driven by half of the features. The first component, displayed in the first 5000 rows, partitions the columns into three groups. The other component is driven by the bottom half of the features and partitions the columns into two groups, both with row spaces orthogonal to the joint. Note that these two individual score

subspaces for $\mathbf{X}$ and $\mathbf{Y}$ are different but not orthogonal. The smallest angle between the individual subspaces is 45°.

This example presents several challenging aspects, which also appear in real data sets such as TCGA, as studied in Section 3.4.1. One is that both the values and the number of the features are orders of magnitude different between $\mathbf{X}$ and $\mathbf{Y}$. Another important challenge is that because the individual spaces are not orthogonal, the individual signals are correlated. Correctly handling these in an integrated manner is a major improvement of AJIVE over earlier methods. In particular, normalization is no longer an issue because AJIVE only uses the low rank initial *scores* (represented as the gray boxes in the SVD shown on the left of Figure 3.1), while signal power appears in the central subblocks and the features only in the left subblocks. Appropriate handling of potential correlation among individual components is done using perturbation theory in Section 3.3.

The noise matrices, the right panels of Figure 3.2, are standard Gaussian random matrices (scaled by 5000 for $\mathbf{X}$) which generates a noisy context for both data blocks and thus a challenge for analysis, as shown in the left panels of Figure 3.2.

Simply concatenating X and Y on columns and performing a singular value decomposition on this concatenated matrix completely fails to give a meaningful joint analysis. PLS and CCA might be used to address the magnitude difference in this example and capture the signal components. However, they target common relationships between two data matrices and therefore are unable to simultaneously extract and distinguish the two types of variation. Moreover, because of its sensitivity to the strength of the signal PLS misclassifies correlated individual components as joint components. As seen in the Section 3.2.2, the original JIVE of Lock et al. (2013) also fails on this toy example.

In this toy example, the selection of the initial low rank parameters $r_{\mathbf{X}} = 2$ and $r_{\mathbf{Y}} = 3$ is unambiguous. The left panel of Figure 3.3 shows this AJIVE-approximation well captures the signal variations within both X and Y. What's more, our method correctly distinguishes the types of variation showing its robustness against heterogeneity across data blocks and correlation between individual data blocks. The approximations of both joint and individual signal are depicted in the remaining panels. A careful study of the impact of initial rank misspecification on the AJIVE results for this toy example is in Section 3.3.2 and 3.3.3.

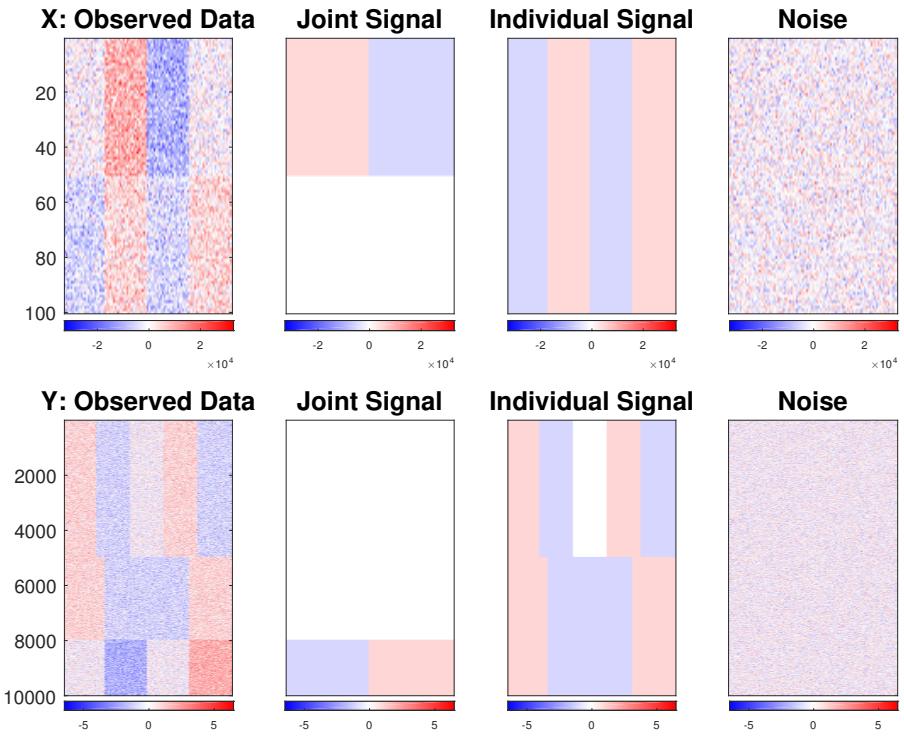


Figure 3.2: Data blocks $\mathbf{X}$ (top) and $\mathbf{Y}$ (bottom) in the toy example. The left panels present the observed data matrices which are a sum of the signal and noise matrices depicted in the remaining panels. Scale is indicated by color bars at the bottom of each sub-plot. These structures are challenging to capture using conventional methods due to very different orders of magnitude and numbers of features.

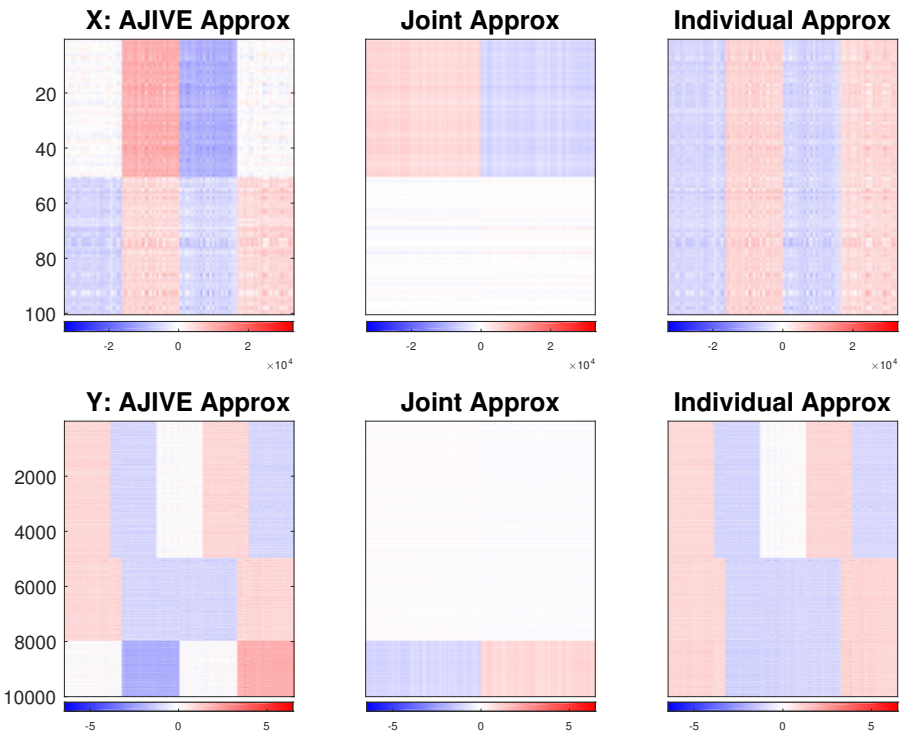


Figure 3.3: AJIVE approximation of the data blocks X and Y in the toy example are shown in the first column, with the joint and individual signal matrices depicted in the remaining columns. Both quite diverse types of variations are well captured for each data block by AJIVE, in contrast to other usual methods as seen in 3.2 and 4.

The AJIVE Matlab software, the related Matlab scripts and associated datasets, which can be used to reproduce all the results in this paper, are available at the GitHub repository `https://github.com/MeileiJiang/AJIVE_Project`.

3.2 JIVE

The details of JIVE approach (Lock et al., 2013) is reviewed in this section.

3.2.1 Model

Let matrices $\{\mathbf{X}_k(d_k \times n), k = 1, \ldots, K\}$ be a set of data blocks for study. The columns are regarded as data objects, one vector of measurements for each experimental subject, while rows are considered as features. All $\mathbf{X}_k$ therefore have the same number of columns, n, and perhaps a different number of rows, d_k. Let $\mathbf{I}_k$ be the matrix representing the individual structure of $\mathbf{X}_k$, and let $\mathbf{J}_k$ be the submatrix of the joint structure matrix that is associated with $\mathbf{X}_k$. Then, the unified JIVE model is

$$\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_K \end{bmatrix} = \begin{bmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_K \end{bmatrix} + \begin{bmatrix} \mathbf{I}_1 \\ \vdots \\ \mathbf{I}_K \end{bmatrix} + \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_K \end{bmatrix}. \tag{3.1}$$

where $\mathbf{E}_k$ are $d_k \times n$ are error matrices of independent entries with zero expectation. Let

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_K \end{bmatrix}$$

denote the joint structure matrix. The JIVE model imposes the rank constraints, i.e.,

$$\text{rank}(\mathbf{J}) = r, \text{rank}(\mathbf{I}_k) = r_k, \text{ for } k = 1, \ldots, K,$$

as well as the orthogonality between the joint matrix and individual matrices, i.e. for $k = 1, \ldots, K$,

$$\mathbf{I}_k \mathbf{J}^\intercal = \mathbf{0}_{d_k \times d}, d = \sum_{k=1}^{K} d_k.$$

This means that sample patterns responsible for joint structure between data types are unrelated to sample patterns responsible for individual structure. Under additional constraints

$$\text{rank}(\mathbf{J}_k + \mathbf{I}_k) = \text{rank}(\mathbf{J}_k) + \text{rank}(\mathbf{I}_k), k = 1, \ldots, K$$
$$\bigcap_{k=1}^{K} \text{row}(\mathbf{I}_k) = \{\vec{0}\}$$

Model 3.1 is identifiable. Lock et al. (2013) also point out that Model 3.1 can be formatted as a factorized model:

$$\mathbf{X}_k = \mathbf{U}_k\mathbf{S} + \mathbf{W}_k\mathbf{S}_k + \mathbf{E}_k, k = 1, \ldots, K. \tag{3.2}$$

The common score matrix $\mathbf{S}$ represents joint structure and summarizes patterns in the samples that explain variability across multiple data types. The loading matrices $\mathbf{U}_k$ indicate how these joint scores are expressed in the variables of data block k. The score matrices $\mathbf{S}_k$ summarize sample patterns individual to data block k, with variable loading $\mathbf{W}_k$.

3.2.2 Estimation

Joint and individual structures are estimated by minimizing the sum of squared error. Let $\mathbf{R}$ be the $d \times n$ residual matrix:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 \\ \vdots \\ \mathbf{R}_K \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 - \mathbf{J}_1 - \mathbf{I}_1 \\ \vdots \\ \mathbf{X}_K - \mathbf{J}_K - \mathbf{I}_K \end{bmatrix}.$$

Lock et al. (2013) estimates the matrices $\mathbf{J}$ and $\mathbf{I}_1, \ldots, \mathbf{I}_K$ by minimizing $\|\mathbf{R}\|^2$ under given ranks $r, r_1, \ldots, r_K$. This is accomplished by the following iterative algorithm:

- Initialize $\mathbf{X}^{\text{Joint}} = \begin{bmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_K \end{bmatrix}$
- Loop:

- Estimate $\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_K \end{bmatrix} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$ by a rank r SVD of $\mathbf{X}^{\text{Joint}}$
- For $k = 1, \ldots, K$:
  - Set $\mathbf{X}_k^{\text{individual}} = \mathbf{X}_k - \mathbf{J}_k$
  - Estimate $\mathbf{I}_k$ by a rank r_k SVD of $\mathbf{X}_k^{\text{individual}}(\mathbf{I} - \mathbf{V}\mathbf{V}')$
  - Set $\mathbf{X}_k^{\text{Joint}} = \mathbf{X}_k - \mathbf{I}_k$
- Set $\mathbf{X}^{\text{Joint}} = \begin{bmatrix} \mathbf{X}_1^{\text{Joint}} \\ \vdots \\ \mathbf{X}_K^{\text{Joint}} \end{bmatrix}$

While this algorithm ensures that the orthogonality constraint is satisfied, it has no guarantee for the rank constraint, which is a necessary condition for the identifiability of Model 3.1. Thus this algorithm can not guarantee achievement of a solution that satisfies the definitions of JIVE.

Typically, this algorithm fails on the toy data set in Section 3.1.1. We implemented the JIVE algorithm using the R package `r.jive` (O'Connell and Lock, 2016) without the orthogonality constraint. The `jive` function provides two options for rank selection: using a permutation test and the Bayesian Information Criterion, respectively. However, neither of them segmented joint signal properly. When using the Bayesian Information Criterion approach, no joint signal was identified and the true joint signals were labeled as noise. The permutation test approach gave a reasonable approximation of the total signal variation within each data block as in the left panel of Figure 3.4. However, the Lock et al. (2013) method gave rank-2 approximations to the joint matrices shown in the middle panel. The approximation consists of the real joint component together with the individual component of $\mathbf{X}$. Consequently, the approximation of the $\mathbf{X}$ individual matrix is a zero matrix and a wrong approximation of the $\mathbf{Y}$ individual matrix is shown in the top half of the right panel. We speculate that failure to correctly apportion the joint and individual variation is caused by the fact that the individual spaces are correlated, because the permutation test does not handle correlated individual signals very well. We also manually specified the correct joint and individual ranks for $\mathbf{X}$ and $\mathbf{Y}$, which results in the correct results.

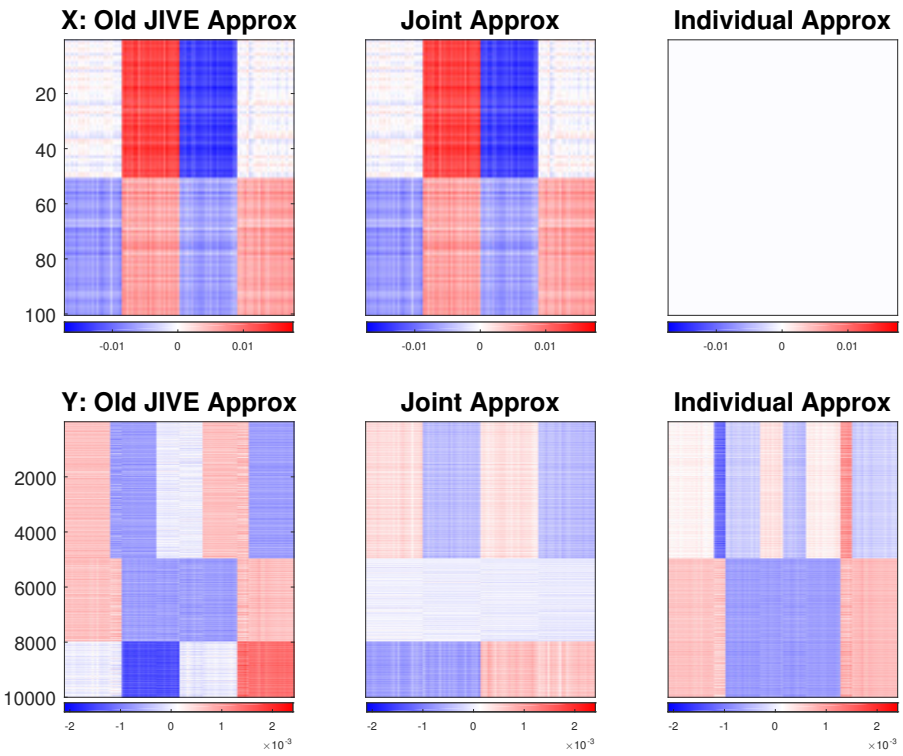


Figure 3.4: The Lock et al. (2013) JIVE method approximation of the data blocks X and Y in the toy example are shown in the first panel of figures. The joint matrix approximations (middle panel) incorrectly contain the individual component of X because of the failure of the permutation test to correctly select ranks in the presence of correlated individual components.

3.3 AJIVE

AJIVE approach (Lock et al., 2013), which is a major improvement over JIVE approach, is reviewed in this section.

3.3.1 Population model

Matrices $\{\mathbf{X}_k,\ k = 1, \ldots, K\}$ each of size $(d_k \times n)$ are a set of data blocks for study, e.g., the colored blocks on the left of Figure 3.1. The columns are regarded as data objects, one vector of measurements for each experimental subject, while rows are considered as features. All $\mathbf{X}_k$s therefore have the same number of columns and perhaps a different number of rows.

Each $\mathbf{X}_k$ is modeled as low rank true underlying signals $\mathbf{A}_k$ perturbed by additive noise matrices $\mathbf{E}_k$. Each low rank signal $\mathbf{A}_k$ is the sum of two matrices containing joint and individual variation, denoted as $\mathbf{J}_k$ and $\mathbf{I}_k$ respectively for each block, viz.

$$\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_K \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_K \end{bmatrix} + \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_K \end{bmatrix} = \begin{bmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_K \end{bmatrix} + \begin{bmatrix} \mathbf{I}_1 \\ \vdots \\ \mathbf{I}_K \end{bmatrix} + \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_K \end{bmatrix}.$$

Our approach focuses on the vectors in the row space of our matrices. In this context these vectors are often called *score vectors* and the row space of the matrix is often called *score subspace*($\subset \mathbb{R}^n$). Therefore, the row spaces of the matrices capturing joint variation, i.e., joint matrices, are defined as sharing a common score subspace denoted as row($\mathbf{J}$)

$$\text{row}(\mathbf{J}_k) = \text{row}(\mathbf{J}), \quad k = 1, \ldots, K.$$

The individual matrices are individual in the sense that they are orthogonal to the joint space, i.e., row($\mathbf{I}_k$) $\perp$ row($\mathbf{J}$),for all $k = 1, \ldots, K$, and the intersection of their score subspaces is the zero vector space, i.e.,

$$\bigcap_{k=1}^{K} \text{row}(\mathbf{I}_k) = \{\vec{0}\}, \quad k = 1, \ldots, K.$$

This means that there is no non-trivial common row pattern in every individual score subspaces across blocks.

To ensure an identifiable variation decomposition we assume $\text{row}(\mathbf{J}) \subset \text{row}(\mathbf{A}_k)$, which also implies $\text{row}(\mathbf{I}_k) \subset \text{row}(\mathbf{A}_k)$, for all $k = 1, \ldots, K$. Note that orthogonality between individual matrices $\{\mathbf{I}_k,\ k = 1, \ldots, K\}$ is *not* assumed as it is not required for the model to be uniquely determined.

Under these assumptions, the model is identifiable in the sense:

Lemma 1. *Given a set of matrices* $\{\mathbf{A}_k,\ k = 1, \ldots, K\}$*, there are unique sets of matrices* $\{\mathbf{J}_k,\ k = 1, \ldots, K\}$*, and* $\{\mathbf{I}_k,\ k = 1, \ldots, K\}$ *so that:*

1. $\mathbf{A}_k = \mathbf{J}_k + \mathbf{I}_k$*, for all* $k = 1, \ldots, K$
2. $\text{row}(\mathbf{J}_k) = \text{row}(\mathbf{J}) \subset \text{row}(\mathbf{A}_k)$*, for all* $k = 1, \ldots, K$
3. $\text{row}(\mathbf{J}) \perp \text{row}(\mathbf{I}_k)$*, for all* $k = 1, \ldots, K$
4. $\bigcap_{k=1}^{K} \text{row}(\mathbf{I}_k) = \{\vec{0}\}$.

Proof of Lemma 1. Define the row subspaces respectively for each matrix $\mathbf{A}_k$ as $\text{row}(\mathbf{A}_k) \subseteq \mathbb{R}^n$. For non-trivial cases, define a subspace $\text{row}(\mathbf{J}) \neq \{\vec{0}\}$ as the intersection of the row spaces $\{\text{row}(\mathbf{A}_1), \ldots, \text{row}(\mathbf{A}_K)\}$, i.e.,

$$\text{row}(J) \triangleq \bigcap_{k=1}^{K} \text{row}(\mathbf{A}_k).$$

For each matrix $\mathbf{A}_k$, two matrices $\mathbf{J}_k$, $\mathbf{I}_k$ can be obtained by projection of $\mathbf{A}_k$ on $\text{row}(\mathbf{J})$ and its orthogonal complement in the row space $\text{row}(\mathbf{A}_k)$. Thus the two matrices satisfy $\mathbf{J}_k + \mathbf{I}_k = \mathbf{A}_k$ and their row subspaces are orthogonal with each other, i.e., $\text{row}(\mathbf{J}) \perp \text{row}(\mathbf{I}_k)$, for all $k \in \{1, \ldots, K\}$. Then the intersection of the row subspaces $\{\text{row}(\mathbf{I}_1), \ldots, \text{row}(\mathbf{I}_K)\}$, $\bigcap_{k=1}^{K} \text{row}(\mathbf{I}_k)$, has a zero projection matrix. Therefore, we have $\bigcap_{k=1}^{K} \text{row}(\mathbf{I}_k) = \{\vec{0}\}$ and have obtained a set of matrices simultaneously satisfying the stated constraints.

On the other hand, it follows from the assumptions that the row space $\text{row}(\mathbf{A}_k)$ is spanned by the union of basis vectors of $\text{row}(\mathbf{J}_k)$ and $\text{row}(\mathbf{I}_k)$, which indicates

$$\text{row}(\mathbf{J}) = \bigcap_{k=1}^{K} \text{row}(\mathbf{A}_k).$$

Accordingly, the matrices $\mathbf{J}_1, \ldots, \mathbf{J}_K$ and $\mathbf{I}_1, \ldots, \mathbf{I}_K$ are also uniquely defined.

□

Lemma 1 is very similar to Theorem 1.1 in Lock et al. (2013). The main difference is that the rank conditions are replaced by conditions on row spaces. In our view, this provides a clearer mathematical framework and more precise understanding of the different types of variation.

The additive noise matrices, $\mathbf{E}_k$, are assumed to follow an isotropic error model where the energy of projection is invariant to direction in both row and column spaces. Important examples include the multivariate standard normal distribution and the matrix multivariate Student t–distribution (Kotz and Nadarajah, 2004). All singular values of each noise matrix are assumed to be smaller than the smallest singular values of each signal to give identifiability. This assumption on the noise distribution here is weaker than the classical i.i.d. Gaussian random matrix, and only comes into play when determining the number of joint components.

The estimation algorithm, which segments the data into joint and individual components in the presence of noise, has three main steps, as follows:

Step 1: Signal Space Initial Extraction. Low rank approximation of each data block, as shown on the left in Figure 3.1. A novel approach together with careful assessment of accuracy using matrix perturbation theory from linear algebra (Stewart and Sun, 1990), is provided in Sections 3.3.2 and 3.3.3.

Step 2: Score Space Segmentation. Initial determination of joint and individual components, as shown in the center of Figure 3.1. Our approach to this is based on an extension of Principal Angle Analysis, and an inferential based graphical diagnostic tool. The two block case is discussed in Section 3.3.3.1, with the multi-block case appearing in Section 3.3.3.2.

Step 3: Final Decomposition and Outputs. Check segmented components still meet initial thresholds in Step 1, and reproject for appropriate outputs, as shown in the right of Figure 3.1. Details of this are in Section 3.3.4.

3.3.2 Step 1: Signal Space Initial Extraction

Even though the signal components $\mathbf{A}_1, \ldots, \mathbf{A}_K$ are low rank, the data matrices $\mathbf{X}_1, \ldots, \mathbf{X}_K$ are usually of full rank due to the presence of noise. SVD works as a signal extraction device in

this step, keeping components with singular values greater than selected thresholds individually for each data block, as discussed in Section 3.3.2.1. The accuracy of this SVD approximation will be carefully estimated in Section 3.3.2.2, and will play an essential role in segmenting the joint space in Step 2.

3.3.2.1 Initial Low Rank Approximation

Each signal block $\mathbf{A}_k$ is estimated using SVD of $\mathbf{X}_k$. Given a threshold t_k, the estimator $\tilde{\mathbf{A}}_k$ (represented in Figure 3.1 as the boxes with dashed colored boundaries on the left) is defined by setting all singular values below t_k to 0. The resulting rank $\tilde{r}_k$ of $\tilde{\mathbf{A}}_k$ is an initial estimator of the signal rank r_k. The reduced rank decompositions of the $\tilde{\mathbf{A}}_k$s are

$$\tilde{\mathbf{A}}_k = \tilde{\mathbf{U}}_k \tilde{\mathbf{\Sigma}}_k \tilde{\mathbf{V}}_k^\top, \tag{3.3}$$

where $\tilde{\mathbf{U}}_k$ contains the left singular vectors that correspond to the largest $\tilde{r}_k$ singular values respectively for each data block. The initial estimate of the signal score space, denoted as $\text{row}(\tilde{\mathbf{A}}_k)$, is spanned by the right singular vectors in $\tilde{\mathbf{V}}_k$ (shown as gray boxes with colored boundaries on the left of Figure 3.1).

When selecting these thresholds, one needs to be aware of a bias/variance like trade-off. Setting the threshold too high will provide an accurate estimation of the parts of the joint space that are included in the low-rank approximation. The downside is that significant portions of the joint signal might be thresholded out. This could be viewed as a low-variance high-bias situation. If the threshold is set low, then it is likely that the joint signal is included in all of the blocks. However, the precision of the segmentation in the next step can deteriorate to the point that individual components, or even worse, noise components, can be selected in the joint space. This can be viewed as the low-bias high-variance situation.

Most off-the-shelf automatic procedures for low-rank matrix approximation have as their stated goal signal reconstruction and prediction, which based on our experience tends toward thresholds that are too small, i.e. input ranks that are too large. This is sensible as adding a little bit more noise usually helps prediction but it has bad effects on signal segmentation. We therefore recommend taking a multi-scale perspective and trying several threshold choices, for example, by

considering several relatively big jumps in a scree plot. A useful inferential graphical device to assist with this choice is developed in Section 3.3.3.

Figure 3.5 shows the scree plots of each data block for the toy example in Section 3.1. The left scree plot for $\mathbf{X}$ clearly indicates a selection of rank $\tilde{r}_1 = 2$ and the right scree plot for $\mathbf{Y}$ points to rank $\tilde{r}_2 = 3$; in both cases those components stand out while the rest of the singular values decay slowly showing no clear jump.

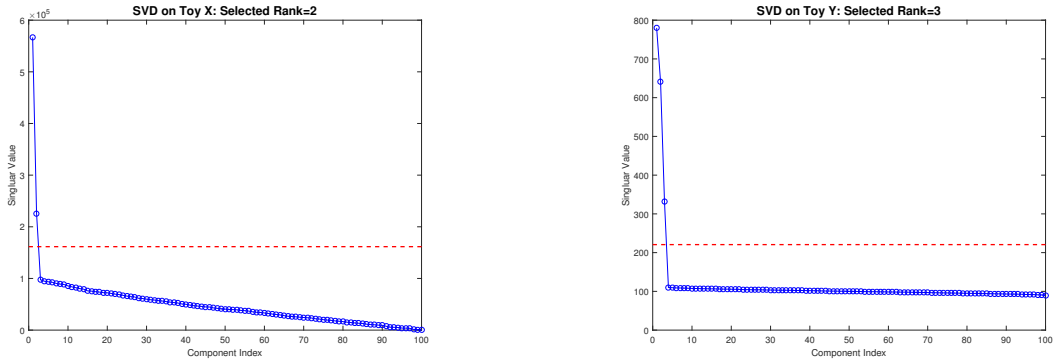


Figure 3.5: Scree plots for the toy data sets $\mathbf{X}$ (left) and $\mathbf{Y}$ (right). Both plots display the singular values associated with a component in descending order versus the index of the component. The components with singular values above the dashed red threshold line are regarded as the initial signal components in the first step of AJIVE.

3.3.2.2 Approximation Accuracy Estimation

A major challenge is segmentation of the joint and individual variation in the presence of noise which individually perturbs each signal. A first step towards addressing this is a careful study of how well $\mathbf{A}_k$ is approximated by $\tilde{\mathbf{A}}_k$ using the *Generalized* $\sin\Theta$ *Theorem* (Wedin, 1972).

Pseudometric Between Subspaces To apply the Generalized $\sin\theta$ Theorem, we use the following pseudometric as a notion of distance between theoretical and perturbed subspaces. Recall that $\text{row}(\mathbf{A}_k)$, $\text{row}(\tilde{\mathbf{A}}_k)$ are respectively the $r_k, \tilde{r}_k$ dimensional score subspaces of $\mathbb{R}^n$ respectively for the matrix $\mathbf{A}_k$ and its approximation $\tilde{\mathbf{A}}_k$. The corresponding projection matrices are $\mathbf{P}_{\mathbf{A}_k}$ and $\mathbf{P}_{\tilde{\mathbf{A}}_k}$, respectively. A pseudometric between the two subspaces can be defined as the difference of the projection matrices under the operator L^2 norm, i.e., $\rho(\text{row}(\mathbf{A}_k), \text{row}(\tilde{\mathbf{A}}_k)) = \|\mathbf{P}_{\mathbf{A}_k} - \mathbf{P}_{\tilde{\mathbf{A}}_k}\|$ (Stewart and Sun, 1990). When $r_k = \tilde{r}_k$, this pseudometric is also a distance between the two subspaces.

An insightful understanding of this pseudometric $\rho(\text{row}(\mathbf{A}_k), \text{row}(\tilde{\mathbf{A}}_k))$ comes from a principal angle analysis (Jordan, 1875; Hotelling, 1936) of the subspaces $\text{row}(\mathbf{A}_k)$ and $\text{row}(\tilde{\mathbf{A}}_k)$. Denote the principal angles between $\text{row}(\mathbf{A}_k)$ and $\text{row}(\tilde{\mathbf{A}}_k)$ as

$$\Theta(\text{row}(\mathbf{A}_k), \text{row}(\tilde{\mathbf{A}}_k)) = \{\theta_{k,1}, \ldots, \theta_{k,r_k \wedge \tilde{r}_k}\} \tag{3.4}$$

with $\theta_{k,1} \leq \theta_{k,2} \ldots \leq \theta_{k,r_k \wedge \tilde{r}_k}$. The pseudometric $\rho(\text{row}(\mathbf{A}_k), \text{row}(\tilde{\mathbf{A}}_k))$ is equal to the sine of the maximal principal angle, i.e., $\sin\theta_{k,r_k \wedge \tilde{r}_k}$. Thus the largest principal angle between two subspaces measures their closeness, i.e., distance.

The pseudometric $\rho(\text{row}(\mathbf{A}_k), \text{row}(\tilde{\mathbf{A}}_k))$ can be also written as

$$\rho(\text{row}(\mathbf{A}_k), \text{row}(\tilde{\mathbf{A}}_k)) = \|(\mathbf{I} - \mathbf{P}_{\mathbf{A}_k})P_{\tilde{\mathbf{A}}_k}\| = \|(\mathbf{I} - \mathbf{P}_{\tilde{\mathbf{A}}_k})\mathbf{P}_{\mathbf{A}_k}\|$$

which gives another useful understanding of this definition. It measures the relative deviation of the signal variation from the theoretical subspace. Accordingly, the similarity/closeness between the subspaces and its perturbation can be written as $\|\mathbf{P}_{\mathbf{A}_k}\mathbf{P}_{\tilde{\mathbf{A}}_k}\|$ and is equal to the cosine of the maximal principal angle defined above, i.e., $\cos\theta_{k,r_k \wedge \tilde{r}_k}$. Hence, $\sin^2\theta_{k,r_k \wedge \tilde{r}_k}$ indicates the proportion of signal deviation and $\cos^2\theta_{k,r_k \wedge \tilde{r}_k}$ tells the proportion of remaining signal in the theoretical subspace.

Wedin Bound For a signal matrix $\mathbf{A}_k$ and its perturbation $\mathbf{X}_k = \mathbf{A}_k + \mathbf{E}_k$, the generalized $\sin\theta$ theorem provides a bound for the distance between the rank $\tilde{r}_k (\leq r_k)$ singular subspaces of $\mathbf{A}_k$ and $\mathbf{X}_k$. This bound quantifies how the theoretical singular subspaces are affected by noise.

Theorem 1 (Wedin, 1972)**.** *Let $\mathbf{A}_k$ be a signal matrix with rank r_k. Letting $\mathbf{A}_{k,1} = \mathbf{U}_{k,1}\mathbf{\Sigma}_{k,1}\mathbf{V}_{k,1}^\top$ denote the rank $\tilde{r}_k$ SVD of $\mathbf{A}_k$, where $\tilde{r}_k \leq r_k$, write $\mathbf{A}_k = \mathbf{A}_{k,1} + \mathbf{A}_{k,0}$. For the perturbation $\mathbf{X}_k = \mathbf{A}_k + \mathbf{E}_k$, a corresponding decomposition can be made as $\mathbf{X}_k = \tilde{\mathbf{A}}_{k,1} + \tilde{\mathbf{E}}_k$, where $\tilde{\mathbf{A}}_{k,1} = \tilde{\mathbf{U}}_{k,1}\tilde{\mathbf{\Sigma}}_{k,1}\tilde{\mathbf{V}}_{k,1}^\top$ is the rank $\tilde{r}_k$ SVD of $\mathbf{X}_k$. Assume that there exists an $\alpha \geq 0$ and a $\delta > 0$ such that for $\sigma_{\min}(\tilde{\mathbf{A}}_{k,1})$ and $\sigma_{\max}(\mathbf{A}_{k,0})$ denoting appropriate minimum and maximum singular values*

$$\sigma_{\min}(\tilde{\mathbf{A}}_{k,1}) \geq \alpha + \delta, \text{ and } \sigma_{\max}(\mathbf{A}_{k,0}) \leq \alpha.$$

Then the distance between the row spaces of $\tilde{\mathbf{A}}_{k,1}$ and $\mathbf{A}_{k,1}$ is bounded by

$$\rho(\text{row}(\tilde{\mathbf{A}}_{k,1}), \text{row}(\mathbf{A}_{k,1})) \leq \frac{\max\left(\|\mathbf{E}_k\tilde{\mathbf{V}}_{k,1}\|, \|\mathbf{E}_k^\top\tilde{\mathbf{U}}_{k,1}\|\right)}{\delta} \wedge 1.$$

In practice we do not observe $\mathbf{A}_{k,0}$ thus δ cannot be estimated in general. A special case of interest for AJIVE is $\tilde{r}_k = r_k$, in which case $\mathbf{A}_{k,0} = 0, \mathbf{A}_k = \mathbf{A}_{k,1}$. The following is an adaptation of the generalized $\sin\theta$ theorem to this case:

Corollary 1 (bound for correctly specified rank)**.** *For each $k = 1, \ldots, K$, the signal matrix $\mathbf{A}_k$ is perturbed by additive noise $\mathbf{E}_k$. Let $\theta_{k,\tilde{r}_k}$ be the largest principal angle for the subspace of signal $\mathbf{A}_k$ and its approximation $\tilde{\mathbf{A}}_k$, where $\tilde{r}_k = r_k$. Denote the SVD of $\tilde{\mathbf{A}}_k$ as $\tilde{\mathbf{U}}_k\tilde{\mathbf{\Sigma}}_k\tilde{\mathbf{V}}_k^\top$. The distance between the subspaces of A_k and $\tilde{\mathbf{A}}_k$, $\rho(\text{row}(\mathbf{A}_k), \text{row}(\tilde{\mathbf{A}}_k))$, i.e., sine of $\theta_{k,\tilde{r}_k}$, is bounded above by*

$$\rho(\text{row}(\mathbf{A}_k), \text{row}(\tilde{\mathbf{A}}_k)) = \sin\theta_{k,\tilde{r}_k} \leq \frac{\max(\|E_k\tilde{\mathbf{V}}_k\|, \|E_k^\top\tilde{\mathbf{U}}_k\|)}{\sigma_{\min}(\tilde{\mathbf{A}}_k)} \wedge 1. \tag{3.5}$$

In this case the bound is driven by the maximal value of noise energy in the column and row spaces and by the estimated smallest signal singular value. This is consistent with the intuition that a deviation distance, i.e., a largest principal angle, is small when the signal is strong and perturbations are weak.

In general, it can be very challenging to correctly estimate the true rank of $\mathbf{A}_k$. If the true rank r_k is not correctly specified, then different applications of the Wedin bound are useful. In particular, when $\mathbf{A}_{k,0}$ is not 0, i.e., $\tilde{r}_k < r_k$, insights come from replacing $\mathbf{E}_k$ by $\mathbf{E}_k + \mathbf{A}_{k,0}$ in the Wedin Bound.

Corollary 2 (bound for under-specified rank)**.** *For each $k = 1, \ldots, K$, the signal matrix $\mathbf{A}_k$ with rank r_k is perturbed by additive noise $\mathbf{E}_k$. Let $\tilde{\mathbf{A}}_k = \tilde{\mathbf{U}}_k \tilde{\mathbf{\Sigma}}_k \tilde{\mathbf{V}}_k^\top$ be the rank $\tilde{r}_k$ SVD approximation of $\mathbf{A}_k$ from the perturbed matrix, where $\tilde{r}_k < r_k$. Denote $\mathbf{A}_k = \mathbf{A}_{k,1} + \mathbf{A}_{k,0}$, where $\mathbf{A}_{k,1}$ is the rank $\tilde{r}_k$ SVD of $\mathbf{A}$. Then the distance between* $\mathrm{row}(\mathbf{A}_{k,1})$ *and* $\mathrm{row}(\tilde{\mathbf{A}}_k)$ *is bounded above by*

$$\rho(\mathrm{row}(\mathbf{A}_{k,1}), \mathrm{row}(\tilde{\mathbf{A}}_k)) \leq \frac{\max\left(\|(\mathbf{E}_k + \mathbf{A}_{k,0})\tilde{\mathbf{V}}_k\|, \|(\mathbf{E}_k + \mathbf{A}_{k,0})^\top \tilde{\mathbf{U}}_k\|\right)}{\sigma_{\min}(\tilde{\mathbf{A}}_k)} \wedge 1.$$

For the other type of initial rank misspecification, $\tilde{r}_k > r_k$, we augment $\mathbf{A}_k$ with appropriate noise components to be able to use the Wedin bound.

Corollary 3 (bound for over-specified rank)**.** *For each $k = 1, \ldots, K$, the signal matrix $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top$ with rank r_k is perturbed by additive noise $\mathbf{E}_k$. Let $\tilde{\mathbf{A}}_k = \tilde{\mathbf{U}}_k \tilde{\mathbf{\Sigma}}_k \tilde{\mathbf{V}}_k^\top$ be the rank $\tilde{r}_k$ SVD of $\mathbf{X}_k$, where $\tilde{r}_k > r_k$. Let $\mathbf{E}_0$ be the rank $\tilde{r}_k - r_k$ SVD of $(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^\top)\mathbf{E}_k(\mathbf{I} - \mathbf{V}_k \mathbf{V}_k^\top)$. Then the pseudometric between* $\mathrm{row}(\mathbf{A}_k)$ *and* $\mathrm{row}(\tilde{\mathbf{A}}_k)$ *is bounded above by*

$$\rho(\mathrm{row}(\mathbf{A}_k), \mathrm{row}(\tilde{\mathbf{A}}_k)) \leq \frac{\max\left(\|(\mathbf{E}_k - \mathbf{E}_0)\tilde{\mathbf{V}}_k\|, \|(\mathbf{E}_k - \mathbf{E}_0)^\top \tilde{\mathbf{U}}_k\|\right)}{\sigma_{\min}(\tilde{\mathbf{A}}_k)} \wedge 1.$$

The bounds in Corollaries 1, 2, 3 provide many useful insights. However, these bounds still cannot be used directly since we do not observe the error matrices $\mathbf{E}_1, \ldots, \mathbf{E}_K$. A re-sampling based estimator of the Wedin bounds is provided in the next paragraph. As seen in Figure 3.6, this estimator appropriately adapts to each of the above three cases. Moreover, Figure 3.6 also indicate that the Wedin bound for over-specified rank is usually very conservative.

Estimation And Evaluation Of The Wedin Bound As mentioned above, the perturbation bounds of each $\theta_{k,r_k \wedge \tilde{r}_k}$ require the estimation of terms $\|\mathbf{E}_k \tilde{\mathbf{V}}_k\|$, $\|\mathbf{E}_k^\top \tilde{\mathbf{U}}_k\|$ for $k = 1, 2$. These terms are measurements of energies of the noise matrices projected onto the signal column and row spaces. Since an isotropic error model is assumed, the *distributions* of energy of the noise matrices in arbitrary fixed directions are equal.Thus, if we sample random subspaces of dimension $\tilde{r}_k$, that are orthogonal to the estimated signal $\tilde{\mathbf{A}}_k$, and use the observed residual $\tilde{\mathbf{E}}_k = \mathbf{X}_k \mathbf{A}_k$, this should provide a good estimator of the distribution of the unobserved terms $\mathbf{E}_k \tilde{\mathbf{V}}_k$, $\mathbf{E}_k^\top \tilde{U}_k$.

In particular, consider the estimation of the term $\|\mathbf{E}_k \tilde{\mathbf{V}}_k\|$. We draw a random subspace of dimension $\tilde{r}_k$ that is orthogonal to $\tilde{\mathbf{V}}_k$, denoted as $\mathbf{V}_k^\star$. The observed data block $\mathbf{X}_k$ is projected onto

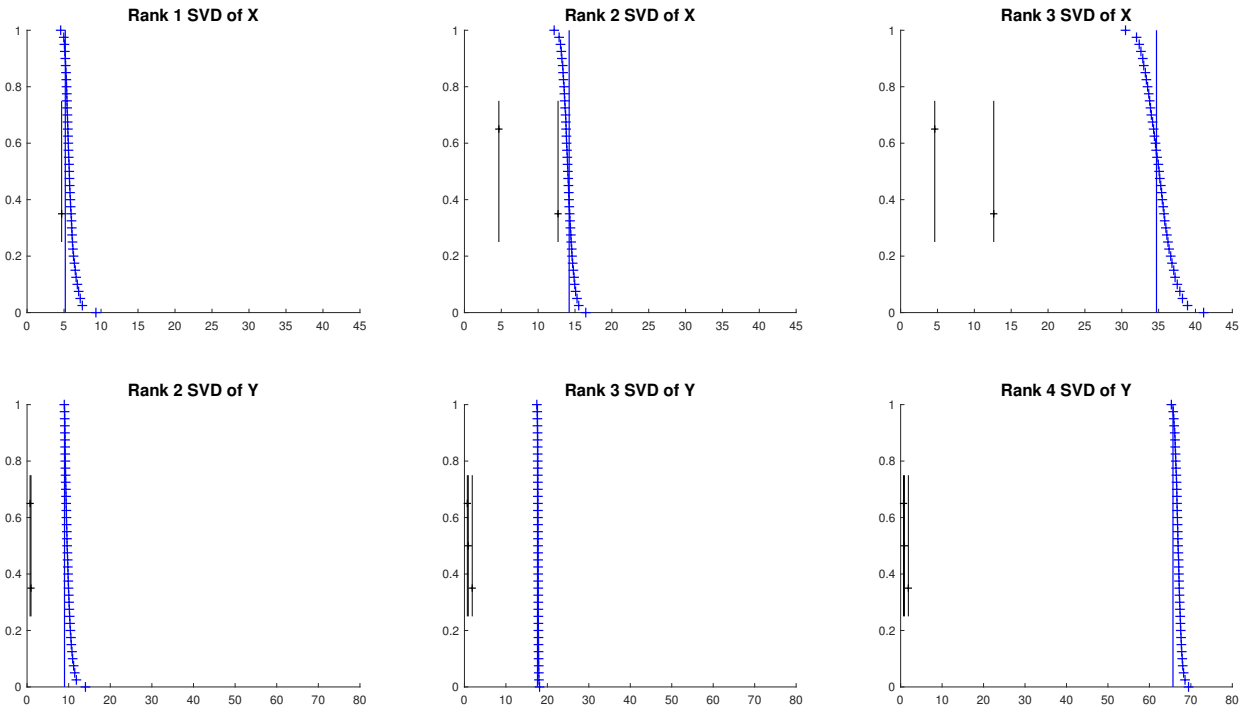


Figure 3.6: Principal angle plots between each singular subspace of the signal matrix $\mathbf{A}_{k,1}$ and its estimator $\tilde{\mathbf{A}}_k$ for the toy dataset. Graphics for $\mathbf{X}$ are on the upper row, with $\mathbf{Y}$ on the lower row. The left, middle and right columns are the under-specified, correctly specified and over-specified signal matrix rank cases respectively. Each x-axis represents the angle. The y-axis shows the values of the survival function of the resampled distribution, which are shown as blue plus signs in the figure. The vertical blue solid line is the theoretical Wedin bound, showing this bound is well estimated. The vertical black solid line segments represent the principal angles $\theta_{k,1}, \ldots, \theta_{k,r_k \wedge \tilde{r}_k}$ between row$(\mathbf{A}_{k,1})$ and row$(\tilde{\mathbf{A}}_k)$. The distance between the black and blue lines reveals when the Wedin bound is tight.

the subspace spanned by $\mathbf{V}_k^\star$, written as $\mathbf{X}_k\mathbf{V}_k^\star$. The distribution (with respect to the $\mathbf{V}_k^\star$ variation) of the operator L^2 norm $\|\mathbf{X}_k\mathbf{V}_k^\star\| = \|\tilde{\mathbf{E}}_k\mathbf{V}_k^\star\|$ approximates the distribution of the unknown $\|\mathbf{E}_k\tilde{\mathbf{V}}_k\|$ because both measure noise energy in essentially random directions. Similarly the estimation of $\|\mathbf{E}_k^\top\tilde{\mathbf{U}}_k\|$ is approximated by $\|\mathbf{X}_k^\top\mathbf{U}_k^\star\|$, where $\mathbf{U}_k^\star$ is a random $\tilde{r}_k$ dimensional subspace orthogonal to $\tilde{\mathbf{U}}_k$. These distributions are used to estimate the Wedin bound by generating 1000 replications of $\|\mathbf{X}_k\mathbf{V}_k^\star\|$ and $\|\mathbf{X}_k^\top\mathbf{U}_k^\star\|$, and plugging these into (3.5). The quantiles of the resulting distributions are used as prediction intervals for the unknown theoretical Wedin bound. Note this random subspace sampling scheme provides a distribution with smaller variance than simply sampling from the remaining singular values of $\mathbf{X}_k$, i.e. using 1000 subspaces each generated by a random sample of $\tilde{r}_k$ remaining singular vectors.

There are two criteria for evaluating the effectiveness of the estimator. First is how well the resampled distributions approximate the underlying theoretical Wedin bounds. This is addressed in Figure 3.6, which is based on the toy example in Section 3.1.1. For each of the matrices $\mathbf{X}$ and $\mathbf{Y}$ (top and bottom rows), the under, correctly, and over specified signal rank cases (Corollaries 2, 1 and 3 respectively) are carefully investigated. In each case the theoretical Wedin bound (calculated using the true underlying quantities, that are only known in a simulation study) are shown as vertical blue lines. Our resampling approach provides an estimated distribution, the survival function (1 - the c.d.f.) of which is shown using blue plus signs. This indicates remarkably effective estimation of the Wedin bound in all three cases.

The second more important criterion is how well the prediction interval covers the actual principal angles between row($\mathbf{A}_k$) and row($\tilde{\mathbf{A}}_k$). These angles are shown as vertical black line segments in Figure 3.6. For the square matrix $\mathbf{X}$, in the under and correctly specified case (top, left, and center), the Wedin bound seems relatively tight. In all other cases, the Wedin bound is conservative.

Figure 3.6 shows one realization of the noise in the toy example. A corresponding simulation study is summarized in Table 3.1. For this we generated 10,000 independent copies of the data sets $\mathbf{X}$ (100×100, true signal rank $r_1 = 2$) and $\mathbf{Y}$ (10000×100, true signal rank $r_2 = 3$). Then for several low rank approximations (columns of Table 3.1) we calculated the estimate of the angle between the true signal and the low rank approximation. Table 3.1 reports the percentage of the times the corresponding quantile of the resampled estimate is bigger than the true angle for the matrix $\mathbf{X}$.

Table 3.1: Coverages of the prediction intervals of the true angle between the signal row($\mathbf{A}_{k,1}$) and its estimator row($\tilde{\mathbf{A}}_k$) for the matrix $\mathbf{X}$ in the toy example. Rows are nominal levels. Columns are ranks of approximation (where 2 is the correct rank). The simulation based on 10000 realizations of $\mathbf{X}$ shows good performance for this square matrix.

	1	**2**	3
50%	91.9%	**63.6%**	100.0%
90%	100.0%	**89.6%**	100.0%
95%	100.0%	**93.7%**	100.0%
99%	100.0%	**98.0%**	100.0%

When the rank is correctly specified, i.e., $\tilde{r}_1 = r_1 = 2$, we see that the performance for the square matrix $\mathbf{X}$ is satisfactory as the empirical percentages are close to the nominal values. When the rank is misspecified, the empirical upper bound is conservative. Corresponding empirical percentages for the high dimension low sample size data set $\mathbf{Y}$ are all 100 %, and thus are not shown. This is caused by the fact that Wedin bound can be very conservative if the matrix is far from square. As seen in Figure 3.7 this can cause identification of spurious joint components. This motivates our development of a diagnostic plot in Section 3.3.3. Recent works of Cai et al. (2018) and O'Rourke et al. (2013) may provide potential approaches for improvement of the Wedin bound.

3.3.3 Step 2: Score Space Segmentation

3.3.3.1 Two Block Case

For a clear introduction to the basic idea of score space segmentation into joint and individual components, the two-block special case ($K = 2$) is first studied. The goal is to use the low rank approximations $\tilde{\mathbf{A}}_k$ from equation (3.3) to obtain estimates of the common joint and individual score subspaces. Due to the presence of noise, the components of row($\tilde{\mathbf{A}}_k$), $k = 1, 2$, corresponding to the underlying joint space, no longer are the same, but should have a relatively small angle. Similarly, the components corresponding to the underlying individual spaces are expected to have a relatively large angle. This motivates the use of principal angle analysis to separate the joint from the individual components.

Principal Angle Analysis One of the ways of computing the principal angles between row($\tilde{\mathbf{A}}_1$) and row($\tilde{\mathbf{A}}_2$) is to perform SVD on a concatenation of their right singular vector matrices (Miao

and Ben-Israel, 1992), i.e.,

$$\mathbf{M} \triangleq \begin{bmatrix} \tilde{\mathbf{V}}_1^\top \\ \tilde{\mathbf{V}}_2^\top \end{bmatrix} = \mathbf{U}_\mathbf{M} \mathbf{\Sigma}_\mathbf{M} \mathbf{V}_\mathbf{M}^\top, \tag{3.6}$$

where the singular values, $\sigma_{\mathbf{M},i}$, on the diagonal of $\mathbf{\Sigma}_\mathbf{M}$, determine the principal angles, $\Phi(\text{row}(\tilde{\mathbf{A}}_1), \text{row}(\tilde{\mathbf{A}}_2)) = \{\phi_1, \ldots, \phi_{\tilde{r}_1 \wedge \tilde{r}_2}\}$, where, for each $i \in \{\tilde{r}_1 \wedge \tilde{r}_2\}$,

$$\phi_i = \arccos((\sigma_{\mathbf{M},i})^2 - 1), \ i = 1, \ldots, \tilde{r}_1 \wedge \tilde{r}_2. \tag{3.7}$$

This SVD decomposition can be understood as a tool that finds pairs of directions in the two subspaces $\text{row}(\tilde{\mathbf{A}}_1)$ and $\text{row}(\tilde{\mathbf{A}}_2)$ of minimum angle, sorted in increasing order. These angles are shown as vertical black line segments in our main diagnostic graphic introduced in Figure 3.7. The first $\tilde{r}_\mathbf{J}$ column vectors in $\mathbf{V}_\mathbf{M}$ will form the orthonormal basis of the estimated joint space, $\text{row}(\mathbf{J}) \subseteq \mathbb{R}^n$. A deeper investigation of the relationship between $\mathbf{V}_\mathbf{M}$ and the canonical correlation vectors in $\mathbf{U}_\mathbf{M}$ appears in Section 4.4. Next we determine which angles are small enough to be labeled as joint components, i.e., the selection of $\tilde{r}_\mathbf{J}$.

Random Direction Bound In order to investigate which principal angles correspond to random directions, we need to estimate the distribution of principal angles generated by random subspaces. This distribution only depends on the initial input ranks of each data block, $\tilde{r}_k$, and the dimension of the row spaces, n. We obtain this distribution by simulation. In particular, $\tilde{\mathbf{V}}_1$ and $\tilde{\mathbf{V}}_2$ are replaced in (3.7) by random subspaces, i.e., each is right multiplied by an independent random orthonormal matrix. The distribution of the smallest principal angle, corresponding to the largest singular value, indicates angles potentially driven by pure noise. We recommend the 5th percentile of the angle distribution as cutoff in practice. Principal angles larger than this are not included in the joint component, which provide 95% confidence that the selected joint space does not have pure noise components. This cutoff is prominently shown in Figure 3.7 as the vertical dot-dashed red line. The c.d.f. of the underlying simulated distribution is shown as red circles.

When the individual spaces are not orthogonal, a sharper threshold based on the Wedin bounds is available.

Threshold Based On The Wedin Bound The following Lemma 2 provides a bound on the largest allowable principal angle of the joint part of the initial estimated spaces.

Lemma 2. *Let ϕ be the largest principal angle between two subspaces that are each a perturbation of the common row space within* $\text{row}(\tilde{\mathbf{A}}_1)$ *and* $\text{row}(\tilde{\mathbf{A}}_2)$*. That angle is bounded by*

$$\sin\phi \leq \sin(\theta_{1,\tilde{r}_1\wedge r_1} + \theta_{2,\tilde{r}_2\wedge r_2})$$

in which $\theta_{1,\tilde{r}_1\wedge r_1}$ and $\theta_{2,\tilde{r}_2\wedge r_2}$ are the angles given in equation (3.4)*.*

Proof of Lemma 2. Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be the projection matrices onto the individually perturbed joint row spaces. And let $\mathbf{P}$ be the projection matrix onto the common joint row space J. Thus, we have

$$\begin{aligned}\sin\theta = \|(\mathbf{I}-\mathbf{P}_1)\mathbf{P}_2\| &\leq \|(\mathbf{I}-\mathbf{P}_1)(\mathbf{I}-\mathbf{P})\mathbf{P}_2\| + \|(\mathbf{I}-\mathbf{P}_1)\mathbf{P}\mathbf{P}_2\| \\ &\leq \|(\mathbf{I}-\mathbf{P}_1)(\mathbf{I}-\mathbf{P})\|\,\|(\mathbf{I}-\mathbf{P})\mathbf{P}_2\| + \|(\mathbf{I}-\mathbf{P}_1)\mathbf{P}\|\,\|\mathbf{P}\mathbf{P}_2\|,\end{aligned}$$

in which $\|(\mathbf{I}-\mathbf{P}_1)\mathbf{P}\| = \sin\theta_{1,1}$, $\|(\mathbf{I}-\mathbf{P}_1)(\mathbf{I}-\mathbf{P})\| = \cos\theta_1$, $\|(\mathbf{I}-\mathbf{P}_2)\mathbf{P}\| = \sin\theta_{2,1}$ and $\|(\mathbf{I}-\mathbf{P}_2)(\mathbf{I}-\mathbf{P})\| = \cos\theta_{2,1}$. Therefore,

$$\sin\phi \leq \cos\theta_{1,1}\sin\theta_{2,1} + \sin\theta_{1,1}\cos\theta_{2,1} = \sin(\theta_{1,1}+\theta_{2,1}).$$

□

As with the theoretical Wedin bound, the unknown $\theta_{1,\tilde{r}_1\wedge r_1}$ and $\theta_{2,\tilde{r}_2\wedge r_2}$ are replaced by distribution estimators of the Wedin bounds. The survival function of the distribution estimator of this upper bound on ϕ is shown in Figure 3.7 using blue plus signs. The vertical dashed blue line is the 95th percentile of this distribution, giving 95% confidence that angles larger do not correspond to joint components of the lower rank approximations in Step 1. The joint rank $\tilde{r}_J$ is selected to be the number of principal angles, ϕ_i in Equation (3.7), that are smaller than both the 5th percentile of the random direction distribution and the 95th percentile of the resampled Wedin bound distribution.

Figure 3.7 illustrates how this diagnostic graphic provides many insights that are useful for initial rank selection. This considers several candidates of initial ranks. Recall for Section 3.1.1, this toy example has one joint component, one individual $\mathbf{X}$ component, and two individual $\mathbf{Y}$

components. The row subspaces of their individual components are not orthogonal and the true principal angle (only known in simulation study) is 45°. Furthermore, PCA of $\mathbf{Y}$ reveals that 79.6% of the joint component appears in the third principal component.

The upper left panel of Figure 3.7 shows the under specified rank case of $\tilde{r}_1 = \tilde{r}_2 = 2$. The principal angles (black lines) are larger than the Wedin bound (blue dashed line), so we conclude neither is joint variation. This is sensible since the true joint signal is mostly contained in the 3rd $\mathbf{Y}$ component. However, both are smaller than the random direction bound (red dashed line), so we conclude each indicates presence of correlated individual spaces.

The correctly specified rank case of $\tilde{r}_1 = 2, \tilde{r}_2 = 3$ is studied in the upper right panel of Figure 3.7. Now the smallest angle is smaller than the blue Wedin bound, suggesting a joint component. The second principal angle is about 45°, which is the angle between the individual spaces. This is above the blue Wedin bound, so it is not joint structure.

The lower left panel considers the over specified initial rank of $\tilde{r}_1 = \tilde{r}_2 = 3$. The over specification results in a loosening of the blue Wedin bound, so that now we can no longer conclude 45°is not joint, i.e., $\tilde{r}_{\mathbf{J}} = 2$ cannot be ruled out for this choice of ranks. Note that there is a third principal angle, larger than the red random direction bound, which thus cannot be distinguished from pure noise, which make sense because $\mathbf{A}_1$ has only rank $r_1 = 2$.

A case where the Wedin bound is useless is shown in the lower right. Here the initial ranks are $\tilde{r}_1 = 2$ and $\tilde{r}_2 = 4$, which results in the blue Wedin bound being actually larger than the red random direction bound. In such cases, the Wedin bound inference is too conservative to be useful. While not always true, the fact that this can be caused by over specification gives a suggestion that the initial ranks may be too large. Further analysis of this is an interesting open problem.

3.3.3.2 Multi-block Case

To generalize the above idea to more than two blocks, we focus on singular values rather than on principal angles in Equation (3.7). In other words, instead of finding an upper bound on an angle, we will focus on a corresponding lower bound on the remaining energy as expressed by the sum of the squared singular values. Hence, an analogous SVD will be used for studying the closeness of multiple initial signal score subspace estimates.

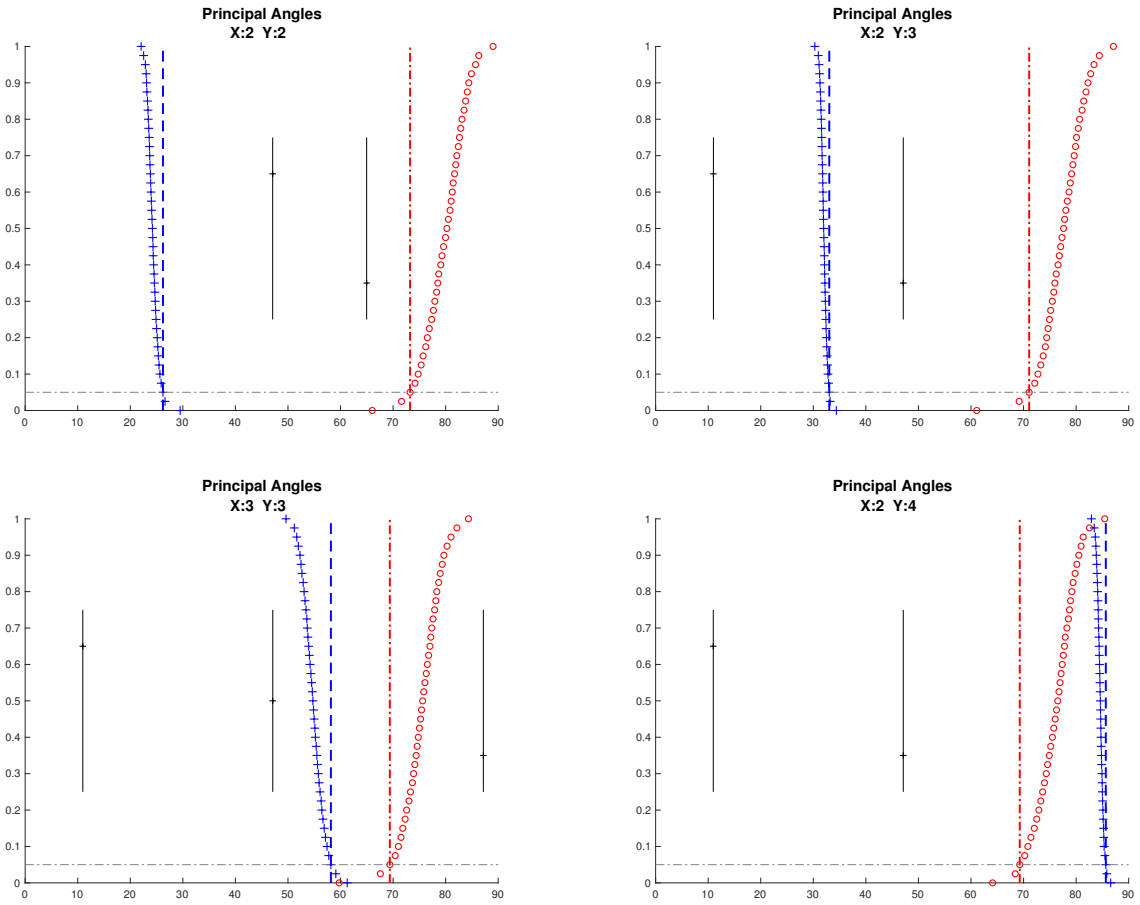


Figure 3.7: Principal angles and angle bounds used for segmentation in Step 2 of AJIVE for various input ranks. In each subfigure, the x-axis shows the angle and the y-axis shows the probabilities of the simulated distributions. The vertical black line segments are the values of the principal angles between row($\tilde{\mathbf{A}}_1$) and row($\tilde{\mathbf{A}}_2$), $\phi_1, \ldots, \phi_{\tilde{r}_1 \wedge \tilde{r}_2}$. The red circles show the values of the cumulative distribution function of the random direction distribution; the red dot-dashed line shows the 5th percentile of these angles. The blue plus signs show the values of the survival functions of the resampled Wedin bounds; the blue dashed line is the 95th percentile of the distribution. This figure contains several diagnostic plots, which provide guidance for rank selection. See Section 3.3.3.1 for details.

For the vertical concatenation of right singular vector matrices

$$\mathbf{M} \triangleq \begin{bmatrix} \tilde{\mathbf{V}}_1^\top \\ \vdots \\ \tilde{\mathbf{V}}_K^\top \end{bmatrix} = \mathbf{U}_\mathbf{M} \mathbf{\Sigma}_\mathbf{M} \mathbf{V}_\mathbf{M}^\top, \tag{3.8}$$

SVD sorts the directions within these K subspaces in increasing order of amount of deviation from the theoretical joint direction. The squared singular value $\sigma^2_{\mathbf{M},i}$ indicates the total amount

of variation explained in the common direction $\mathbf{V}_{\mathbf{M},i}^{\top}$ in the score subspace of $\mathbb{R}^n$. A large value of $\sigma_{\mathbf{M},i}^2$ (close to K) suggests that there is a set of K basis vectors within each subspace that are close to each other and thus are potential noisy versions of a common joint score vector. As in Section 3.3.3.1, the random direction bound and the Wedin bound for these singular values are used for segmentation of this joint and individual components in the multi-block case.

Random Direction Bound The extension of the random direction bound in Section 3.3.3.1 is straightforward. The distribution of the largest squared singular value in (3.8) generated by random subspaces is also obtained by simulation. As in the two block case, each $\tilde{\mathbf{V}}_k$ in $\mathbf{M}$ is replaced by an independent random subspace, i.e., right multiplied by an independent orthonormal matrix. The simulated distribution of the largest singular value of $\mathbf{M}$ indicates singular values potentially driven by pure noise. For the toy example, the values of the survival function of this distribution are shown as red circles in Figure 3.8, a singular value analog of Figure 3.7. The 5th percentile of this distribution, shown as the vertical red dot-dashed line in Figure 3.8, is used as the random direction bound for squared singular value, which provides 95% confidence that the squared singular values larger than this cutoff are not generated by random subspaces.

Threshold Based On The Wedin Bound Next is the lower bound for segmentation of the joint space based on the Wedin bound.

Lemma 3. *Let $\theta_{k,\tilde{r}_k \wedge r_k}$ be the largest principal angle between the theoretical subspace* $\mathrm{row}(\mathbf{A}_k)$ *and its estimation* $\mathrm{row}(\tilde{\mathbf{A}}_k)$ *for K data blocks from equation* (3.4)*. The squared singular values ($\sigma_{\mathbf{M},i}^2$) corresponding to the estimates of the joint components satisfy*

$$\sigma_{\mathbf{M},i}^2 \geq K - \sum_{k=1}^{K} \sin^2 \theta_{k,\tilde{r}_k \wedge r_k} \geq K - \sum_{k=1}^{K} \big(\frac{\max(\|\mathbf{E}_k \tilde{\mathbf{V}}_k\|, \|\mathbf{E}_k^{\top} \tilde{\mathbf{U}}_k\|)}{\sigma_{\min}(\tilde{\mathbf{A}}_k)} \wedge 1 \big)^2. \tag{3.9}$$

Proof of Lemma 3. Notation from (3.6) and (3.8) is used here. For each singular value $\sigma_{\mathbf{M},i}$, it can be formulated as a sequential optimization problem i.e

$$\sigma_{\mathbf{M},i}^2 = \max_{\mathbf{Q}} \|\mathbf{M}\mathbf{Q}\|_F^2 = \max_{\mathbf{Q}} \sum_{k=1}^{K} \|\tilde{\mathbf{V}}_1^{\top} \mathbf{Q}\|_F^2,$$

where $\mathbf{Q}$ is a rank-1 projection matrix that is orthogonal to the previous $i-1$ optima, i.e., $\mathbf{Q}_1, \ldots, \mathbf{Q}_{i-1}$. The $\mathbf{Q}$ that maximizes the Frobenius norm of $\mathbf{MQ}$ is denoted as $\mathbf{Q}_i$.

For an arbitrary component in the theoretical joint score subspace row($\mathbf{J}$), write its projection matrix as $\mathbf{P}_{\mathbf{J}}^{(1)}$. The Frobenius norm of $\mathbf{M}$ projected onto $\mathbf{P}_{\mathbf{J}}^{(1)}$ is

$$\|\mathbf{M}\mathbf{P}_{\mathbf{J}}^{(1)}\|_F^2 = \begin{bmatrix} \tilde{\mathbf{V}}_1^\top \mathbf{P}_{\mathbf{J}}^{(1)} \\ \vdots \\ \tilde{\mathbf{V}}_K^\top \mathbf{P}_{\mathbf{J}}^{(1)} \end{bmatrix}_F^2 \geq \begin{bmatrix} \cos\theta_1 \\ \vdots \\ \cos\theta_K \end{bmatrix}_F^2 = \sum_{k=1}^{K} \cos^2\theta_k$$

Considering the mechanism of SVD, $\sigma_{\mathbf{M},1}^2$ is the maximal norm obtained from the optimal projection matrix $\mathbf{Q}_1 \subseteq \bigcup_{k=1}^{K} \text{row}(\tilde{\mathbf{A}}_k) \subseteq \mathbb{R}^n$. If all $\tilde{\mathbf{A}}_k$ contain all components obtained by noise perturbation of the common row space row($\mathbf{J}$), then we have

$$\sigma_{\mathbf{M},1}^2 \geq \|\mathbf{M}\mathbf{P}_{\mathbf{J}}^{(1)}\|_F^2 \geq \sum_{k=1}^{K} \cos^2\theta_k.$$

to be considered as a component of the joint score subspace.

This argument can be applied sequentially. For the $\mathbf{Q}_2 \in \mathbf{Q}_1^\perp \cap \{\bigcup_{k=1}^{K} \text{row}(\tilde{\mathbf{A}}_k)\}$, there exist a non-empty joint subspace ($\subseteq$ row($\mathbf{J}$)) such that all $\mathbf{Q}_1^\perp \cap \text{row}(\tilde{\mathbf{A}}_k)$ contain perturbed directions of a joint component other than the one above. Therefore this joint component with projection matrix $\mathbf{P}_{\mathbf{J}}^{(2)}$ should have

$$\sigma_{\mathbf{M},2}^2 \geq \|\mathbf{M}\mathbf{P}_{\mathbf{J}}^{(2)}\|_F^2 \geq \sum_{k=1}^{K} \cos^2\theta_k.$$

Thus the singular values corresponding to the joint components satisfies (3.9) and this procedure can continue through at least $r_{\mathbf{J}}$ steps. □

This lower bound is independent of the variation magnitudes. This property makes AJIVE insensitive to scale heterogeneity across each block when extracting joint variation information.

As in Section 3.3.2.2, all the terms $\|\mathbf{E}_k\tilde{\mathbf{V}}_k\|$, $\|\mathbf{E}_k^\top\tilde{\mathbf{U}}_k\|$ are resampled to derive a distribution estimator for the lower bound in (3.9), which can provide a prediction interval as well. Figure 3.8 shows the values of the c.d.f. of this upper bound as blue dots for the toy example. As in the

two-block case, if there are $\tilde{r}_{\mathbf{J}}$ singular values larger than both this lower bound and the random direction bound, the first $\tilde{r}_{\mathbf{J}}$ right singular vectors are used as the basis of the estimator of row($\mathbf{J}$).

In the two-block case, Figure 3.8 contains essentially the same information as Figure 3.7, thus the same insights are available. Since principal angles between multiple subspaces are not defined, Figure 3.8 provides appropriate generalization to the multi-block case, see Figure 3.9.

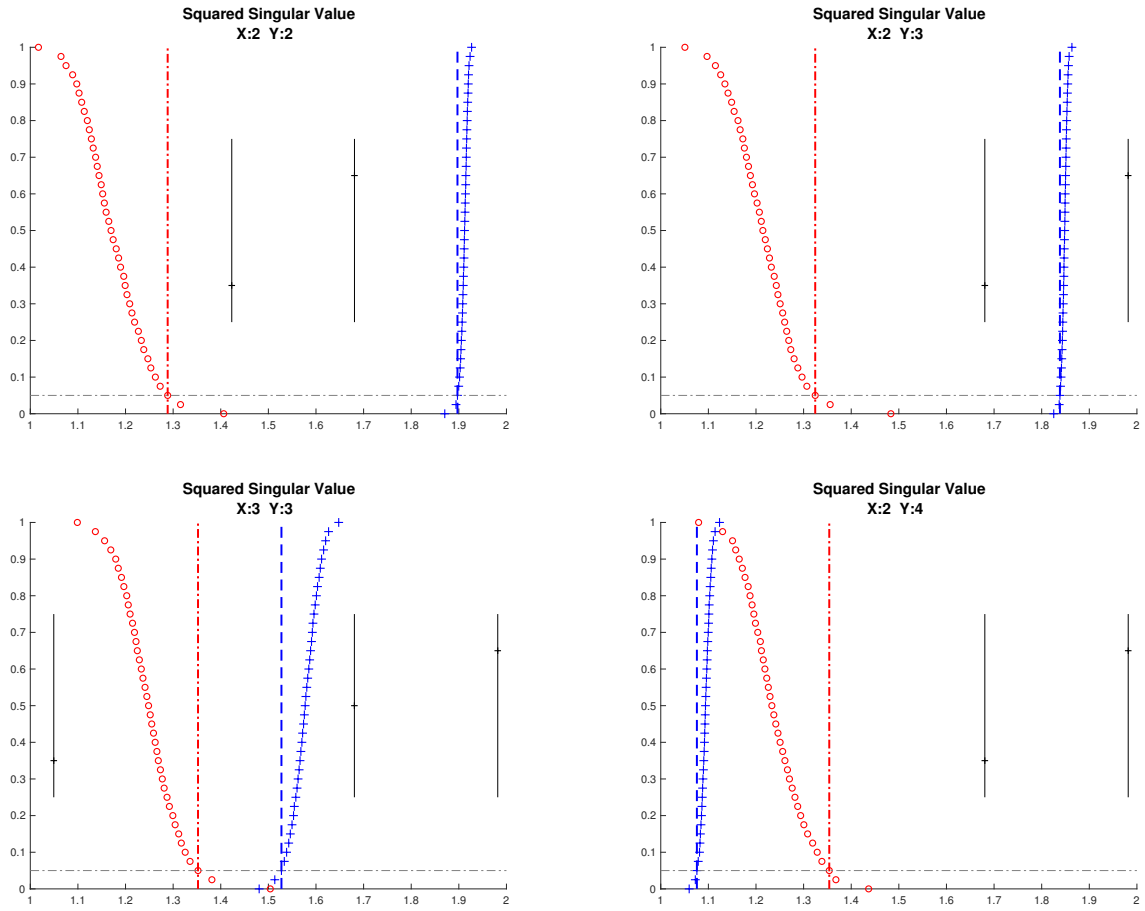


Figure 3.8: Squared singular values in (3.8) and bounds for Step 2 of AJIVE for various rank choices. The black vertical line segments shows the first $\tilde{r}_1 \wedge \tilde{r}_2$ squared singular values of $\mathbf{M}$ in equation (3.8). The values of the survival function of the random direction bounds are shown as the red circles and the red dot-dashed line is the 95th percentile of this distribution, which is the random direction bound. The values of the c.d.f of the Wedin bound are shown as the blue plus signs and the 5th percentile (blue dashed line) is used for a prediction interval for the Wedin bound. In the two-block case presented here this contains the essentially same information as in Figure 3.7. For the multi-block case it is the major diagnostic graphic.

3.3.4 Step 3: Final Decomposition And Outputs

Based on the estimate of the joint row space, matrices containing joint variation in each data block can be reconstructed by projecting $\mathbf{X}_k$ onto this estimated space. Define the matrix $\tilde{\mathbf{V}}_\mathbf{J}$ as $[\vec{v}_{\mathbf{M},1}, \ldots, \vec{v}_{\mathbf{M},\tilde{r}_\mathbf{J}}]$, where $\vec{v}_{\mathbf{M},i}$ is the ith column in the matrix $\mathbf{V}_\mathbf{M}$. To ensure that all components continue to satisfy the identifiability constraints from Section 3.3.2.1, we check that, for all the blocks, each $\|\mathbf{X}_k \vec{v}_{\mathbf{M},i}\|$ is also above the corresponding threshold used in Step 1. If the constraint is not satisfied for any block, that component is removed from $\tilde{\mathbf{V}}_\mathbf{J}$. A real example of this happens in Section 3.4.1. An important point is that this removal can happen even when there is a common joint structure in all but a few blocks.

Denote $\hat{\mathbf{V}}_\mathbf{J}$ as the matrix $\tilde{\mathbf{V}}_\mathbf{J}$ after this removal and $\hat{r}_\mathbf{J}$ as the final joint rank. The projection matrix onto the final estimated joint space $\text{row}(\hat{\mathbf{J}})$ is $\mathbf{P}_\mathbf{J} = \hat{\mathbf{V}}_\mathbf{J}\hat{\mathbf{V}}_\mathbf{J}^\top$, represented as the red rectangle in Figure 3.1. The estimates of the joint variation matrices in block $k = 1, \ldots, K$ is $\hat{\mathbf{J}}_k = \mathbf{X}_k\mathbf{P}_\mathbf{J}$.

The row space of joint structure is orthogonal to the row spaces of each individual structure. Therefore, the original data blocks are projected to the orthogonal space of $\text{row}(\hat{\mathbf{J}})$. The projection matrix onto the orthogonal space of $\text{row}(\hat{\mathbf{J}})$ is $\mathbf{P}_\mathbf{J}^\perp = \mathbf{I} - \mathbf{P}_\mathbf{J}$ and the projections of each data block are denoted as $\mathbf{X}_k^\perp$ respectively for each block, i.e.,

$$\mathbf{X}_k^\perp = \mathbf{X}_k\mathbf{P}_\mathbf{J}^\perp.$$

These projections are represented as the circled minus signs in Figure 3.1.

Finally we rethreshold this projection by performing SVD on $\mathbf{X}_1^\perp, \ldots, \mathbf{X}_K^\perp$. The components with singular values larger than the first thresholds from Section 3.3.2.1 are kept as the individual components, denoted as $\hat{\mathbf{I}}_1, \ldots, \hat{\mathbf{I}}_K$. The remaining components of each SVD are regarded as an estimate of the noise matrices.

By taking a direct sum of the estimated row spaces of each type of variation, denoted by $\oplus$, the estimated signal row spaces are

$$\text{row}(\hat{\mathbf{A}}_k) = \text{row}(\hat{\mathbf{J}}) \oplus \text{row}(\hat{\mathbf{I}}_k)$$

with rank $\hat{r}_k = \hat{r}_\mathbf{J} + \hat{r}_{\mathbf{I}_k}$ respectively for each $k = 1, \ldots, K$.

Due to this adjustment of directions of the joint components, these final estimates of signal row spaces may be different from those obtained in the initial signal extraction step. Note that even the estimates of rank $\hat{r}_k$ might also differ from the initial estimates $\tilde{r}_k$.

Given the variation decompositions of each AJIVE component, as shown on the right side of Figure 3.1, several types of post AJIVE representations are available for representing the joint and individual variation patterns. There are three important matrix representations of the information in the AJIVE joint output, i.e., the boxes on the right in Figure 3.1, with differing uses in post AJIVE analyses.

1. *Full Matrix Representation.* For applications where the original features are the main focus (such as finding driving genes) the full $d_k \times n$ matrix representations $\hat{\mathbf{J}}_k$ and $\hat{\mathbf{I}}_k$, $k = 1, \ldots, K$ are most useful. Thus this AJIVE output is the product of all three blocks in each dashed box on the right side of Figure 3.1. Examples of these outputs are shown in the two right columns of Figure 3.3.

2. *Block Specific Representation.* For applications where the relationships between subjects are the main focus (such as discrimination between subtypes) large computational gains are available by using lower dimensional representations. These are based on SVDs as indicated in the right side of Figure 3.1, i.e., for each $k = 1, \ldots, K$

$$\hat{\mathbf{J}}_k = \hat{\mathbf{U}}_J^k \hat{\mathbf{\Sigma}}_J^k \hat{\mathbf{V}}_{\mathbf{J}}^{k\top}, \hat{\mathbf{I}}_k = \hat{\mathbf{U}}_{\mathbf{I}}^k \hat{\mathbf{\Sigma}}_{\mathbf{I}}^k \hat{\mathbf{V}}_{\mathbf{I}}^{k\top}. \tag{3.10}$$

   The resulting AJIVE outputs include the joint and individual *Block Specific Score* (BSS) matrices $\hat{\mathbf{\Sigma}}_J^k \hat{\mathbf{V}}_{\mathbf{J}}^{k\top} (\hat{r}_{\mathbf{J}} \times n)$, $\hat{\mathbf{\Sigma}}_{\mathbf{I}}^k \hat{\mathbf{V}}_{\mathbf{I}}^{k\top} (\hat{r}_{\mathbf{I}_k} \times n)$ respectively. This results in no loss of information when rotation invariant methods are used. The corresponding *Block Specific Loading* matrices are $\hat{\mathbf{U}}_{\mathbf{J}}^k (d_k \times \hat{r}_{\mathbf{J}})$ and $\hat{\mathbf{U}}_{\mathbf{I}}^k (d_k \times \hat{r}_{\mathbf{I}_k})$.

3. *Common Normalized Representation.* Although $\text{row}(\hat{\mathbf{V}}_{\mathbf{J}}^{k\top})$ in (3.10) are the same, the matrices are different. In particular, the rows in (3.10) can be completely different across k, because they are driven by the pattern of the singular values in each $\hat{\mathbf{\Sigma}}_{\mathbf{J}}^k$. In some applications, correspondence of components across data blocks is important. In this case the analysis should use a common basis of $\text{row}(\hat{\mathbf{J}})$, namely $\hat{\mathbf{V}}_{\mathbf{J}}^{\top} (\hat{r}_{\mathbf{J}} \times n)$, called the *Common Normalized*

Scores (CNS). This is shown as the gray rectangular near the center of Figure 3.1. To get the corresponding loadings, we regress $\hat{\mathbf{J}}_k$ on each score vector in $\hat{\mathbf{V}}_{\mathbf{J}}^{\top}$ (which is computed as $\hat{\mathbf{J}}_k\hat{\mathbf{V}}_{\mathbf{J}}$) following by normalization. By doing this, there is no guarantee of orthogonality between CNS loading vectors. However, the loadings are linked across blocks by their common scores. For studying scale free individual spaces, use the *Individual Normalized Scores* (INS) $\hat{\mathbf{V}}_{\mathbf{I}}^{k\top}$ ($\hat{r}_{\mathbf{I}_k} \times n$). The individual loading matrices $\hat{\mathbf{U}}_{\mathbf{I}}^{k}$ are the same as the block specific individual loadings.

The relationship between Block Specific Representation and Common Normalized Representation is analogous to that of the traditional covariance, i.e., PLS, and correlation, i.e., CCA, modes of analysis. The default output in the AJIVE software is the Common Normalized Representation.

3.4 Data Analysis

In this section, we apply AJIVE to two real data sets, TCGA breast cancer in Section 3.4.1 and Spanish mortality in Section 3.4.2.

3.4.1 TCGA Data

A prominent goal of modern cancer research, of which The Cancer Genome Atlas (Network et al., 2012) is a major resource, is the combination of biological insights from multiple types of measurements made on common subjects.

TCGA provides prototypical data sets for the application of AJIVE. Here we study the 616 breast cancer tumor samples from Ciriello et al. (2015), which had a common measurement set. For each tumor sample, there are measurements of 16615 gene expression features (GE), 24174 copy number variations features (CN), 187 reverse phase protein array features (RPPA) and 18256 mutation features (Mutation). These data sources have very different dimensions and scalings.

The tumor samples are classified into four molecular subtypes: Basal-like, HER2, Luminal A and Luminal B. An integrative analysis targets the association among the features of these four disparate data sources that jointly quantify the differences between tumor subtypes. In addition, identification of driving features for each source and subtype is obtained from studying loadings.

Scree plots were used to find a set of interesting candidates for the initial ranks selected in Step 1. Various combinations of them were investigated using the diagnostic graphic. Four interesting cases are shown in Figure 3.9. The upper left panel of Figure 3.9 is a case where the input ranks are too small resulting in no joint components being identified, i.e., all the black lines are smaller than the dashed blue estimated Wedin bound. The upper right panel shows a case where only one joint component is identified. In addition to the joint component identified in the upper right panel, the lower left panel contains a second potential joint component close to the Wedin bound. The lower right panel shows a case where the Wedin bound becomes too small since the input ranks are too large. Many components are suggested as joint here, but these are dubious because the Wedin bound is smaller than the random direction bound. Between the two viable choices, in the upper right and the lower left, we investigate the latter in detail, as it best highlights important fine points of the AJIVE algorithm. In particular, we choose low rank approximations of dimensions 20 (GE), 16 (CN), 15 (RPPA) and 27 (Mutation). However, detailed analysis of the upper right panel results in essentially the same final joint component. After selection of the threshold in step 1, it took AJIVE 298 seconds (5.0 minutes, on Macbook Pro Mid 2012, 2.9 GHz) to finish steps 2 and 3.

In the second AJIVE step, the one sided 95% prediction interval suggested selection of two joint components. However, the third step indicated dropping one joint component, because the norm of the projection of the mutation data on that direction, i.e., the second CNS, is below the threshold from Step 1. This result of one joint component was consistent with the expectation of cancer researchers, who believe the mutation component has only one interesting mode of variation. A careful study of all such projections shows that the other data types, i.e., GE, CN and RPPA, do have a common second joint component as discussed at the end of this section. The association between the CNS and genetic subtype differences is visualized in the left panel of Figure 3.10. The dots are a jitter plot of the patients, using colors and symbols to distinguish the subtypes (Blue for Basal-like, cyan for HER2, red for Luminal A and magenta for Luminal B). Each symbol is a data point whose horizontal coordinate is the value and vertical coordinate is the height based on data ordering. The curves are Gaussian kernel density estimates, i.e., smoothed histograms, which show the distribution of the subtypes.

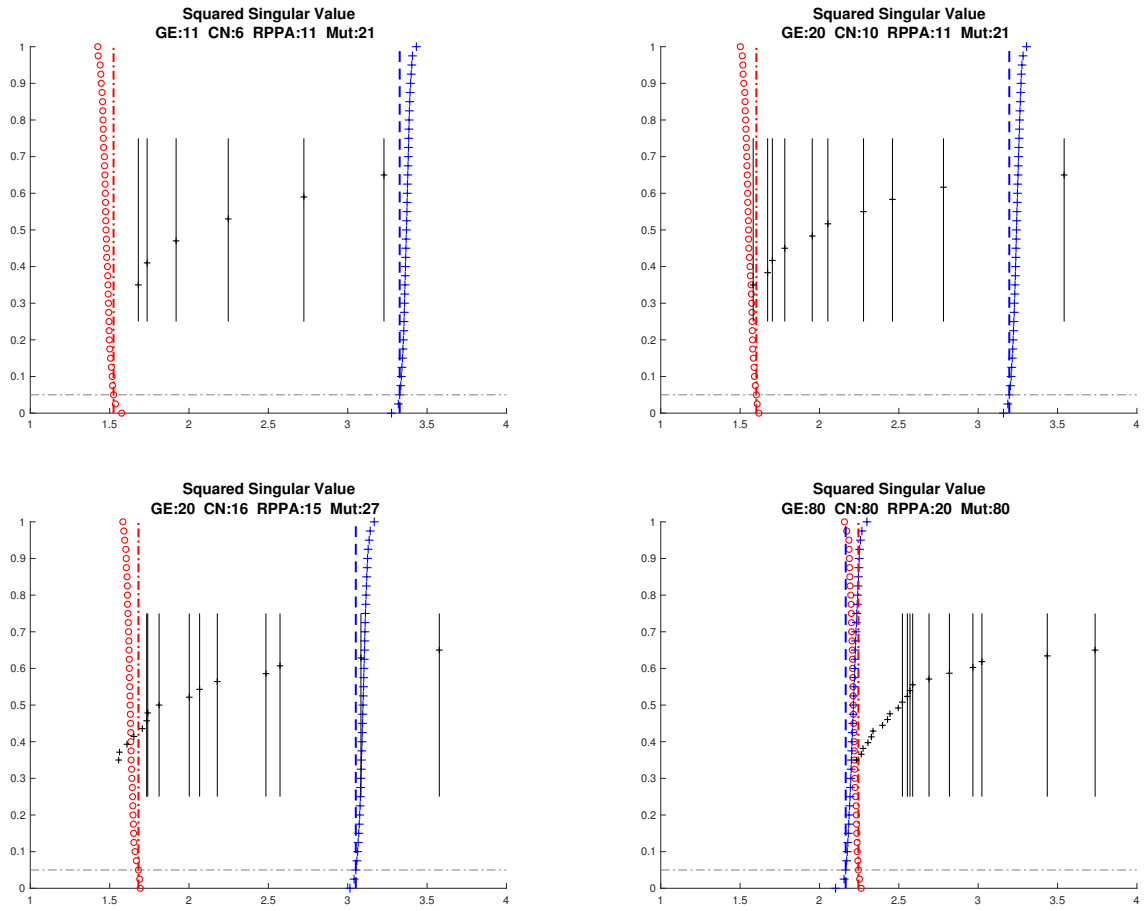


Figure 3.9: Squared singular value diagnostic graphics for TCGA dataset over various rank choices. Indicates that there are one joint component among four data blocks and one joint component among three data blocks.

The clear separation among density estimates suggest that this joint variation component is strongly connected with the subtype difference between Luminal A versus the other subtypes. To quantify this subtype difference, a test is performed using the CNS of this joint component evaluated by the DiProPerm hypothesis test (Wei et al., 2016) based on 100 permutations. Strength of the evidence is usually measured by permutation p-values. However, in this context empirical p-values are frequently zero. Thus a more interpretable measure of strength of the evidence is the DiProPerm z-score. This is 26.54 for this CNS. An area under the receiver operating characteristic (ROC) curve (AUC) (Hanley and McNeil, 1982) of 0.878, is also obtained to reflect the classification accuracy. These numbers confirm the strong Luminal A property shared by these four data types.

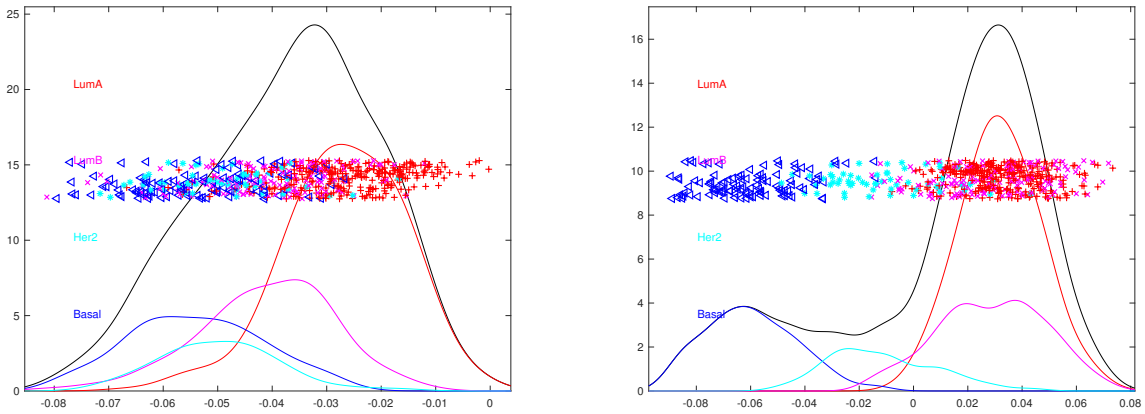


Figure 3.10: Left: Kernel density estimates of the CNS among GE, CN, RPPA and mutation. The clear separation among Luminal A versus Her2 and Basal indicates that these four data blocks share a very strong Luminal A property captured in this joint variation component; Right: The CNS from applying AJIVE to the individual matrices of GE, CN, and RPPA. The clear separation indicates that these contain a joint variation component that is consistent with the subtype difference between Basal versus the others.

A further understanding can be obtained by identifying the feature set of each data type which jointly works with the others in characterizing the Luminal A property. By studying the loading coefficients, important mutation features TP53,TTN and PIK3CA are identified which are well known features from previous studies. Similarly the strong role played by GATA3 in RPPA is well known, and is connected with the large GATA3 mutation loading. A less well known result of this analysis is the genes appearing with large GE loadings. Many of these were not flagged in earlier studies, which had focused on subgroup separation, instead of joint behavior.

As noted in the discussion of Step 2 above, all four data types have only one significant joint component. However, the individual components for all of GE, CN and RPPA seem to have 3-way joint components. This is investigated by performing a second AJIVE analysis. In particular, we apply the second and third step to the 3 individual variation matrices from the initial analysis. Notice that all individual matrices are low rank and thus the first step is not necessary. The AJIVE analysis results in one joint variation component which is displayed in the right panel of Figure 3.10. This joint variation component clearly shows the differences among Basal, HER2 and Luminal subtypes. In particular, a subtype difference between Basal-like versus the others is quantified using the DiProPerm z-score (31.60) and the AUC (0.996). Considering the fact that the AUC of the classification between Basal-like versus the others using all the original separate

GE features is 0.999, this single joint component contains almost all the variation information for separating Basal-like from the others. This hierarchical application of AJIVE reveals an important joint component that is specific to GE, CN and RPPA but not to Mutation.

3.4.2 Spanish Mortality Data

A quite different data set from the Human Mortality Database is studied here, which consists of both Spanish males and females. For each gender data block, there is a matrix of *mortality*, defined as the number of people who died divided by the total, for a given age group and year. Because mortality varies by several orders of magnitude, the $\log_{10}$ mortality is studied here. Each row represents an age group from 0 to 95, and each column represents a year between 1908 and 2002. In order to associate the historical events with the variations of mortality, columns, i.e., mortality as a function of age, are considered as the common set of data objects of each gender block. Marron and Alonso (2014b) performed analysis on the male block and showed interesting interpretations related to Spanish history. Here we are looking for a deeper analysis which integrates both males and females by exploring joint and individual variation patterns.

AJIVE is applied to the two gender blocks centered by subtracting the mean of each age group. The principal angle diagnostic graphics introduced in Section 3.3.3 are provided for this mortality dataset over various rank choices in Figure 3.11 to guide the selection of initial ranks in Step 1. The upper left panel shows the case $\tilde{r}_1 = \tilde{r}_2 = 1$. The only principal angle is larger than the 95th percentile of the resampled Wedin bound and thus we conclude that no joint space is identified. The upper right panel shows the effect of increasing the initial rank choices to $\tilde{r}_1 = \tilde{r}_2 = 2$. In this case, the first principal angle becomes smaller. Because it is smaller than the Wedin bound it is identified as a joint component. The second principal angle is still larger than the Wedin bound. Thus we concluded that only one joint component is identified in this case. In the lower left panel we increase the input rank of male mortality to 3. The second principal angle becomes much smaller, in particular smaller than the Wedin bound, and thus is also labeled as joint component. This indicates that the 3rd principal component of male mortality contains joint information. The lower right panel shows the case where $\tilde{r}_1 = 4, \tilde{r}_2 = 5$. In this case the two smallest principal angle are unchanged (and still joint). Two more principal angle appear. One is larger than the random

direction bound, and thus cannot be distinguished from pure noise. The other is just inside the boundary of the much increased Wedin bound suggesting correlation among individual components. Based on these, the choice $\tilde{r}_1 = 3, \tilde{r}_2 = 2$ is used in the subsequent analysis.

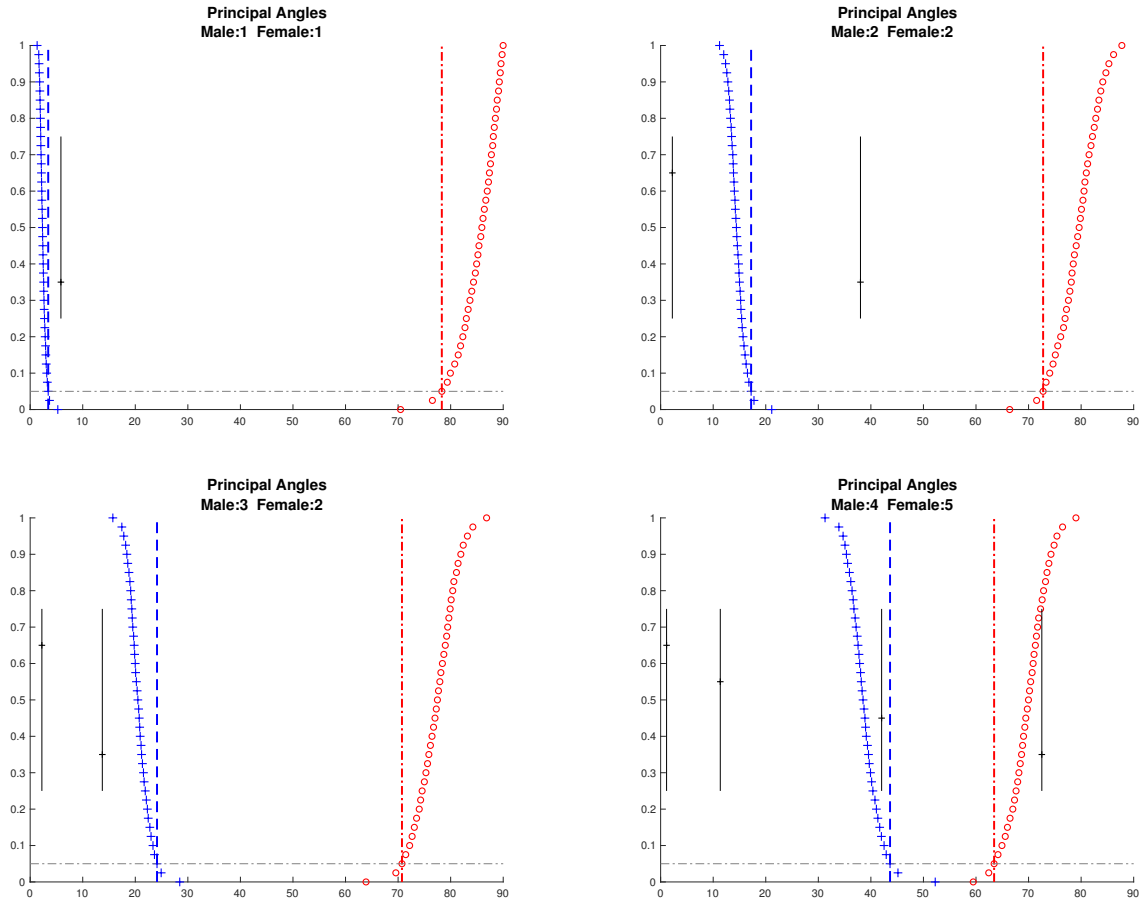


Figure 3.11: Principal angle diagnostic graphics for Spanish mortality data set over various rank choices. Provides the rationale of the rank choice, $\tilde{r}_1 = 3, \tilde{r}_2 = 2$.

The resulting AJIVE gives 2 joint components and 1 individual component for the male. Since the loading matrices provide important information on the effect of different age groups, block specific analysis together with loading matrices is most informative here.

Figure 3.12 shows a view of the first joint components for the males (left) and females (right) that is very different from the heat map views used in Section 3.1.1. While these components are matrices, additional insights come from plotting the rows of the matrices as curves over year (top) and the columns as curves over age (bottom). The curves over year (top) are colored using a heat color scheme, indexing age (black = 0 through red = 40 to yellow = 95 as shown in the vertical

color bar on the bottom left). The curves over age (bottom) are colored using a rainbow color scheme (magenta = 1908 through green = 1960 to red = 2002, shown in the horizontal color bar in the top) and use the vertical axis as domain with horizontal axis as range to highlight the fact that these are column vectors. Additional visual cues to the matrix structure are the horizontal rainbow color bar in the top panel, showing that year indexes columns of the data matrix and the vertical heat color bar (bottom) showing that age indexes rows of the component matrix. Because this is a single component, i.e., a rank one approximation of the data, each curve is a multiple of a single eigenvector. The corresponding coefficients are shown on the right. In conventional PCA/SVD terminology, the upper block specific coefficients are called *loadings*, and are in fact the entries of the left eigenvectors (colored using the heat color scale on the bottom). Similarly, the lower coefficients are called *scores* and are the entries of the right eigenvectors, colored using the rainbow bar shown in the top.

The scores plots together with the rows as curves plots in Figure 3.12 indicate a dramatic improvement in mortality over time for both males and females. The scores plots are bimodal indicating rapid overall improvement in mortality around the 1950s. This is also visible as the steepest part in the rows as curves plot. Thus the first mode of joint variation is driven by overall improvement in mortality. In addition to the overall improvement, the rows as curves and scores plots also show two major mortality events, the global flu pandemic of 1918 and the Spanish Civil war in the late 1930s. The loading plots together with the columns as curves plots present the different impacts of this common variation on different age groups for males and females. The loadings plot for males suggests the improvement in mortality is gradually increasing from older towards younger age groups. In contrast, the female block has a bimodal kernel density estimate of the loadings. This shows that females of child bearing age have received large benefits from improving health care. This effect is similarly visible from comparing the female versus male columns as curves.

The second block specific components of joint variation within each gender are similarly visualized in Figure 3.13. This common variation reflects the contrast between the years around 1950 and the years around 1980 which can be told from the curves in the left top and the colors in the right bottom subplots in both male and female panels. In the scores plots, the green circles, seen on the left end, represent the years around 1950 when automobile penetration started. And the

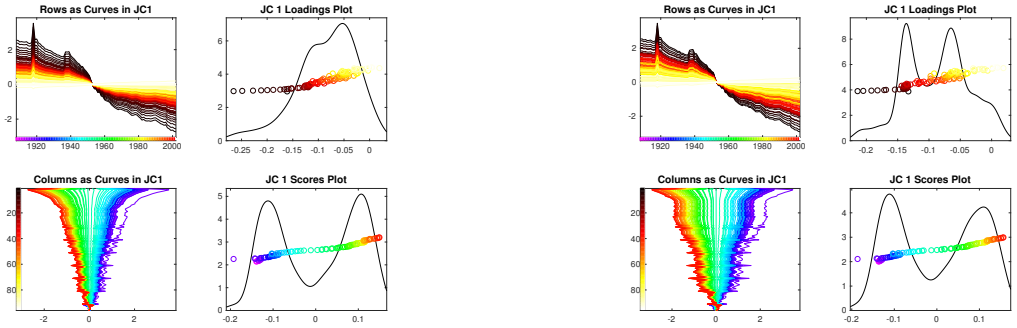


Figure 3.12: The first block specific joint components of male (left panel) and female (right panel) contain the common modes of variation caused by the overall improvement across different age groups, as can be seen from the scores plots in the right bottom of each panel. The dramatic decrease happened around the 1950s shown in the columns plots. The degree of decrease varies over age groups.

orange to red circles on the right end correspond to recent years, and much improved car and road safety. The loadings plot for males shows that these automobile events had a stronger influence on the 20-45 years old males in terms of both larger values and a second peak in the kernel density estimate. Although this contrast can also be seen in the loadings plot of females, it is not as strong as for the male block. Both loadings plots show an interesting outlier, the babies of age zero. We speculate this shows an improvement in post-natal care that coincidently happened around the same time.

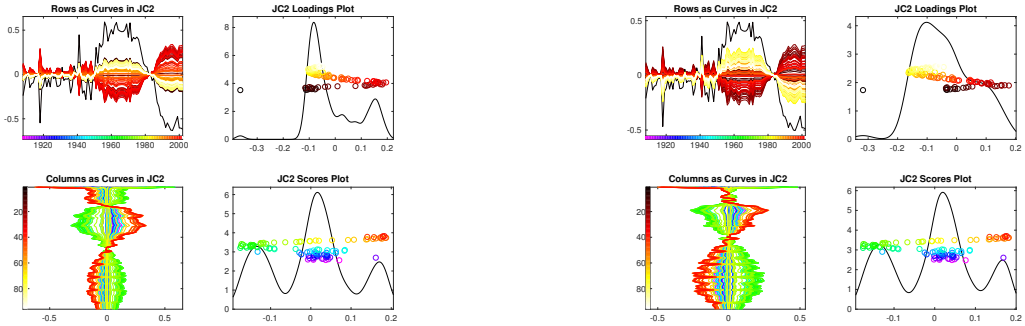


Figure 3.13: The second joint components of male (left) and female (right) contain the common modes of variation driven by the increase in fatalities caused by automobile penetration and later improvement due to safety improvements. This can be seen from the scores plots in the right bottom. The loadings plots show that this automobile event exerted a significantly stronger impact on the 20-45 males.

Another interesting result comes from the studying of the first individual components for males, shown in Figure 3.14. In the scores plot of males, the blue circles stand out from the rest, corresponding to the years of the Spanish civil war when a significant spike can be seen in the rows as curves plot. Young to middle age groups (typical military age) are affected more than the others as seen in the loadings plot and columns as curves plot.

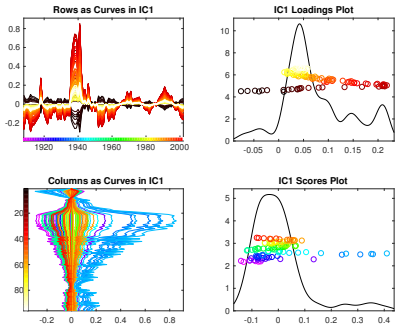


Figure 3.14: The individual component of male contains the variation driven by the Spanish civil war which can be seen from the blue circles on the right end of the right bottom plot. The Spanish civil war mainly affected the young to middle age males.

CHAPTER 4

Relationship Between AJIVE And Existing Integrative Methods

In this chapter, related existing integrative methods are explicitly discussed and compared to AJIVE. Suppose that $\mathbf{X}_k, k = 1, \cdots, K$ are $(d_k \times n)$ row centered data matrices, with SVD decompositions

$$\mathbf{X}_k = \mathbf{U}_{\mathbf{X}_k} \mathbf{\Sigma}_{\mathbf{X}_k} \mathbf{V}_{\mathbf{X}_k}, \tag{4.1}$$

where the $\mathbf{\Sigma}_{\mathbf{X}_k}$ contain no zeros on their diagonal. To be compatible with AJIVE, we will consider these algorithms from a non-standard viewpoint using row spaces. Let $\mathcal{Q}_k \subset \mathbb{R}^n$ be the row spaces of the low rank signal matrix $\mathbf{A}_k, k = 1, \cdots, K$. Assume that $\mathcal{Q}_k$ is a non-trivial subspace with $r_k = \dim(\mathcal{Q}_k) \in (0, n)$. We will also use the following notation: for column vectors $\vec{a}_1 \in \mathbb{R}^{d_1}, \vec{a}_2 \in \mathbb{R}^{d_2}$

$$\langle \vec{a}_1^\intercal \mathbf{X}_1, \vec{a}_2^\intercal \mathbf{X}_2 \rangle = \vec{a}_1^\intercal \mathbf{X}_1 \mathbf{X}_2^\intercal \vec{a}_2 = \mathrm{Cov}(\vec{a}_1^\intercal \mathbf{X}_1, \vec{a}_2^\intercal \mathbf{X}_2) = \sqrt{\mathrm{Var}(\vec{a}_1^\intercal \mathbf{X}_1)\,\mathrm{Var}(\vec{a}_2^\intercal \mathbf{X}_2)}\,\mathrm{Corr}(\vec{a}_1^\intercal \mathbf{X}_1, \vec{a}_2^\intercal \mathbf{X}_2).$$

4.1 SVD of the concatenated data blocks

A naive integrative analysis approach is concatenating all data blocks and performing thresholded SVD/PCA. However, this approach generally won't be able to segment the joint and individual signals, because the first several singular vectors (principal components) of this SVD are influenced by both variance within data blocks and correlation between the data blocks. More precisely, let

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_K \end{bmatrix}$$

be the concatenation of all data blocks. Then often suitable thresholding, the singular vector pairs $\{(\vec{u}^{(i)}, \vec{v}^{(i)}) : i = 1, \cdots, r_{\mathbf{J}}\}$ from SVD/PCA of $\mathbf{X}$ are the sequence of solutions of the following

maximization problems

$$
\begin{aligned}
(\vec{u}^{(i)}, \vec{v}^{(i)}) = \underset{\vec{u}=(\vec{u}_1^\intercal,\cdots,\vec{u}_K^\intercal)^\intercal,\vec{v}}{\arg\max}\ \vec{u}^\intercal \mathbf{X}\vec{v} = \sum_{1\le k\le K} \vec{u}_k^\intercal \mathbf{X}_k \vec{v} \\
\text{subject to the constraints: } \vec{u} \perp \vec{u}^{(j)}, \vec{v} \perp \vec{v}^{(j)}, j = 1, \cdots, i-1 \\
\|\vec{u}\|^2 = \sum_{1\le k\le K} \|\vec{u}_k\|^2 = 1, \|\vec{v}\|^2 = 1
\end{aligned}
\tag{4.2}
$$

To understand how within block and between block variation appear in this, we use the bound on the objective function in optimization problem (4.2),

$$
\begin{aligned}
(\vec{u}^\intercal \mathbf{X}\vec{v})^2 \le \|\vec{u}^\intercal \mathbf{X}\|^2 \|\vec{v}\|^2 = \|\vec{u}^\intercal \mathbf{X}\|^2 \| \sum_{1\le k\le K} \vec{u}_k^\intercal \mathbf{X}_k\|^2 \\
&= \sum_{1\le k,l\le K} \vec{u}_k^\intercal \mathbf{X}_k \mathbf{X}_l^\intercal \vec{u}_l \\
&= \sum_{1\le k\le K} \|\vec{u}_k^\intercal \mathbf{X}_k\|^2 + 2 \sum_{1\le k<l\le K} \vec{u}_k^\intercal \mathbf{X}_k \mathbf{X}_l^\intercal \vec{u}_l \\
&= \sum_{1\le k\le K} \mathrm{Var}(\vec{u}_k^\intercal \mathbf{X}_k) + 2 \sum_{1\le k<l\le K} \langle \vec{u}_k^\intercal \mathbf{X}_k, \vec{u}_l^\intercal \mathbf{X}_l \rangle
\end{aligned}
\tag{4.3}
$$

Note that the first term of this bound captures within block variance and the second term captures covariance between block. For each $\vec{u}$, equality in (4.3) can be achieved for some $\vec{v}$. Thus for the left singular vector $\vec{u}^{(i)}$, we have

$$
\begin{aligned}
\vec{u}^{(i)} = \begin{bmatrix} \vec{u}_1^{(i)} \\ \vdots \\ \vec{u}_K^{(i)} \end{bmatrix} = \underset{\vec{u}=(\vec{u}_1^\intercal,\cdots,\vec{u}_K^\intercal)^\intercal}{\arg\max} \sum_{1\le k\le K} \mathrm{Var}(\vec{u}_k^\intercal \mathbf{X}_k) + 2 \sum_{1\le k<l\le K} \langle \vec{u}_k^\intercal \mathbf{X}_k, \vec{u}_l^\intercal \mathbf{X}_l \rangle \\
\text{subject to the constraints: } \vec{u} \perp \vec{u}^{(j)}, j = 1, \cdots, i-1, \\
\|\vec{u}\|^2 = \sum_{1\le k\le K} \|\vec{u}_k\|^2 = 1.
\end{aligned}
\tag{4.4}
$$

As we can see in optimization problem (4.4), the objective function is affected by both variance within data blocks (first term) and covariance between data blocks (second term). Thus this approach can be driven by either strong within block variation in the first term (the individual

components of AJIVE), or by covariance across block in the second term (the joint components of AJIVE).

Figure 4.1 shows the results for three choices of rank. The rank-2 approximation essentially captures the joint variation component and the individual variation component of X, but the Y components are hard to interpret. The bottom 2000 rows show the joint variation but the top half of Y reveals signal from the individual component of X. One might hope that the Y individual components would show up in the rank-3 and rank-4 approximations. However, because the noise in the X matrix is so large, a random noise component from X dominates the Y signal, so the important latter component disappears from this low-rank representation unlike the AJIVE result in Figure 3.3. In this example, this naive approach completely fails to give a meaningful joint analysis.

4.2 Partial least squares (PLS)

The two-blocks PLS, invented by Wold (1975), finds linear combinations of rows of two matrices, for example $\mathbf{X}_1$ and $\mathbf{X}_2$, which maximize their sample covariance. More precisely, the PLS identifies a set of pairs of principal vectors, indexed by i, obtained sequentially from the following maximization problems:

$$\begin{aligned} \{\vec{a}_1^{(i)}, \vec{a}_2^{(i)}\} &= \underset{\vec{a}_1 \in \mathbb{R}^{d_1}, \vec{a}_2 \in \mathbb{R}^{d_2}}{\arg\max} \langle \vec{a}_1^\intercal \mathbf{X}_1, \vec{a}_2^\intercal \mathbf{X}_2 \rangle \\ \text{subject to the constraints: } & \|\vec{a}_1\| = 1, \|\vec{a}_2\| = 1, \\ & \langle \vec{a}_1^\intercal \mathbf{X}_1, (\vec{a}_1^{(j)})^\intercal \mathbf{X}_1 \rangle = 0, \ \langle \vec{a}_2^\intercal \mathbf{X}_2, (\vec{a}_2^{(j)})^\intercal \mathbf{X}_2 \rangle = 0, \ j = 1, \cdots, i-1. \end{aligned} \tag{4.5}$$

Unlike AJIVE, the directions from PLS are influenced by both variance within data blocks and correlation between data blocks. In particular, if the signal strength of the individual structure is sufficiently large it will be classified as a joint structure by being found ahead of the joint structure in the AJIVE sense.

Figure 4.2 presents the PLS approximations with different numbers of components selected. PLS completely fails to separate joint and individual components. Instead it provides mixtures of

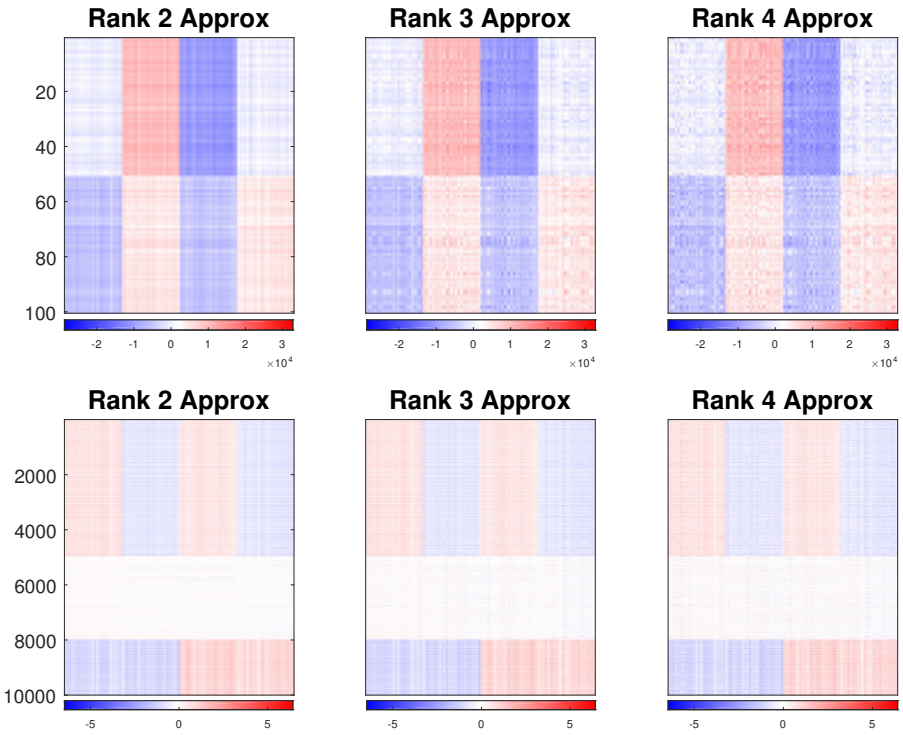


Figure 4.1: Shows the concatenation SVD approximation of each block for rank 2 (left), 3 (center) and 4 (right). Although block X has a relatively accurate approximation when the rank is chosen as 2, the individual pattern in block Y has never been captured due to the heterogeneity between X and Y.

the joint, and some of the individual components. Increasing the rank of the PLS approximation only includes more noise.

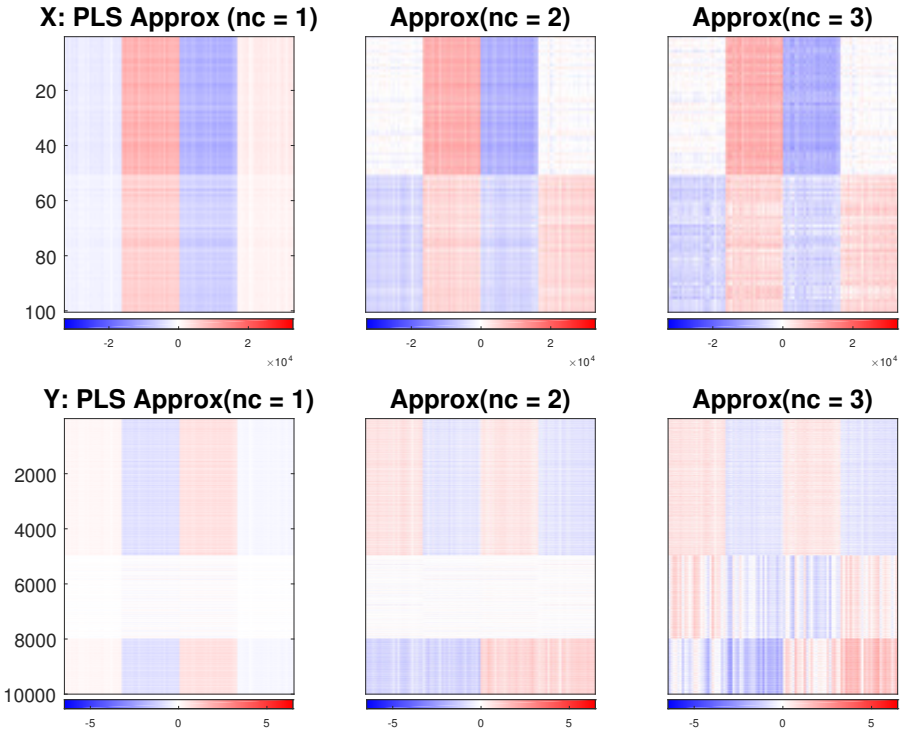


Figure 4.2: PLS approximations of each block for numbers of components as 1 (left), 2 (center) and 3 (right). PLS fails to distinguish the joint and individual variation structure.

multi-block PLS For the general case of multi-block, several approaches to multi-block PLS (mPLS) have been developed. A comprehensive summary of these generalizations can be found in Hanafi and Kiers (2006). A typical form of mPLS is the *MAXDIFF* criterion with equal sums of squares for weight vectors. For the ith set of weight vectors $\{\vec{a}_1^{(i)}, \cdots, \vec{a}_K^{(i)}\}$, where $\vec{a}_i \in \mathbb{R}^{d_k}$,

$$\begin{aligned} \{\vec{a}_1^{(i)}, \cdots, \vec{a}_K^{(i)}\} = \underset{\vec{a}_1, \cdots, \vec{a}_K}{\arg\max} \sum_{1 \le k < l \le K} \langle \vec{a}_k^\intercal \mathbf{X}_k, \vec{a}_l^\intercal \mathbf{X}_l \rangle \\ \text{subject to the constraints: } \|\vec{a}_k\|^2 = 1, \vec{a}_k^\intercal \mathbf{X}_k \perp (\vec{a}_k^{(j)})^\intercal \mathbf{X}_k \\ j = 1, \cdots, i-1, k = 1, \cdots, K. \end{aligned} \tag{4.6}$$

This version of mPLS is related to the SVD of the concatenated data block in Section 4.1. In fact, the objective function in the optimization problem (4.6) is the second term of the objective function in the optimization problem (4.4). And the optimization problem (4.4) is referred as the *MAXBET* criterion, in Hanafi and Kiers (2006), with total sum of squares of the weight vectors equal to 1. Both MAXDIFF and MAXBET are affected by within-block variation and correlation among data blocks, while MAXBET puts more emphasis on describing within-block variation. Taking into account within-block variance, most of the variation found by the solutions of MAXDIFF and MAXBET comes from the true signal subspace row($\mathbf{A}_k$). However, they feel both the joint and individual components, and thus tend not to effectively segment the joint space from each data block.

4.3 Principal angle analysis (PAA)

Principal angle analysis for two blocks (Hotelling, 1936) plays an essential role in AJIVE to segment the joint and individual variations from each data block. Let $\mathcal{X}$ and $\mathcal{Y}$ be two subspaces in $\mathbb{R}^n$. Assume $\dim(\mathcal{X}) = l \le \dim(\mathcal{Y}) = m \le n$. And let $\mathbf{Q}_\mathcal{X}$ and $\mathbf{Q}_\mathcal{Y}$ be the corresponding orthonormal basis matrices of $\mathcal{X}$ and $\mathcal{Y}$. Then the principal angles between $\mathcal{X}$ and $\mathcal{Y}$, $\{\theta_1, \cdots, \theta_l\}$,

are defined by $\theta_i = \arccos(\vec{u}_i^\intercal \vec{v}_i)$, where $\{\vec{u}_i, \vec{v}_i\}$ are solved sequentially for $i = 1, \cdots, l$

$$\begin{aligned} \{\vec{u}_i, \vec{v}_i\} &= \underset{\vec{u}\in\mathcal{X}, \vec{v}\in\mathcal{Y}}{\arg\max} \operatorname{Cov}(\vec{u}, \vec{v}) \\ \text{subject to the constraints: } & \|\vec{u}\| = \|\vec{v}\| = 1 \\ & \vec{u}^\intercal \vec{u}_j = 0, \vec{v}^\intercal \vec{v}_j = 0, j = 1, \cdots, i-1 \end{aligned} \tag{4.7}$$

The corresponding paired vectors $\{\vec{u}_i, \vec{v}_i\}$ are called the ith pair of *principal vectors*. By definition, it is easy to see

$$0 \le \theta_1 \le \cdots \le \theta_l \le \pi/2$$

When $l = m = 1$, the principal angle is the angle between the two lines $\mathcal{X}$ and $\mathcal{Y}$. Moreover, suppose $\dim(\mathcal{X} \cap \mathcal{Y}) = d$, it can be shown that the first d principal angles are 0,

$$0 = \theta_1 = \cdots = \theta_d < \theta_{d+1} \le \cdots \le \theta_l \le \pi/2$$

When $l = m$, the largest principal angle is related to the distance between subspaces $\mathcal{X}$ and $\mathcal{Y}$ defined by

$$\rho(\mathcal{X}, \mathcal{Y}) = \sin(\theta_l) = \|\mathbf{P}_\mathcal{X} - \mathbf{P}_\mathcal{Y}\| = \|(\mathbf{I} - \mathbf{P}_\mathcal{X})\mathbf{P}_\mathcal{Y}\| = \|(\mathbf{I} - \mathbf{P}_\mathcal{Y})\mathbf{P}_\mathcal{X}\|$$

where $\mathbf{P}_\mathcal{X} = \mathbf{Q}_\mathcal{X}\mathbf{Q}_\mathcal{X}^\intercal$ and $\mathbf{P}_\mathcal{Y} = \mathbf{Q}_\mathcal{Y}\mathbf{Q}_\mathcal{Y}^\intercal$ are the corresponding projection matrices of $\mathcal{X}$ and $\mathcal{Y}$. In the general case that $l \ne m$, the function $\rho(\mathcal{X}, \mathcal{Y})$ is a pseudometric on subspaces.

Accordingly, the similarity/closeness between $\mathcal{X}$ and $\mathcal{Y}$ can be written as

$$\|\mathbf{P}_\mathcal{X}\mathbf{P}_\mathcal{Y}\| = \cos(\theta_l)$$

As seen in Section 4.4, this quantity is the largest canonical correlation between $\mathcal{X}$ and $\mathcal{Y}$.

The following theorem in Björck and Golub (1973) gives a useful computation algorithm of *principal angle analysis*(PAA).

Theorem 2. *Let the columns of $\mathbf{Q}_\mathcal{X} \in \mathbb{R}^{n\times l}$ and $\mathbf{Q}_\mathcal{Y} \in \mathbb{R}^{n\times m}$ be orthonormal bases for $\mathcal{X}$ and $\mathcal{Y}$ respectively, $l \le m$ and let*

$$\sigma_1 \ge \cdots \ge \sigma_l \ge 0$$

be the singular values of $\mathbf{M} = \mathbf{Q}_{\mathcal{X}}^{\intercal}\mathbf{Q}_{\mathcal{Y}} = \mathbf{U}_{\mathbf{M}}\mathbf{\Sigma}_{\mathbf{M}}\mathbf{V}_{\mathbf{M}}^{\intercal} \in \mathbb{R}^{l\times m}$.

Then for $i = 1, \cdots, l$, *the principal angle* θ_i, *and the corresponding principal vectors* $\{\vec{u}_i, \vec{v}_i\}$ *are given by*

$$\begin{aligned} \vec{u}_i &= ith \text{ column of } \mathbf{Q}_{\mathcal{X}}\mathbf{U}_{\mathbf{M}} = \mathbf{Q}_{\mathcal{X}}\mathbf{U}_{\mathbf{M},i}, \\ \vec{v}_i &= ith \text{ column of } \mathbf{Q}_{\mathcal{Y}}\mathbf{V}_{\mathbf{M}} = \mathbf{Q}_{\mathcal{Y}}\mathbf{V}_{\mathbf{M},i}. \\ \cos\theta_i &= \sigma_i = \vec{u}_i^{\intercal}\vec{v}_i. \end{aligned} \tag{4.8}$$

and

$$\sigma_1 = \cdots = \sigma_d = 1 > \sigma_{d+1} \geq \sigma_k \iff \dim \mathcal{X} \cap \mathcal{Y} = d \tag{4.9}$$

AJIVE utilizes an alternative approach to PAA for computation of principal angles, principal vectors and a basis of the joint space, which is based on the following theory combining the results from Miao and Ben-Israel (1992) and Horn and Johnson (2012).

Theorem 3. *Let the columns of* $\mathbf{Q}_{\mathcal{X}} \in \mathbb{R}^{n\times l}$ *and* $\mathbf{Q}_{\mathcal{Y}} \in \mathbb{R}^{n\times m}$ *be orthonormal bases for* $\mathcal{X}$ *and* $\mathcal{Y}$ *respectively. Assume* $l \leq m$ *and* $(l+m) < n$. *Denote*

$$\mathbf{Q} := \begin{bmatrix} \mathbf{Q}_{\mathcal{X}}^{\intercal} \\ \mathbf{Q}_{\mathcal{Y}}^{\intercal} \end{bmatrix} \in \mathbb{R}^{(l+m)\times n}.$$

Let the non-zero singular values of

$$\mathbf{M} = \mathbf{Q}_{\mathcal{X}}^{\intercal}\mathbf{Q}_{\mathcal{Y}} = \overbrace{\mathbf{U}_{\mathbf{M}}}^{l\times l}\overbrace{\mathbf{\Sigma}_{\mathbf{M}}}^{l\times m}\overbrace{\mathbf{V}_{\mathbf{M}}^{\intercal}}^{m\times m} \in \mathbb{R}^{l\times m}$$

be

$$\sigma_1 \geq \cdots \geq \sigma_l > 0.$$

And denote $\mathbf{V}_{\mathbf{M}} = (\overbrace{\mathbf{V}_1}^{l}, \overbrace{\mathbf{V}_2}^{m-l})$. *Then the singular values of* $\mathbf{Q}$ *are*

$$\sqrt{1+\sigma_1}, \cdots, \sqrt{1+\sigma_l}, \underbrace{1, \cdots, 1}_{m-l}, \sqrt{1-\sigma_l}, \cdots, \sqrt{1-\sigma_1}$$

and $\mathbf{Q} = \mathbf{U_Q}\mathbf{\Sigma_Q}\mathbf{V_Q^\intercal}$ *with the diagonal matrix* $\mathbf{\Sigma_Q}$ *given by*

$$\mathbf{\Sigma_Q} = diag\{\sqrt{1+\sigma_1}, \cdots, \sqrt{1+\sigma_l}, \underbrace{1, \cdots, 1}_{m-l}, \sqrt{1-\sigma_l}, \cdots, \sqrt{1-\sigma_1}\} \in \mathbb{R}^{(l+m)\times(l+m)}$$

and the orthogonal matrices $\mathbf{U_Q}$ *and* $\mathbf{V_Q}$ *given by*

$$\mathbf{U_Q} = \begin{pmatrix} \hat{\mathbf{U}}_\mathbf{M} & \mathbf{0}_{l,m-l} & \hat{\mathbf{U}}_\mathbf{M} \\ \hat{\mathbf{V}}_\mathbf{M} & \mathbf{V}_2 & -\hat{\mathbf{V}}_\mathbf{M} \end{pmatrix} \in \mathbb{R}^{(m+l)\times(m+l)}$$

$$\mathbf{V_Q} = \mathbf{\Sigma_Q^{-1}} \begin{pmatrix} \mathbf{Q}_\mathcal{X}\hat{\mathbf{U}}_\mathbf{M} + \mathbf{Q}_\mathcal{Y}\hat{\mathbf{V}}_\mathbf{M} & \mathbf{Q}_\mathcal{Y}\mathbf{V}_2 \end{pmatrix} \in \mathbb{R}^{(l+m)\times n}$$

where

$$\hat{\mathbf{U}}_\mathbf{M} = \mathbf{U_M}/\sqrt{2}, \hat{\mathbf{V}}_\mathbf{M} = \mathbf{V}_1/\sqrt{2}$$

Notice that in Theorem 3, the first l columns of $\mathbf{U_Q}$ is the concatenation of the first l columns of $\mathbf{U_M}$ and $\mathbf{V_M}$ (scaled by $\sqrt{2}$). Thus we can obtain both principal angles and principal vectors from this approach. However, AJIVE constructs the basis of the joint space from the first l columns of the right singular matrix $\mathbf{V_Q}$, which is essentially the average of the corresponding principal vectors.

In the context of AJIVE, identifying the joint basis matrix of $\mathbf{A}_1$ and $\mathbf{A}_2$, i.e., the intersection of $\mathcal{Q}_1$ and $\mathcal{Q}_2$, is equivalent to finding the principal vectors corresponding to 0 angles. In practice, due to the existence of noise, PAA is applied on low rank approximations of $\mathbf{X}_1$ and $\mathbf{X}_2$, i.e. $\tilde{\mathbf{A}}_1$ and $\tilde{\mathbf{A}}_2$, and the principal vectors corresponding to small angles are found by thresholding based on the generalized $\sin\theta$ theorem, see Section 3.3.2.2 of Chapter 3 for details.

4.4 Canonical correlation analysis (CCA)

Similar to PLS, the two-block CCA finds linear combinations of rows of $\mathbf{X}_1$ and $\mathbf{X}_2$ maximizing their sample correlation. In particular, CCA identifies a set of pairs of canonical vectors obtained

sequentially from the optimization problem:

$$\begin{aligned} \{\vec{a}_1^{(i)}, \vec{a}_2^{(i)}\} &= \underset{\vec{a}_1 \in \mathbb{R}^{d_1}, \vec{a}_2 \in \mathbb{R}^{d_2}}{\arg\max} \langle \vec{a}_1^\intercal \mathbf{X}_1, \vec{a}_2^\intercal \mathbf{X}_2 \rangle \\ \text{subject to the constraints: } & \|\vec{a}_1^\intercal \mathbf{X}_1\| = 1, \|\vec{a}_2^\intercal \mathbf{X}_2\| = 1 \\ & \langle \vec{a}_1^\intercal \mathbf{X}_1, (\vec{a}_1^{(j)})^\intercal \mathbf{X}_1 \rangle = 0, \ \langle \vec{a}_2^\intercal \mathbf{X}_2, (\vec{a}_2^{(j)})^\intercal \mathbf{X}_2 \rangle = 0, \ j = 1, \cdots, i-1. \end{aligned} \tag{4.10}$$

This form makes the relationship between (4.5) and (4.10) clear (differing only constraining the loading loadings $\|\vec{a}_k\| = 1$ versus the scores $\|\vec{a}_k^\intercal \mathbf{X}_k\| = 1$) and is equivalent to the usual formulation of optimizing the correlation.

There is an important relationship between CCA and PAA (Hotelling, 1936; Björck and Golub, 1973), i.e., if $\rho_i = \langle (\vec{a}_1^{(i)})^\intercal \mathbf{X}_1, (\vec{a}_2^{(i)})^\intercal \mathbf{X}_2 \rangle$ is the ith canonical correlation, $\rho_i = \cos(\theta_i)$, where θ_i is the ith principal angle between row spaces of $\mathbf{X}_1$ and $\mathbf{X}_2$. The principal vector pairs $\{\vec{x}_{1,i}, \vec{x}_{2,i}\} = \{\mathbf{X}_1^\intercal \vec{a}_1^{(i)}, \mathbf{X}_2^\intercal \vec{a}_2^{(i)}\}$ are often obtained through SVD of $\mathbf{V}_{\mathbf{X}_1}^\intercal \mathbf{V}_{\mathbf{X}_2}$. In particular, let $\vec{u}_{\mathbf{X}_1,i}, \vec{u}_{\mathbf{X}_2,i}$ be the ith left and right singular vectors of $\mathbf{V}_{\mathbf{X}_1}^\intercal \mathbf{V}_{\mathbf{X}_2}$. Then, the ith pair of principal vectors are

$$\vec{x}_{1,i} = \mathbf{V}_{\mathbf{X}_1} \vec{u}_{\mathbf{X}_1,i}, \quad \vec{x}_{2,i} = \mathbf{V}_{\mathbf{X}_2} \vec{u}_{\mathbf{X}_2,i}.$$

An issue with CCA of high-dimensional data is related to the fact that CCA is interested in the canonical vectors $\vec{a}_i$ rather than the principal vectors $\vec{x}_i$. In particular, when $d_1 > n, d_2 > n$, the values of $\vec{a}_i$ in (4.10) are not identifiable due to the singularity of $\mathbf{X}_1\mathbf{X}_1^\intercal$ and $\mathbf{X}_2\mathbf{X}_2^\intercal$. Several approaches have been taken to solve this problem. One approach is to use the Moore-Penrose pseudo inverse in place of the inverse of $\mathbf{X}_1\mathbf{X}_1^\intercal$ and $\mathbf{X}_2\mathbf{X}_2^\intercal$. A second approach is to add a ridge penalty on $\mathbf{X}_1\mathbf{X}_1^\intercal$ and $\mathbf{X}_2\mathbf{X}_2^\intercal$ (Vinod, 1976). A third approach called *penalized CCA* is to add penalty functions on $\{\vec{a}_1^{(i)}, \vec{a}_2^{(i)}\}$, such as an ℓ_1 penalty (Parkhomenko et al., 2007; Lê Cao et al., 2009), an elastic net (Waaijenborg et al., 2008) or a fused lasso (Witten et al., 2009). Another approach called *diagonal penalized CCA* is to replace $\mathbf{X}_1\mathbf{X}_1^\intercal$ and $\mathbf{X}_2\mathbf{X}_2^\intercal$ by $\text{diag}(\mathbf{X}_1\mathbf{X}_1^\intercal)$ and $\text{diag}(\mathbf{X}_2\mathbf{X}_2^\intercal)$ (Parkhomenko et al., 2009; Witten et al., 2009).

Another important issue with CCA, which is directly related to AJIVE, is that when $d_1 > n, d_2 > n$, CCA is generally driven by noise. Lee (2007); Samarov (2009); Lee (2016) study the asymptotic behavior of CCA in the high-dimension low sample size context and point out the

inconsistency phenomenon in this case. One solution to this issue, used by AJIVE and COBE, is to replace X_k by its low rank approximation $\tilde{\mathbf{A}}_k$, $k = 1, 2$. The ith principal vectors are $\vec{p}_i = \tilde{\mathbf{V}}_1 \vec{u}_{1,i}$, $\vec{q}_i = \tilde{\mathbf{V}}_2 \vec{u}_{2,i}$, where $\vec{u}_{j,i}$ is the ith singular vector of $\tilde{\mathbf{U}}_i$ of the SVD of $\tilde{\mathbf{V}}_1^\intercal \tilde{\mathbf{V}}_2$ respectively.

As described in Section 4.3, AJIVE uses an equivalent principal angle calculation based on SVD of $\mathbf{M} = [\tilde{\mathbf{V}}_1, \tilde{\mathbf{V}}_2]^\intercal = \mathbf{U}_\mathbf{M} \mathbf{\Sigma}_\mathbf{M} \mathbf{V}_\mathbf{M}^\intercal$ (Miao and Ben-Israel, 1992). AJIVE uses the transpose of the ith right singular vector, $\vec{v}_{\mathbf{M},i}$, as the estimated ith basis vector of the joint space, provided that the ith principal angle is smaller than the threshold derived in Section 3.3.3.1. Moreover, if the ith principal angle has a value distinct from other principal angles, then the ith left singular vector of $\mathbf{M}$ can be written as $\vec{u}_{\mathbf{M},i} = [\vec{u}_{1,i}^\intercal, \vec{u}_{2,i}^\intercal]^\intercal / \sqrt{2}$. Consequently

$$\vec{v}_{\mathbf{M},i} = \frac{1}{\sigma_{\mathbf{M},i}} \mathbf{M}^\intercal \vec{u}_{\mathbf{M},i} = \frac{1}{\sqrt{2}\sigma_{\mathbf{M},i}} (\tilde{\mathbf{V}}_1 \vec{u}_{1,i} + \tilde{\mathbf{V}}_2 \vec{u}_{2,i}) = \frac{1}{\sqrt{2}\sigma_{\mathbf{M},i}} (\vec{p}_i + \vec{q}_i).$$

This shows that the AJIVE direction $\vec{v}_{\mathbf{M},i}$ is the scaled sum (essentially the average) of the ith pair of principal vectors.

CCA applied to the low rank approximations $\tilde{\mathbf{A}}_k$ and AJIVE are therefore closely related. However, AJIVE provides one joint vector per two distinct principal vectors that by the virtue of being an average should be a better estimate of the joint space than either of the principal vectors. More importantly, AJIVE uses a theoretically sound threshold of the principal angles that allows us to segment individual and joint variation.

multiset CCA The AJIVE formulation allows for a natural extension to multi-block situations. Several approaches of Multiset Canonical Correlation Analysis (mCCA) have been developed as extensions of CCA (Horst, 1961; Kettenring, 1971; Nielsen, 2002). There is no general consensus on which of these extensions is preferable. We point out that AJIVE is closely related to one of the mCCA approaches discussed in (Nielsen, 2002).

This version of mCCA is defined using the optimization problem for the ith set of canonical vectors $\{\vec{a}_1^{(i)}, \cdots, \vec{a}_K^{(i)}\}$ and corresponding principal vectors (also called canonical variables)

$\{\mathbf{X}_1^\intercal \vec{a}_1^{(i)}, \cdots, \mathbf{X}_K^\intercal \vec{a}_K^{(i)}\}$:

$$\begin{aligned} \{\vec{a}_1^{(i)}, \cdots, \vec{a}_K^{(i)}\} &= \underset{\vec{a}_1,\cdots,\vec{a}_K}{\arg\max} \sum_{1\le k,l\le K} \langle \vec{a}_k^\intercal \mathbf{X}_k, \vec{a}_l^\intercal \mathbf{X}_l \rangle \\ \text{subject to the constraints: } & \sum_{k=1}^{K} \|\vec{a}_k^\intercal \mathbf{X}_k\|_2^2 = 1, \\ & \langle \vec{a}_k^\intercal \mathbf{X}_k, (\vec{a}_k^{(j)})^\intercal \mathbf{X}_k \rangle = 0, \quad k = 1, \cdots, K, \quad j = 1, \cdots, i-1. \end{aligned} \tag{4.11}$$

Notice that the constraint in (4.11) is different than the perhaps more natural $\|\vec{a}_k^\intercal \mathbf{X}_k\|_2^2 = 1$ for all k.

If the ith singular value corresponding to the AJIVE direction $\vec{v}_{\mathbf{M},i}$ has a value distinct from other singular values in the AJIVE SVD, then calculations similar to the two block case show that the ith basis vector of the joint space from AJIVE

$$\vec{v}_{\mathbf{M},i} = \frac{1}{\sigma_{\mathbf{M},i}} \sum_{k=1}^{K} \mathbf{X}_k^\intercal \vec{a}_k^{(i)}$$

is again a scaled sum (essentially average) of the corresponding canonical variables.

4.5 Flag mean

The Grassmann manifold $\mathrm{Gr}(\mathcal{V}, p)$ is a manifold whose points parametrize the subspaces of dimension p inside the vector space $\mathcal{V}$. Typically, we denote by $\mathrm{Gr}(n, p)$ the Grassmann manifold of p dimensional subspaces of $\mathbb{R}^n$. Let $(\mathbb{R}^{d\times n})^\circ$ be the open submanifold of full row rank $d \times n$ matrices. For each $\mathbf{Y} \in (\mathbb{R}^{d\times n})^\circ$, let $[\mathbf{Y}]$ denote the transpose of the row space of $\mathbf{Y}$. Suppose that $[\mathbf{X}] \in \mathrm{Gr}(n, p_1), [\mathbf{Y}] \in \mathrm{Gr}(n, p_2)$ for $p_1 < p_2$. Let $d_{pF}([\mathbf{X}], [\mathbf{Y}]) : \mathrm{Gr}(n, p_1) \times \mathrm{Gr}(n, p_2) \to \mathbb{R}$ be a pseudometric, which is defined as the ℓ_2-norm of the vector of the sines of the p_1 principal angles between $[\mathbf{X}]$ and $[\mathbf{Y}]$, i.e.

$$d_{pF}([\mathbf{X}], [\mathbf{Y}]) = \sqrt{\sum_{i=1}^{p_1} (\sin\theta_i)^2}$$

where $0 \le \theta_1 \le \cdots \le \theta_{p_1} \le \pi/2$ are principal angles between $[\mathbf{X}]$ and $[\mathbf{Y}]$.

A set of matrices of the same dimension and rank can be treated as a collection of points on a single Grassman manifold. Draper et al. (2014) proposes a *flag mean* representation for data

consisting of subspaces of $\mathbb{R}^n$ of differing dimensions, i.e. a data cloud living on a disjoint union of Grassmann manifolds. Let $\mathcal{D} = \{[\mathbf{X}_i]\}_{i=1}^K$ be a finite collection of subspaces of $\mathbb{R}^n$. The flag mean is a nested sequence of vector spaces that best fits the data according to an optimization criterion based on the projection Frobenius norm. In particular,

$$\begin{gathered}[\vec{u}^{(i)}] := \underset{[\vec{u}]\in \mathrm{Gr}(n,1)}{\arg\min} \sum_{[\mathbf{X}_k]\in\mathcal{D}} d_{pF}([\vec{u}],[\mathbf{X}_k])^2 \\ \text{subject to the constraints: } [\vec{u}] \perp [\vec{u}^{(j)}], j < i \end{gathered} \tag{4.12}$$

This leads to the set $\{[\vec{u}^{(1)}], \cdots, [\vec{u}^{(r)}]\}$ where r denotes the dimension of the span of the elements in $\mathcal{D}$. If the non-zero singular values of $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_K]$ are all distinct, then the flag is:

$$\|\mu_{pF}\|(\mathcal{D}) := [\vec{u}^{(1)}] \subset [\vec{u}^{(1)}|\vec{u}^{(2)}] \subset \cdots \subset [\vec{u}^{(1)}|\ldots|\vec{u}^{(r)}]$$

where $|$ denote horizontal concatenation. The flag mean is a sequence of subspaces, where the ith subspace is generated by $[\vec{u}^{(1)}|\ldots|\vec{u}^{(i)}]$. Note in the case $K = 1$, the flag mean is the sequence of subspaces generated by the PCA eigenvectors.

Draper et al. (2014) show that $\vec{u}^{(i)}$, the solution in optimization problem (4.12), is also the average of the ith set of canonical variables in mCCA derived from the optimization problem (4.11). This is closely related to AJIVE since the joint space determined there is the flag mean of rank $r_{\mathbf{J}}$. An important contribution of AJIVE is determination of $r_{\mathbf{J}}$ using perturbation theory.

4.6 Common Orthogonal Basis Extraction (COBE)

Zhou et al. (2016) proposed a compelling optimization problem for finding the common orthogonal basis (COBE). It is based on iteratively solving

$$\begin{gathered}\bar{a}_i = \underset{\bar{a},\ \vec{z}_{i,k}, k=1,\cdots,K}{\arg\min} \sum_{k=1}^{K} \|\tilde{\mathbf{V}}_k \vec{z}_{i,k} - \bar{a}\|^2 \\ \text{subject to the constraints: } \|\bar{a}\|_2 = 1, \langle \bar{a}, \bar{a}_j \rangle = 0,\ j = 1, \ldots, i-1. \end{gathered} \tag{4.13}$$

To compare COBE to AJIVE we first simplify the objective function of (4.13) to

$$\begin{aligned}\sum_{k=1}^{K} \|\tilde{\mathbf{V}}_k \vec{z}_{i,k} - \bar{a}_i\|_2^2 &= \sum_{k=1}^{K} \|\tilde{\mathbf{V}}_k \vec{z}_{i,k}\|_2^2 + K\|\bar{a}_i\|_2^2 - 2\sum_{k=1}^{K} \langle \tilde{\mathbf{V}}_k z_{i,k}, \bar{a}_i \rangle \\ &= \|\vec{z}_i\|_2^2 + K\|\bar{a}_i\|_2^2 - 2\vec{z}_i^{\intercal} M \bar{a}_i.\end{aligned}$$

where $\vec{z}_i = [\vec{z}_{i,1}^{\intercal}, \cdots, \vec{z}_{i,K}^{\intercal}]^{\intercal}$. If we fix the value of $\|\vec{z}_i\|$ we see that the solution to the optimization problem (4.13) is the same as SVD of $\mathbf{M}$ with $\bar{a}_i = \vec{v}_{\mathbf{M},i}$. Moreover this solution is invariant in $\|\vec{z}_i\|$.

Thus the optimization problem (4.13) gives the same result as AJIVE. However, because AJIVE uses well optimized SVD rather than a heuristic iteration algorithm, AJIVE is much faster than the COBE algorithm. Moreover, COBE lacks any principle based standard on how to choose the threshold for selecting the joint space.

We finally remark that the Zhou et al. (2016) method COBE correctly segments the toy example. However it takes significantly (39 times) longer time than AJIVE to do so.

CHAPTER 5

Perturbation Analysis For A Given Direction

5.1 Introduction

The AJIVE method (Feng et al., 2017) introduced in Chapter 3 is a major improvement over the original JIVE method (Lock et al., 2013). Relative to existing integrative methods, AJIVE combines (generalized) principal angle analysis with perturbation theory, which makes it much more effective in capturing the joint and individual variation within each data block. However, there are still some limitations of the current AJIVE implementation.

First, the framework of perturbation analysis in AJIVE is too conservative. The perturbation analysis of approximation accuracy of signal space extraction and principal angle analysis is based on the Wedin bound (Wedin, 1972). As seen in Figure 3.6, the Wedin bound is very conservative for long matrices (the number of rows is much larger than the number of columns) in the estimation of the discrepancy between the row spaces of signal and estimated signal. This is due to the fact that Wedin's bound is uniform for both the left and right singular spaces. Cai et al. (2018) construct separate rate-optimal perturbation bounds for the left and right singular subspaces, which provides improvement of the Wedin bound in some cases. However, their perturbation bounds cannot be directly estimated from the data matrix.

Second, the framework of perturbation analysis in AJIVE is "subspace oriented" rather than "vector oriented". In Step 2 of AJIVE in Section 3.3.3, for each pair (group) of principal vectors, their principal angle (corresponding singular value in $\mathbf{M}$) is compared to the angle bound from Lemma 2 (Lemma 3) to determine whether they correspond to a joint space basis vector. This bound is based on Wedin bound, which is uniformed for all principal vectors in each signal space. Our analysis shows that the vector in the signal spaces would have less perturbation when the signal matrix has larger variance in that direction. Thus improved statistical inference is available for doing a separate perturbation analysis for each AJIVE direction.

Third, AJIVE lacks principled and efficient guideline for rank selection at Step 1 as in Section 3.3.2 of Chapter 3. Diagnostic plot in Section 3.3.3 provides useful information for rank selection of signal spaces. However, it can be very inefficient to check all rank combinations of each data block in this way. Recent results from Shabalin and Nobel (2013) and Gavish and Donoho (2017) provide insights of signal extraction in Step 1.

Last but not the least, the AJIVE algorithm can only identify the variation shared by all data blocks. The long term goal to develop an efficient algorithm which can capture partially shared structure across some but not all data blocks.

All these considerations motivate us to develop a new framework of perturbation analysis for a given direction. There is some literature discussing the perturbation angle for singular vectors, see Benaych-Georges and Nadakuditi (2012), Gavish and Donoho (2014), and Fan et al. (2016). But for our long term goal of identifying partially shared structure, we need to develop perturbation analysis for any given direction; we hope to answer whether a given direction could lie in the signal space. To the best of our knowledge, there is no existing literature on this.

The rest of this chapter is organized as follows. Section 5.2 discusses our framework of perturbation analysis for a given direction. Section 5.2.1 introduces the settings of the population model. Section 5.2.2 discusses the signal space extraction based on the general singular value shrinkage estimator. Section 5.2.3 reviews some useful asymptotic results from the random matrix theory. Section 5.2.4 introduces the methodology to infer whether a direction belongs to the row space of the underlying signal matrix. Section 5.3 describes the computation algorithm to estimate the angle range. Section 5.4 shows the simulation results under different settings. Several attempts to adjust the AJIVE directions are also discussed in Section 5.4.3. Section 5.5 studies the application of this perturbation analysis framework on the TCGA dataset in Section 3.4.1.

5.2 Perturbation analysis framework for a given direction

5.2.1 Population model

A matrix $\mathbf{X}$ of size $d \times n$ is a data block for study. Using terminology from Marron and Alonso (2014b), the columns are regarded as data objects, one vector of d measurements for each experimental subject, while rows are considered as features. The data matrix $\mathbf{X}$ is modeled as a

low-rank true underlying signal matrix $\mathbf{A}$ perturbed by an additive noise matrix $\mathbf{E}$, viz.

$$\mathbf{X} = \mathbf{A} + \mathbf{E} = \mathbf{A} + \sigma\mathbf{Z}. \tag{5.1}$$

The entries of $\mathbf{Z}$ are assumed to be i.i.d. samples following a distribution with zero mean, unit variance and finite fourth moment. Moreover, we assume the distribution of $\mathbf{Z}$ is *orthogonal invariant*, i.e. for any orthogonal matrices $\mathbf{O}_1 \in \mathbb{R}^{d\times d}, \mathbf{O}_2 \in \mathbb{R}^{n\times n}$, the distributions of $\mathbf{Z}$ and $\mathbf{O}_1\mathbf{Z}\mathbf{O}_2$ are the same. This assumption can be replaced by the relationship between $\mathbf{A}$ and $\mathbf{E}$ is "isotropic".

For a given direction $\mathbf{a}^\star \in \mathbb{R}^n$, we hope to make inference whether $\mathbf{a}^\star$ could be in the row space of the signal matrix $\mathbf{A}$. This is based on an estimator of the angle between $\mathbf{a}^\star$ and row($\mathbf{A}$). Denote this angle and the projection on row($\mathbf{A}$) as θ and $\mathbf{a}$ respectively. Let $\mathbf{V}$ be the right singular matrix of $\mathbf{A}$. Then,

$$\mathbf{a} = \mathbf{V}\mathbf{V}^\top\mathbf{a}^\star, \theta = \arccos(\frac{\mathbf{a}^\top\mathbf{a}^\star}{\|\mathbf{a}\|\|\mathbf{a}^\star\|}).$$

This estimation is challenging since $\mathbf{A}$ (and $\mathbf{V}$) is usually unobservable in practice.

Suppose $\hat{\mathbf{A}}$ is an estimator of the signal matrix $\mathbf{A}$ from the data matrix $\mathbf{X}$. Then useful information comes from calculating the angle between $\mathbf{a}^\star$ and row($\hat{\mathbf{A}}$). Denote this angle and the projection of $\mathbf{a}^\star$ on row($\hat{\mathbf{A}}$) as $\hat{\theta}$ and $\hat{\mathbf{a}}$ respectively, and let $\hat{\mathbf{V}}$ be the right singular matrix of $\hat{\mathbf{A}}$. Then,

$$\hat{\mathbf{a}} = \hat{\mathbf{V}}\hat{\mathbf{V}}^\top\mathbf{a}^\star, \hat{\theta} = \arccos(\frac{\hat{\mathbf{a}}^\top\mathbf{a}^\star}{\|\hat{\mathbf{a}}\|\|\mathbf{a}^\star\|})$$

Due to the presence of the noise matrix $\mathbf{E}$, there exists some discrepancy between θ and $\hat{\theta}$. In order to quantify this discrepancy, the perturbation effects of $\mathbf{a}$ and $\hat{\mathbf{a}}$ are estimated separately. We define the *direction specific perturbation angle* for $\mathbf{a}$, $\theta_1^\star$, which is the angle between $\mathbf{a}$ and row($\hat{\mathbf{A}}$). Similarly, the direction specific perturbation angle for $\hat{\mathbf{a}}$, $\theta_2^\star$, is defined as the angle between $\hat{\mathbf{a}}$ and row($\mathbf{A}$).

Then we can bound the range of θ, the angle between $\mathbf{a}^\star$ and row($\mathbf{A}$), with $\hat{\theta}$, $\theta_1^\star$, $\theta_2^\star$ by the following theorem.

Theorem 4. *Let $\mathbf{X} = \mathbf{A} + \mathbf{Z}$ be a data matrix of size $d \times n$, which is the sum of the signal matrix $\mathbf{A}$ and the noise matrix $\mathbf{Z}$. Suppose $\hat{\mathbf{A}}$ is an estimator of $\mathbf{A}$. For a given vector $\mathbf{a}^\star \in \mathbb{R}^n$, let $\theta, \hat{\theta}, \theta_1^\star$*

and $\theta_2^\star$ be defined as above. Then we have

$$(\hat{\theta} - \theta_1^\star)_+ \leq \theta \leq \hat{\theta} + \theta_2^\star. \tag{5.2}$$

Proof. Firstly, we want to prove $\hat{\theta} \leq \theta_1^\star + \theta$.

Let $\hat{\mathbf{a}}_1$ be the projection of $\mathbf{a}$ on row($\hat{\mathbf{A}}$). Then $\theta_1^\star$ is the angle between $\mathbf{a}$ and $\hat{\mathbf{a}}_1$. Applying Lemma 4, we get

$$\theta_1^\star + \theta = \measuredangle(\mathbf{a}, \hat{\mathbf{a}}_1) + \measuredangle(\mathbf{a}, \mathbf{a}^\star) \geq \measuredangle(\hat{\mathbf{a}}_1, \mathbf{a}^\star) \geq \measuredangle(\hat{\mathbf{a}}, \mathbf{a}^\star) = \hat{\theta}$$

since $\hat{\theta}$ is the smallest angle between vectors in row($\hat{\mathbf{A}}$) and $\mathbf{a}^\star$.

Similarly, we can show

$$\theta \leq \hat{\theta} + \theta_2^\star.$$

Combining these two equalities, we have proven the main result (5.2). □

Lemma 4. *Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ be three vectors in $\mathbb{R}^n$. Let $\theta_{i,j}$ be the angle between $\mathbf{v}_i$ and $\mathbf{v}_j$. Then, we have the following inequality:*

$$\theta_{1,2} \leq \theta_{1,3} + \theta_{2,3}.$$

Proof. First, we consider the special case that $\mathbf{v}_3$ lies in the subspace spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$. If $\mathbf{v}_3$ lies between $\mathbf{v}_1$ and $\mathbf{v}_2$, then

$$\theta_{1,2} = \theta_{1,3} + \theta_{2,3}.$$

If $\mathbf{v}_3$ lies outside of $\mathbf{v}_1$ and $\mathbf{v}_2$, then

$$\theta_{1,2} = |\theta_{1,3} - \theta_{2,3}| \leq \theta_{1,3} + \theta_{2,3}.$$

Next, we consider the general case that $\mathbf{v}_3$ lies outside of the subspace spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$. Denote the projection of $\mathbf{v}_3$ on the subspace spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$ as $\hat{\mathbf{v}_3}$. Analogously, we define the angle $\hat{\theta}_{i,3} = \measuredangle(\mathbf{v}_i, \hat{\mathbf{v}_3}), i = 1, 2$. Then, following the argument above, we have

$$\theta_{1,2} \leq \hat{\theta}_{1,3} + \hat{\theta}_{2,3}.$$

Moreover, the subspace spanned by $\mathbf{v}_3$ and $\hat{\mathbf{v}}_3$ is orthogonal to the subspace spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$, and $\hat{\mathbf{v}}_3$ is on the intersection of two subspaces. Then $\hat{\mathbf{v}}_3$ is also the projection of $\mathbf{v}_1$ on the subspace spanned by $\mathbf{v}_3$ and $\hat{\mathbf{v}}_3$. Then $\hat{\theta}_{1,3} = \measuredangle(\mathbf{v}_1, \hat{\mathbf{v}}_3) \leq \measuredangle(\mathbf{v}_1, \mathbf{v}_3) = \theta_{1,3}$ since the projection angle is the smallest angle between $\mathbf{v}_1$ and the subspace spanned by $\mathbf{v}_3$ and $\hat{\mathbf{v}}_3$. Similarly, we have $\hat{\theta}_{2,3} \leq \theta_{2,3}$. Thus,

$$\theta_{1,2} \leq \hat{\theta}_{1,3} + \hat{\theta}_{2,3} \leq \theta_{1,3} + \theta_{2,3}.$$

□

Therefore, in order to bound the range of θ, it is essential to propose good estimators $\hat{\mathbf{A}}$, $\theta_1^\star$ and $\theta_2^\star$. In Section 5.2.2, we will discuss the methodology of estimation of $\mathbf{A}$ based on results from Shabalin and Nobel (2013) and Gavish and Donoho (2017). Section 5.2.4 and 5.3 discuss the estimation of the direction specific perturbation angle for any given direction.

5.2.2 Signal space extraction

It is a ubiquitous challenge in modern scientific research to recover a low rank signal from a data matrix. For the additive noise model as in (5.1), the standard practice is to apply truncated Singular Value Decomposition (SVD) to $\mathbf{X}$. This is accomplished by writing the SVD of $\mathbf{X}$ in the form

$$\mathbf{X} = \begin{bmatrix} \hat{\mathbf{U}} & \hat{\mathbf{U}}_0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{\Sigma}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{\Sigma}}_0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{V}}^\top \\ \hat{\mathbf{V}}_0^\top \end{bmatrix} = \sum_{j=1}^{d \wedge n} \hat{\lambda}_j \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^\top,$$

where $\hat{r}$ is the estimator of $r = \operatorname{rank} \mathbf{A}$, $\hat{\mathbf{U}} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_{\hat{r}} \end{bmatrix}$ and $\hat{\mathbf{V}} = \begin{bmatrix} \hat{\mathbf{v}}_1 & \cdots & \hat{\mathbf{v}}_{\hat{r}} \end{bmatrix}$ are rank $\hat{r}$ left and right singular matrices respectively, and $\hat{\mathbf{\Sigma}} = \operatorname{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{r}}\}$. Then the conventional estimator of $\mathbf{A}$ is

$$\hat{\mathbf{A}}_{\hat{r}} = \hat{\mathbf{U}} \hat{\mathbf{\Sigma}} \hat{\mathbf{V}}^\top = \sum_{j=1}^{\hat{r}} \hat{\lambda}_j \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^\top. \tag{5.3}$$

Denote the corresponding SVD of $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top = \sum_{j=1}^{r} \lambda_j \mathbf{u}_j \mathbf{v}_j^\top.$$

Due to the additional energy contributed by the noise matrix $\mathbf{E}$, the singular values of $\hat{\mathbf{A}}_{\hat{r}}$ are usually inflated relative to $\mathbf{A}$, i.e.

$$\hat{\lambda}_j \geq \lambda_j, \text{ for } j = 1, \ldots, \hat{r}.$$

Moreover, if $\hat{\mathbf{A}}_{\hat{r}}$ is used as the estimator of $\mathbf{A}$, the corresponding estimated noise matrix,

$$\hat{\mathbf{E}} = \mathbf{X} - \hat{\mathbf{A}}_{\hat{r}} = \hat{\mathbf{U}}_0 \hat{\mathbf{\Sigma}}_0 \hat{\mathbf{V}}_0^\top,$$

will have no energy in directions in the subspace $\text{row}(\hat{\mathbf{A}}_{\hat{r}})$. This would cause bias in the forthcoming estimation in Section 5.3.

Singular value shrinkage In order to overcome this issue, we consider the *general singular value shrinkage estimator* (Cai et al., 2010; Shabalin and Nobel, 2013; Gavish and Donoho, 2014, 2017),

$$\hat{\mathbf{A}}_\eta = \hat{\mathbf{A}}_\eta^{(d \vee n, \sigma)}(\mathbf{X}) = \sum_{j=1}^{d \wedge n} \eta^{(d \vee n, \sigma)}(\hat{\lambda}_j) \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^\top, \tag{5.4}$$

where $\eta^{(d \vee n, \sigma)}$ is a mapping from $[0, +\infty)$ to $[0, +\infty)$. For convenience of specifying shrinkage functions without depending on σ and $d \vee n$, the general model (5.1) is calibrated to the standard model:

$$\mathbf{X} = \mathbf{A} + \frac{1}{\sqrt{d \vee n}} \mathbf{Z}. \tag{5.5}$$

Then the singular value shrinkage estimator from the general model (5.1), $\hat{\mathbf{A}}_\eta^{(d \vee n, \sigma)}(\mathbf{X})$, is connected with the estimator from the standard model (5.5),

$$\hat{\mathbf{A}}_\eta(\mathbf{X}) = \sum_{j=1}^{d \wedge n} \eta(\hat{\lambda}_j) \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^\top,$$

by the following equation,

$$\hat{\mathbf{A}}_\eta^{(d \vee n, \sigma)}(\mathbf{X}) = \sigma \sqrt{d \vee n} \cdot \hat{\mathbf{A}}_\eta\left(\frac{\mathbf{X}}{\sigma \sqrt{d \vee n}}\right),$$

and their shrinkage functions are linked by

$$\eta^{(d\vee n,\sigma)}(\lambda) = \sigma\sqrt{d\vee n}\cdot\eta(\frac{\lambda}{\sigma\sqrt{d\vee n}}). \tag{5.6}$$

Popular choices of η include hard and soft thresholding with the following shrinkage rules

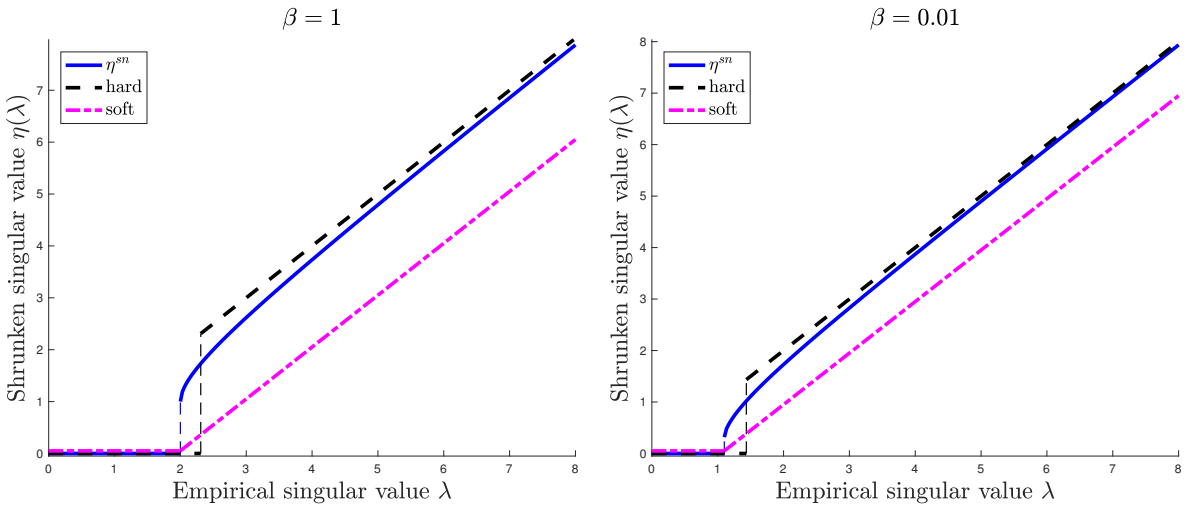
$$\eta_s^{\text{soft}}(\lambda) = \max(0,\lambda - s),$$
$$\eta_\tau^{\text{hard}}(\lambda) = \lambda\cdot\mathbb{1}_{\{\lambda\geq\tau\}}.$$

It is easy to see that hard thresholding leads to the truncated SVD (5.3). Shabalin and Nobel (2013) propose an optimal form of η in the sense of minimizing the operator L_2 norm loss $L(\hat{\mathbf{A}}_\eta,\mathbf{A}) = \|\hat{\mathbf{A}}_\eta - \mathbf{A}\|_2$ under the assumption that $\mathbf{Z}$ is a standard random Gaussian matrix. Gavish and Donoho (2017) extend the results of Shabalin and Nobel (2013) to more general loss functions and non-Gaussian noise matrices, e.g. as in Model (5.1). In the context of our application, we use the optimal shrinkage function with respect to the operator L_2 norm loss function from Shabalin and Nobel (2013), i.e.

$$\eta^\star(\lambda) = \eta^{sn}(\lambda;\beta) = \begin{cases} \frac{1}{\sqrt{2}}\sqrt{\lambda^2-\beta-1+\sqrt{(\lambda^2-\beta-1)^2-4\beta}}, & \text{if } \lambda\geq 1+\sqrt{\beta}, \\ 0, & \text{otherwise,} \end{cases} \tag{5.7}$$

where $\beta = \frac{d\wedge n}{d\vee n}$.

Figure 5.1 shows the graphs of several shrinkage rules for different values of β under the standard model (5.5). The graphs of the optimal shrinkage function η^{sn} are shown as blue solid lines. The graphs of the optimally tuned soft and hard threshold functions discussed in Gavish and Donoho (2014) are shown as the black dashed lines and purple dot-dashed lines respectively. The left and right panels show the cases of square matrix and non-square matrix with the dimension ratio of the toy dataset in Section 3.1.1 respectively. The figure shows that the optimal shrinkage function lies essentially between the optimally tuned soft and hard threshold functions. Note that the threshold for zeroing eigenvalues by η^{sn} is the same as for the optimally tuned soft thresholding, and is largest in the square matrix ($\beta = 1$) case.



(a) Optimal shrinkage function for square matrix.

(b) Optimal shrinkage function for non-square matrix.

Figure 5.1: The graphs of different shrinkage functions under the standard model (5.5). The horizontal axis represents empirical singular value. The vertical axis shows shrunken singular values. The blue solid lines show the graphs of the optimal shrinkage function η^{sn}. The black dashed lines and the purple dot-dashed lines show the graphs of the optimally tuned soft and hard threshold functions. The left and right panel shows the case of square and non-square matrices respectively. This figure is produced by modifying code from the supplement of Gavish and Donoho (2017).

Estimation of noise level In practice, σ in the general model (5.1) is usually not known and needs to be estimated. Assuming $\hat{r} < (d \wedge n)/2$, we use the estimator in Gavish and Donoho (2014), which is simple and robust,

$$\hat{\sigma} = \frac{\lambda_{\text{med}}(\mathbf{X})}{\sqrt{\mu_\beta \cdot (d \vee n)}}, \tag{5.8}$$

where $\lambda_{\text{med}}(\mathbf{X})$ is the median of the first $d \wedge n$ singular values of $\mathbf{X}$, and μ_β is the median of the famous Marčenko–Pastur distribution (Marčenko and Pastur, 1967) with density function

$$p(s) = \frac{\sqrt{((1+\sqrt{\beta})^2 - s)(s - (1-\sqrt{\beta})^2)}}{2\pi\beta s} \mathbb{1}_{[(1-\sqrt{\beta})^2, \quad (1+\sqrt{\beta})^2]}(s).$$

With the formulas in (5.4), (5.6), (5.7) and (5.8), we can estimate the rank of the signal $\mathbf{A}$ as well as the energy of the noise matrix in each principal component direction. More specifically, our estimators of $\mathbf{A}$ and $\mathbf{E}$ are

$$\begin{aligned} \hat{\mathbf{A}} &= \sum_{j=1}^{\hat{r}} \tilde{\lambda}_j \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^\top = \hat{\mathbf{U}} \tilde{\mathbf{\Sigma}} \hat{\mathbf{V}}^\top, \\ \hat{\mathbf{E}} &= \mathbf{X} - \hat{\mathbf{A}}, \end{aligned} \tag{5.9}$$

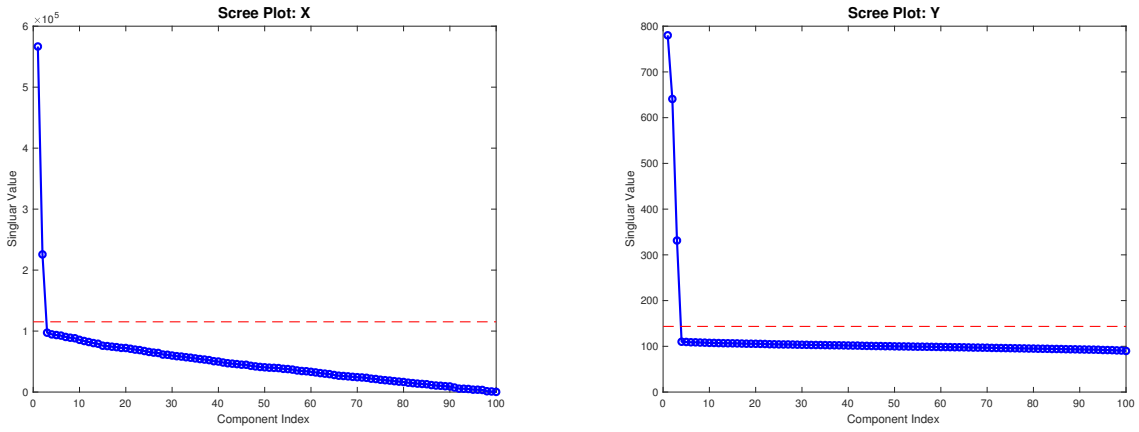


Figure 5.2: Scree plots along with corresponding singular value thresholds for the toy data sets $\mathbf{X}$ (left) and $\mathbf{Y}$ (right). The red dashed lines show the values of singular value threshold. The components with singular values above the red dashed threshold lines are regarded as the signal components.

where

$$\begin{aligned} \hat{r} &= \sum_{j=1}^{d\wedge n} \mathbb{1}_{\{\hat{\lambda}_j \geq \hat{\sigma}\sqrt{d\vee n}(1+\sqrt{\beta})\}} = \sum_{j=1}^{d\wedge n} \mathbb{1}_{\{\hat{\lambda}_j \geq \hat{\sigma}(\sqrt{d}+\sqrt{n})\}}, \\ \tilde{\lambda}_j &= \hat{\sigma}\sqrt{d\vee n}\cdot \eta^{sn}\left(\frac{\hat{\lambda}_j}{\hat{\sigma}\sqrt{d\vee n}}\right), \\ \tilde{\mathbf{\Sigma}} &= \mathrm{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{\hat{r}}). \end{aligned} \tag{5.10}$$

As we can see in (5.10), the threshold for non-zero signal singular values is $\hat{\sigma}(\sqrt{d}+\sqrt{n})$. Thus the estimated noise level $\hat{\sigma}$ is directly related to the estimation of the signal rank $\hat{r}$. If $\hat{\sigma}$ is overestimated, $\hat{r}$ could be underestimated; and vice versa. Applying the estimator $\hat{\sigma}$ to the toy dataset in Section 3.1.1 of Chapter 3, the singular value thresholds of each data block are shown as red dashed lines in Figure 5.2, which lead to correct estimation of the signal ranks of $\mathbf{X}$ and $\mathbf{Y}$, i.e. $\hat{r}_{\mathbf{X}} = 2, \hat{r}_{\mathbf{Y}} = 3$.

5.2.3 Review of useful asymptotic results

To provide insights into the formulas (5.7), a related asymptotic framework and relevant results from random matrix theory are reviewed in this subsection.

Asymptotic framework Without loss of generality, we assume $d \geq n$ for $\mathbf{X} \in \mathbb{R}^{d\times n}$; otherwise we can transpose the matrix. For convenient of asymptotic analysis, the standard model,

$$\mathbf{X} = \mathbf{A} + \frac{1}{\sqrt{d}}\mathbf{Z},$$

is used in this subsection. This enables us to follow the asymptotic framework in Shabalin and Nobel (2013) and Gavish and Donoho (2017). We stress that we are thinking of a sequence model by adding subscripts,

$$\mathbf{X}_n = \mathbf{A}_n + \frac{1}{\sqrt{d_n}}\mathbf{Z}_n \in \mathbb{R}^{d_n\times n}, \tag{5.11}$$

where $\mathbf{A}_n$, $\mathbf{Z}_n$ satisfy the following assumptions:

1. *Invariant white noise.* Each noise matrix $\mathbf{Z}_n$ follows the assumptions in Model (5.1).
2. *Fixed signal singular values* $(\lambda_1, \cdots, \lambda_r)$. The rank of $\mathbf{A}_n$ is fixed to be r. The non-zero singular values of $\mathbf{A}_n$ are also fixed such that $\lambda_1 > \cdots > \lambda_r > 0$.
3. *Asymptotic aspect ratio* β. The sequence d_n satisfies $\frac{n}{d_n} \to \beta \in (0, 1]$, as $n \to +\infty$.

Asymptotic behavior Under the above asymptotic framework, according to the Marčenko–Pastur law (Marčenko and Pastur, 1967), the distribution of the singular values of the noise matrix $\mathbf{Z}_n$ will weakly converge to *the generalized quarter-circle distribution*

$$g(s) = \frac{\sqrt{(s^2 - (1-\sqrt{\beta})^2)((1+\sqrt{\beta})^2 - s^2)}}{\pi\beta s}\mathbb{1}_{[1-\sqrt{\beta},1+\sqrt{\beta}]}(s) \tag{5.12}$$

When the $\mathbf{Z}_n$s are square matrices, i.e. $\beta = 1$, we get the special case, the quarter circle law,

$$g(s) = \frac{1}{\pi}\sqrt{4 - s^2}\mathbb{1}_{[0,2]}(s) \tag{5.13}$$

Moreover, Geman (1980) shows that

$$\lim_{n\to+\infty} \lambda_1(\mathbf{Z}_n) \overset{\text{a.s.}}{=} 1 + \sqrt{\beta}$$

The value $1 + \sqrt{\beta}$ is called the *bulk edge*, which is also the threshold of signal singular values in (5.4).

Summarizing the seminal results from Baik and Silverstein (2006), Nadler et al. (2008), Dozier and Silverstein (2007), Paul (2007), and Benaych-Georges and Nadakuditi (2012), we state the following two theorems, which describe the relationship of singular values and singular vectors between the data matrix $\mathbf{X}_n$ and the signal matrix $\mathbf{A}_n$.

Theorem 5 (Asymptotic location of the empirical singular values)**.** *Let $\hat{\lambda}_j^{(n)}$ be the jth singular value of $\mathbf{X}_n$. For each $1 \leq j \leq r$,*

$$\lim_{n\to+\infty} \hat{\lambda}_j^{(n)} \overset{a.s.}{=} \begin{cases} l(\lambda_j;\beta), & \text{if } \lambda_j \geq \beta^{1/4}, \\ 1+\sqrt{\beta}, & \text{if } \lambda_j < \beta^{1/4}, \end{cases} \tag{5.14}$$

where

$$l(\lambda;\beta) = \sqrt{1+\lambda^2+\beta+\frac{\beta}{\lambda^2}}. \tag{5.15}$$

For each $r < j \leq n$, $\lim_{n\to+\infty} \hat{\lambda}_j^{(n)} \overset{a.s.}{=} 1+\sqrt{\beta}$.

Note that $\eta^{sn}(\lambda;\beta) = l^{-1}(\lambda;\beta)$ for $\lambda \geq 1+\sqrt{\beta}$, where $\eta^{sn}(\lambda;\beta)$ is the optimal shrinkage function in (5.7). Theorem 5 describes the singular value inflation of the data matrix relative to the signal matrix. The second branch of (5.14) together with the last part of Theorem 5 imply that the signal singular values less than $\beta^{1/4}$ are not distinguishable from the noise. The phenomenon is observed in the simulation study of Section 5.4.2.

Theorem 6 (Asymptotic angle between signal and empirical singular vectors)**.** *For $1 \leq i \neq j \leq r$, if $\lambda_i \geq \beta^{1/4}$, we have*

$$\lim_{n\to+\infty} |\langle \mathbf{u}_i^{(n)}, \hat{\mathbf{u}}_j^{(n)} \rangle| \overset{a.s.}{=} \begin{cases} c_1(\lambda_i;\beta), & \text{if } i = j, \\ 0, & \text{if } i \neq j, \end{cases} \tag{5.16}$$

and

$$\lim_{n\to+\infty} |\langle \mathbf{v}_i^{(n)}, \hat{\mathbf{v}}_j^{(n)} \rangle| \overset{a.s.}{=} \begin{cases} c_2(\lambda_i;\beta), & \text{if } i = j, \\ 0, & \text{if } i \neq j, \end{cases} \tag{5.17}$$

where

$$\begin{aligned} c_1(\lambda;\beta) &= \sqrt{\frac{\lambda^4-\beta}{\lambda^4+\lambda^2}}, \\ c_2(\lambda;\beta) &= \sqrt{\frac{\lambda^4-\beta}{\lambda^4+\beta\lambda^2}}. \end{aligned} \tag{5.18}$$

If $\lambda_i < \beta^{1/4}$, we have

$$\lim_{n\to+\infty} |\langle \mathbf{u}_i^{(n)}, \hat{\mathbf{u}}_j^{(n)} \rangle| = \lim_{n\to+\infty} |\langle \mathbf{v}_i^{(n)}, \hat{\mathbf{v}}_j^{(n)} \rangle| \overset{a.s.}{=} 0.$$

Results (5.16) and (5.17) are examples of conical limiting structure as studied in the Figure 2 of Shen et al. (2016). The final part of Theorem 6 implies that the estimated singular vectors associated with signal singular value $\lambda_i < \beta^{1/4}$ contain no useful information. Theorem 5 and Theorem 6 establish the validity of the data singular value threshold $\hat{\sigma}\sqrt{d \vee n}(1 + \sqrt{\beta})$.

5.2.4 Direction specific perturbation angle estimation

For any given direction $\mathbf{a}^\star \in \mathbb{R}^n$, its projections on row($\mathbf{A}$) and row($\hat{\mathbf{A}}$) are

$$\mathbf{a} = \mathbf{V}\mathbf{V}^\top \mathbf{a}^\star = \mathbf{V}\omega_a = \sum_{j=1}^{r} \omega_j \mathbf{v}_j,$$
$$\hat{\mathbf{a}} = \hat{\mathbf{V}}\hat{\mathbf{V}}^\top \mathbf{a}^\star = \hat{\mathbf{V}}\hat{\boldsymbol{\omega}}_a = \sum_{j=1}^{\hat{r}} \hat{\omega}_j \hat{\mathbf{v}}_j,$$

where $\boldsymbol{\omega}_a = \mathbf{V}^\top \mathbf{a}^\star, \hat{\boldsymbol{\omega}}_a = \hat{\mathbf{V}}^\top \mathbf{a}^\star$.

Recall the direction specific perturbation angles in Section 5.2.1, $\theta_1^\star$ and $\theta_2^\star$, which are the perturbation angles of $\mathbf{a}$ and $\hat{\mathbf{a}}$ respectively. Denote the projection of $\mathbf{a}$ on row($\hat{\mathbf{A}}$) and the projection of $\hat{\mathbf{a}}$ on row($\mathbf{A}$) as

$$\hat{\mathbf{a}}_1 = \hat{\mathbf{V}}\hat{\mathbf{V}}^\top \mathbf{a} = \hat{\mathbf{V}}\hat{\mathbf{V}}^\top \mathbf{V}\boldsymbol{\omega}_a,$$
$$\mathbf{a}_1 = \mathbf{V}\mathbf{V}^\top \hat{\mathbf{a}} = \mathbf{V}\mathbf{V}^\top \hat{\mathbf{V}}\hat{\boldsymbol{\omega}}_a.$$

Then we have

$$\begin{aligned}
\theta_1^\star &= \arccos\left(\frac{\mathbf{a}^\top \hat{\mathbf{a}}_1}{\|\mathbf{a}\|\|\hat{\mathbf{a}}_1\|}\right) = \arccos\left(\frac{\boldsymbol{\omega}_a^\top \mathbf{V}^\top \hat{\mathbf{V}}\hat{\mathbf{V}}^\top \mathbf{V}\boldsymbol{\omega}_a}{\|\boldsymbol{\omega}_a\|\|\hat{\mathbf{V}}^\top \mathbf{V}\boldsymbol{\omega}_a\|}\right) = \arccos\left(\frac{\|\hat{\mathbf{V}}^\top \mathbf{V}\boldsymbol{\omega}_a\|}{\|\boldsymbol{\omega}_a\|}\right), \\
\theta_2^\star &= \arccos\left(\frac{\hat{\mathbf{a}}^\top \mathbf{a}_1}{\|\hat{\mathbf{a}}\|\|\mathbf{a}_1\|}\right) = \arccos\left(\frac{\hat{\boldsymbol{\omega}}_a^\top \hat{\mathbf{V}}^\top \mathbf{V}\mathbf{V}^\top \hat{\mathbf{V}}\hat{\boldsymbol{\omega}}_a}{\|\hat{\boldsymbol{\omega}}_a\|\|\mathbf{V}^\top \hat{\mathbf{V}}\hat{\boldsymbol{\omega}}_a\|}\right) = \arccos\left(\frac{\|\mathbf{V}^\top \hat{\mathbf{V}}\hat{\boldsymbol{\omega}}_a\|}{\|\hat{\boldsymbol{\omega}}_a\|}\right).
\end{aligned} \tag{5.19}$$

Notice that the singular values of $\mathbf{V}^\top \hat{\mathbf{V}}$ (and $\hat{\mathbf{V}}^\top \mathbf{V}$) are connected to the principal angles between row($\hat{\mathbf{A}}$) and row($\mathbf{A}$). Denote the singular values of $\mathbf{V}^\top \hat{\mathbf{V}}$ (and $\hat{\mathbf{V}}^\top \mathbf{V}$) as $\sigma_1, \ldots, \sigma_{r\wedge\hat{r}}$. Then $\sigma_i = \cos(\phi_i)$, where ϕ_i is the ith principal angle between row($\hat{\mathbf{A}}$) and row($\mathbf{A}$) defined in Section 3.3.2.2.

This indicates that

$$\phi_1 \leq \theta_1^\star, \theta_2^\star \leq \phi_{r\wedge\hat{r}}.$$

Under the asymptotic framework of Section 5.2.3, applying the results of Theorem 6, $\mathbf{V}^\top\hat{\mathbf{V}}$ is approximated as

$$\mathbf{V}^\top\hat{\mathbf{V}} \approx \mathrm{diag}(c_2(\lambda_1;\beta),\ldots,c_2(\lambda_{r\wedge\hat{r}};\beta)).$$

This implies that, roughly speaking,

$$\sigma_i = \cos(\phi_i) \approx c_2(\lambda_i;\beta) = \sqrt{\frac{\lambda_i^4-\beta}{\lambda_i^4+\beta\lambda_i^2}},$$

and the corresponding principal vectors are approximately the ith principal components in $\mathbf{V}$ and $\hat{\mathbf{V}}$, i.e. $\mathbf{v}_i$ and $\hat{\mathbf{v}}_i$. Following the results in (5.19), the perturbation angle for the ith principal component $\mathbf{v}_i$ of $\mathbf{V}$ is roughly ϕ_i. In other words, the principal component with smaller variance usually has larger perturbation angle. In general, $\theta_1^\star$ and $\theta_2^\star$ are connected to principal angles, $\phi_1,\ldots,\phi_{r\wedge\hat{r}}$, by

$$\cos(\theta_1^\star)^2 \approx \sum_{j=1}^{r\wedge\hat{r}} \omega_{a,j}^{\star 2}\cos(\phi_j)^2$$
$$\cos(\theta_2^\star)^2 \approx \sum_{j=1}^{r\wedge\hat{r}} \hat{\omega}_{a,j}^{\star 2}\cos(\phi_j)^2$$

where $\boldsymbol{\omega}_a^\star = \boldsymbol{\omega}_a/\|\boldsymbol{\omega}_a\|, \hat{\boldsymbol{\omega}}_a^\star = \hat{\boldsymbol{\omega}}_a/\|\hat{\boldsymbol{\omega}}_a\|$. Both the perturbation bounds in Wedin (1972) and Cai et al. (2018) only capture the largest principal angle, $\phi_{r\wedge\hat{r}}$; thus they are usually conservative for the direction perturbation angle in Equation (5.19).

Theorem 7 shows that the distributions of $\mathbf{V}^\top\hat{\mathbf{V}}$ and $\hat{\mathbf{V}}^\top\mathbf{V}$ are orthogonal invariant under our model assumptions. This motivates us to use a bootstrapping procedure to estimate the distribution of $\mathbf{V}^\top\hat{\mathbf{V}}$ and $\hat{\mathbf{V}}^\top\mathbf{V}$.

Theorem 7. *Let $\mathbf{X} = \mathbf{A} + \mathbf{E}$ be a data matrix, which follows the model assumptions in* (5.1)*. Let $\mathbf{V}$ be the right singular matrix of $\mathbf{A}$, and denote $\hat{\mathbf{V}}$ as the rank $\hat{r}$ right singular matrix of $\hat{\mathbf{A}}$, where $\hat{r}$ is the estimator of $r = \mathrm{rank}(\mathbf{A})$. For any given orthogonal matrices $\mathbf{O}_1 \in \mathbb{R}^{d\times d}, \mathbf{O}_2 \in \mathbb{R}^{n\times n}$, let $\mathbf{A}^\star = \mathbf{O}_1\mathbf{A}\mathbf{O}_2$ and $\mathbf{X}^\star = \mathbf{A}^\star + \mathbf{E}$. Analogous to $\mathbf{V}$ and $\hat{\mathbf{V}}$, let $\mathbf{V}^\star$ and $\hat{\mathbf{V}}^\star$ be the rank r right*

singular matrix of $\mathbf{A}^\star$ *and the rank* $\hat{r}$ *right singular matrix of* $\mathbf{X}^\star$ *respectively. Then the distribution of* $\mathbf{V}^\top\hat{\mathbf{V}}$ *and* $\hat{\mathbf{V}}^\top\mathbf{V}$ *are the same as the distributions of* $\mathbf{V}^{\star\top}\hat{\mathbf{V}}^\star$ *and* $\hat{\mathbf{V}}^{\star\top}\mathbf{V}^\star$ *respectively.*

Proof. Since $\mathbf{E}$ is assumed to be orthogonal invariant, $\mathbf{E}^\circ = \mathbf{O}_1\mathbf{E}\mathbf{O}_2$ has the same distribution as $\mathbf{E}$. Therefore $\mathbf{X}^\circ = \mathbf{A}^\star + \mathbf{E}^\circ$ has the same distribution as $\mathbf{X}^\star$. Let $\mathbf{V}^\circ$ and $\hat{\mathbf{V}}^\circ$ be the rank r right singular matrix of $\mathbf{A}^\circ$ and the rank $\hat{r}$ right singular matrix of $\mathbf{X}^\star$ respectively. On the one hand, $\mathbf{V}^{\circ\top}\hat{\mathbf{V}}^\circ$ has the same distribution as $\mathbf{V}^{\star\top}\hat{\mathbf{V}}^\star$.

On the other hand,

$$\begin{aligned}\mathbf{X}^\circ = \mathbf{A}^\star + \mathbf{E}^\circ &= \mathbf{O}_1\mathbf{A}\mathbf{O}_2 + \mathbf{O}_1\mathbf{E}\mathbf{O}_2 \\ &= \mathbf{O}_1\mathbf{X}\mathbf{O}_2.\end{aligned}$$

The SVD of $\mathbf{X}^\circ$ is related to that of $\mathbf{X}$ by $\mathbf{V}^\circ = \mathbf{O}_2\mathbf{V}, \hat{\mathbf{V}}^\circ = \mathbf{O}_2\hat{\mathbf{V}}$. Thus $\mathbf{V}^{\circ\top}\hat{\mathbf{V}}^\circ = \mathbf{V}^\top\mathbf{O}_2^\top\mathbf{O}_2\hat{\mathbf{V}} = \mathbf{V}^\top\hat{\mathbf{V}}$. Therefore, the distribution of $\mathbf{V}^\top\hat{\mathbf{V}}$ is the same as the distribution of $\mathbf{V}^{\star\top}\hat{\mathbf{V}}^\star$.

The proof of $\hat{\mathbf{V}}^\top\mathbf{V}$ follows by transposition. □

This establishes the validity of our computation procedure in Section 5.3. Note that Theorem 7 is still valid if we assume the relationship between $\mathbf{A}$ and $\mathbf{E}$ is essentially isotropic instead of assuming the distribution of $\mathbf{Z}$ is orthogonal invariant, because it directly implies $\mathbf{X}^\star = \mathbf{O}_1\mathbf{A}\mathbf{O}_2 + \mathbf{E}$ has the same distribution as $\mathbf{X}$. Moreover, Benaych-Georges and Nadakuditi (2012) also show that all the asymptotic results in Section 5.2.3 are valid when either $\mathbf{A}$ is non-random and $\mathbf{E}$ is orthogonal invariant, or the relationship between $\mathbf{A}$ and $\mathbf{E}$ is essentially isotropic.

With the estimators of $\mathbf{V}^\top\hat{\mathbf{V}}$ and $\hat{\mathbf{V}}^\top\mathbf{V}$, the estimation of $\theta_2^\star$ is straightforward, since $\hat{\boldsymbol{\omega}}_a$ is observable and thus can be directly plugged into (5.19). It is more challenging to estimate $\theta_1^\star$ since $\boldsymbol{\omega}_a$ in (5.19) is not observable. In order to still use (5.2) to bound the range of θ, we can bound $\theta_1^\star$ by $\phi_{\hat{r}}$ and thus

$$(\hat{\theta} - \phi_{\hat{r}})_+ \leq \theta \leq \hat{\theta} + \theta_2^\star. \tag{5.20}$$

One might think we can estimate $\boldsymbol{\omega}_a$ from $\hat{\boldsymbol{\omega}}_a$. However, even in the asymptotic limit situation, the uncertainty of $\boldsymbol{\omega}_a$ is still not eliminated. For example, we consider

$$\mathbf{V} = \begin{bmatrix} \frac{1}{\sqrt{2}}\sin(\phi_1) & \frac{1}{\sqrt{2}}\cos(\phi_1) & 0 & \frac{1}{\sqrt{2}}\sin(\phi_1) & \frac{1}{\sqrt{2}}\cos(\phi_1) & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}^\top,$$
$$\hat{\mathbf{V}} = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}}\sin(\phi_2) & 0 & \frac{1}{\sqrt{2}}\cos(\phi_2) & -\frac{1}{\sqrt{2}}\sin(\phi_2) & 0 & \frac{1}{\sqrt{2}}\cos(\phi_2) \end{bmatrix}^\top,$$

where $\phi_i = \arccos(c_2(\lambda_i); \beta), i = 1, 2$, and $\lambda_1 > \lambda_2$. Then we have

$$\mathbf{V}^\top \hat{\mathbf{V}} = \begin{bmatrix} \cos(\phi_1) & 0 \\ 0 & \cos(\phi_2) \end{bmatrix}$$

Note that $\mathbf{V}$ and $\hat{\mathbf{V}}$ satisfy all the asymptotic structures in Theorem 5 and Theorem 6. However, consider $\mathbf{a}^\star = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^\top \in \mathbb{R}^6$. This gives

$$\boldsymbol{\omega}_a = \mathbf{V}^\top \mathbf{a}^\star = \begin{bmatrix} \frac{1}{\sqrt{2}}\sin(\phi_1) & 0 \end{bmatrix}^\top,$$
$$\hat{\boldsymbol{\omega}}_a = \hat{\mathbf{V}}^\top \mathbf{a}^\star = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}}\sin(\phi_2) \end{bmatrix}^\top.$$

Thus even though the projection of $\mathbf{a}^\star$ on row($\mathbf{A}$) is merely in the direction of $\mathbf{v}_1$, its projection on row($\hat{\mathbf{A}}$) is in the direction of $\hat{\mathbf{v}}_2$. Therefore, the range in (5.20) is the best which we can infer from only observing $\mathbf{a}^\star$ and $\mathbf{X}$.

Random direction angle Analogous to the random direction bound in Section 3.3.3.1 of Chapter 3, we estimate the distribution of the angles between a random direction and a rank $\hat{r}$ subspace in $\mathbb{R}^n$. This distribution only depends on the estimated signal rank $\hat{r}$, and the dimension of the row space, n. We obtain this distribution by simulation. In particular, a direction is randomly selected from the unit sphere in $\mathbb{R}^n$, which can be generated by normalizing an n-dimensional standard Gaussian vector. Then we calculate the angle between the generated direction and $\hat{\mathbf{V}}$. We recommend the 5th percentile of the angle distribution as a point estimator in practice, which is denoted as θ_0 and referred as random angle bound.

The angle range for $\theta = \measuredangle(\mathbf{a}^\star, \text{row}(\mathbf{A}))$ in (5.20), combining with the random angle bound is used to infer whether $\mathbf{a}^\star$ lies in the signal row space row($\mathbf{A}$). For a given direction $\mathbf{a}^\star \in \mathbb{R}^n$, if the lower bound on θ, i.e. $(\hat{\theta} - \phi_{\hat{r}})_+$, is 0, we can not reject the hypothesis that $\mathbf{a}^\star \in \text{row}(\mathbf{A})$. If the upper bound on θ, i.e. $\hat{\theta} + \theta_2^\star$, is larger than θ_0, we can not reject the hypothesis that $\mathbf{a}^\star$ is not associated with row($\mathbf{A}$).

However, when the angle range in (5.20) is too large such that the lower bound is 0 and the upper bound is larger than θ_0, we cannot gain any useful information from it. A typical case is that the projection $\hat{\mathbf{a}}$ lies on the principal components of $\hat{\mathbf{V}}$ with small variance such that $\theta_2^\star > \frac{\theta_0}{2}$. In this case, by definition, $\phi_{\hat{r}} \geq \theta_2^\star > \frac{\theta_0}{2}$. Then for the intermediate values of $\hat{\theta}$ such that $(\theta_0 - \theta_2^\star)_+ \leq \hat{\theta} \leq \phi_{\hat{r}}$, the angle range covers both 0 and θ_0. In other words, although these principal components are related to signals, they are not very useful in this case. In our application, the judgment for these intermediate values of $\hat{\theta}$ is the most crucial.

In order to make the angle range of θ in (5.20) still informative under this situation, we adjust $\hat{\mathbf{V}}$ to $\bar{\mathbf{V}}$ by filtering out the principal components with corresponding $\theta_2^\star$ larger than $\frac{\theta_0}{2}$. Then $\hat{\mathbf{U}}$ and $\tilde{\mathbf{\Sigma}}$ are adjusted to $\bar{\mathbf{U}}$ and $\bar{\mathbf{\Sigma}}$ accordingly. Thus our adjusted signal matrix estimator is

$$\bar{\mathbf{A}} = \bar{\mathbf{U}}\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^\top. \tag{5.21}$$

Let $\bar{r}$ be the rank of $\bar{\mathbf{A}}$. Note that for the adjusted signal matrix estimator, the largest principal angle between row($\bar{\mathbf{A}}$) and row($\mathbf{A}$), $\phi_{\bar{r}}$, is less than $\frac{\theta_0}{2}$. Then the max length of the angle range in (5.20) is surely less than θ_0, and thus it can not cover 0 and θ_0 at the same time. The estimated noise matrix $\hat{\mathbf{E}} = X - \hat{\mathbf{A}}$ in (5.9) stays unchanged. Next $\hat{\mathbf{E}}$ and $\bar{\mathbf{A}}$ are used to estimated the bound for θ of $\mathbf{a}^\star$ as in Section 5.3.

5.3 Algorithm

This section describes the algorithm for estimating the range of the angle, $\theta = \measuredangle(\mathbf{a}^\star, \text{row}(\mathbf{A}))$, for an arbitrary given direction $\mathbf{a}^\star \in \mathbb{R}^n$ based on (5.20). Since $\hat{\theta}$ can be calculated directly, the essential part of the algorithm is to estimate $\phi_{\hat{r}}$ and $\theta_2^\star$. Because each $\mathbf{a}^\star$ has its specific score $\hat{\boldsymbol{\omega}}_a$ on the row($\hat{\mathbf{A}}$), its corresponding $\theta_2^\star$ is direction driven. As shown in the formula in (5.19), they have the matrix $\mathbf{V}^\top\hat{\mathbf{V}}$ in common, which has an orthogonal invariant distribution as stated

in Theorem 7. In order to avoid repeated computation, the first part of the algorithm generates a batch of bootstrap samples of $\mathbf{V}^\top\hat{\mathbf{V}}$, which are cached. The second part of the algorithm utilizes the cached samples of $\mathbf{V}^\top\hat{\mathbf{V}}$ to bound the angle range for different directions.

Algorithm part 1

Step 1: Signal Matrix Recovery. Apply the opitmal singular value shrinkage estimator (5.9) on the data matrix $\mathbf{X}$ to get estimators $\hat{\mathbf{A}} = \hat{\mathbf{U}}\tilde{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top$ and $\hat{\mathbf{E}} = \mathbf{X} - \hat{\mathbf{A}}$. Adjust $\hat{\mathbf{A}}$ to $\bar{\mathbf{A}}$ as in (5.21).

Step 2: Bootstrap Estimation. Generate a random left singular matrix $\mathbf{U}^\star$ and a random right singular matrix $\mathbf{V}^\star$. Obtain the new signal matrix $\mathbf{A}^\star = \mathbf{U}^\star\tilde{\mathbf{\Sigma}}\mathbf{V}^{\star\top}$ and the new data matrix $\mathbf{X}^\star = \mathbf{A}^\star + \hat{\mathbf{E}}$. Let $\hat{\mathbf{V}}^\star$ be the rank $\hat{r}$ ($\bar{r}$ if using $\bar{\mathbf{A}}$) right singular matrix of $\mathbf{X}^\star$. Then $\mathbf{V}^{\star\top}\hat{\mathbf{V}}^\star$ is one realization from the bootstrap distribution modeling the distribution of $\mathbf{V}^\top\hat{\mathbf{V}}$.

Step 3: Cache Results. Repeat Step 2 M times and cache each bootstrap sample $\mathbf{V}^{\star\top}\hat{\mathbf{V}}^\star$.

Algorithm part 2

For an arbitrary direction $\mathbf{a}^\star \in \mathbb{R}^n$, the upper bound and lower bound on θ can be quickly calculated now.

Estimation of the upper bound on θ. Compute its score on $\text{row}(\hat{\mathbf{A}})$, $\hat{\boldsymbol{\omega}}_a = \hat{\mathbf{V}}^\top\mathbf{a}^\star$. For each bootstrap sample of $\mathbf{V}^\top\hat{\mathbf{V}}$ from Algorithm 1, $\mathbf{V}^{\star\top}\hat{\mathbf{V}}^\star$, calculate the corresponding $\theta_2^\star$ based on the formula (5.19). This is a bootstrap replication of $\theta_2^\star$. Using the 95th percentile of the resampling bootstrap distribution, $\tilde{\theta}_2^\star$, as a point estimator, gives the upper bound $\hat{\theta} + \tilde{\theta}_2^\star$.

Estimation of the lower bound on θ. For each bootstrap sample of $\mathbf{V}^\top\hat{\mathbf{V}}$, a bootstrap replication for $\phi_{\hat{r}}$ can be obtained by calculating its smallest singular value. Using the 95th percentile of the resampled bootstrap distribution, $\tilde{\phi}_{\hat{r}}$, we can get the lower bound $(\hat{\theta} - \tilde{\phi}_{\hat{r}})_+$.

5.4 Simulation study

In this section we carry out numerical experiments to demonstrate the effectiveness of the computation algorithm in Section 5.3. The toy dataset in Section 3.1.1 of Chapter 3, which contains both a square matrix and a non-square matrix, will be studied in the first part and the third part. The first part of the simulation study in Section 5.4.1 illustrates the estimation of the direction specific perturbation angle, $\theta_2^\star$, under rank correct specification as well as rank misspecification cases. The second part of the simulation study in Section 5.4.2 analyzes new toy examples similar

to the toy example in Shabalin and Nobel (2013) with both Gaussian and non-Gaussian noise. The third part of the simulation study in Section 5.4.3 shows the application of this perturbation framework to the AJIVE directions of the toy dataset.

5.4.1 Direction specific perturbation angle estimation

As discussed in Section 5.2.2, the estimated signal rank is affected by the estimation of the noise level σ. In this section, we will analyze the behavior of the algorithm for estimating $\theta_2^\star$ under both correct and incorrect rank specification, which is driven by the estimation of σ. In this subsection, only the theoretical values of the perturbation bounds from Wedin (1972) and Cai et al. (2018) are compared to the direction specific perturbation angles. In practice, the values of these perturbation bounds depend on the underlying true signal matrix and neither can not be directly applied. Feng et al. (2017) propose a practical and effective algorithm to estimate the Wedin bound.

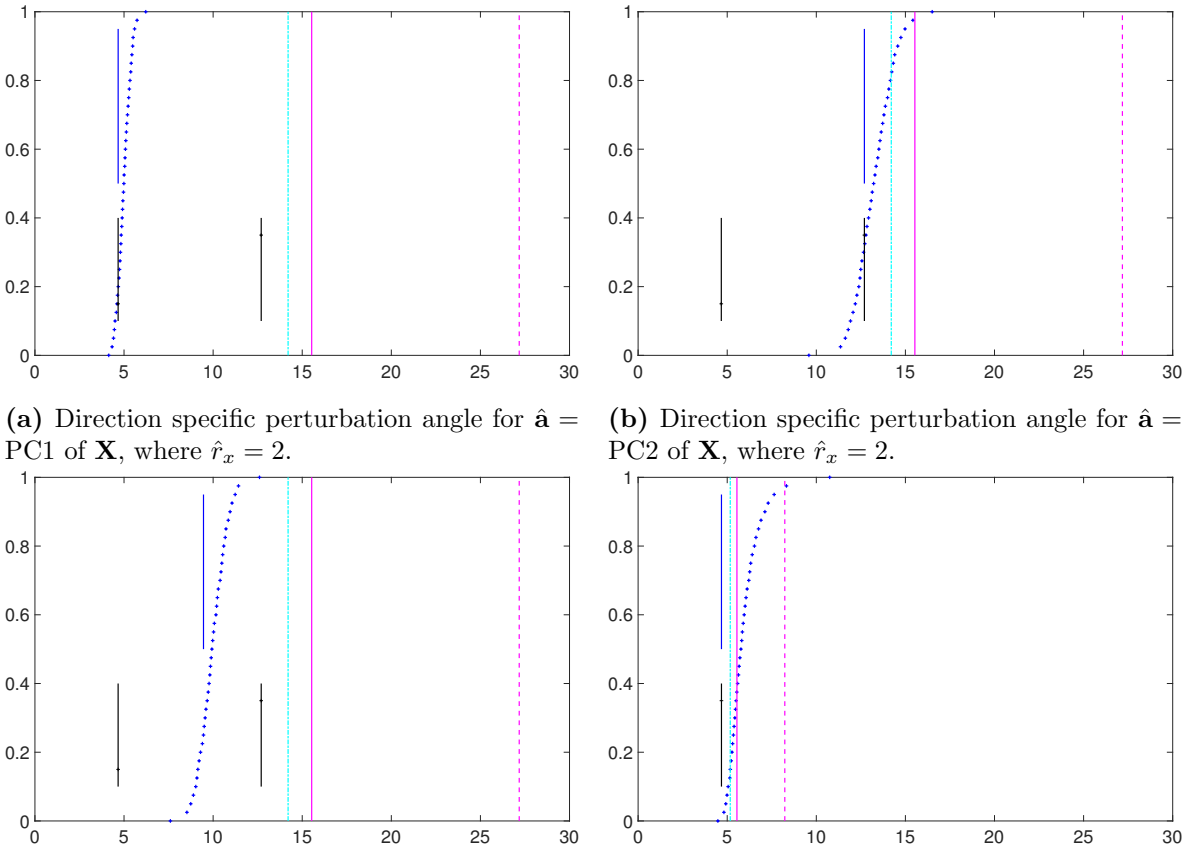
Results under correct rank specification

Applying the estimator $\hat{\sigma}$ in (5.8), we get $\hat{\sigma}_x = 4.9877 \times 10^3, \hat{\sigma}_y = 1.0005$, while the true values are $\sigma_x = 5 \times 10^3, \sigma_y = 1$. So it is not surprising that the signal rank of each data block are correctly specified.

Figure 5.3 illustrates how well the bootstrap distributions estimate the direction specific perturbation angles. The blue solid lines on the upper part of each panel show the values of the true direction specific perturbation angle for different vectors in row($\hat{\mathbf{A}}_1$), the signal row space of $\mathbf{X}$. The blue plus signs show the c.d.f. values of the bootstrapping distribution of each $\theta_2^\star$. Figure 5.3 shows that the true values of $\theta_2^\star$ are always in the range of the resampled bootstrap distribution. The vertical black lines on the lower part of each panel show the values of principal angles between the estimated signal space and the true signal spaces. Panel 5.3(a) and 5.3(b) show that the perturbation angles for PC1 and PC2 of $\mathbf{X}$ are overlapped with the principal angles ϕ_1, ϕ_2 respectively. Panel 5.3(c) shows the perturbation angles for the vector $\hat{\mathbf{a}} = (\text{PC1} + \text{PC2})/\sqrt{2}$ lies between ϕ_1 and ϕ_2. All these are consistent with the conclusions in Section 5.2.4.

The vertical cyan dot-dashed line shows the value of the Wedin bound of the perturbation angle defined in Section 3.3.2.2. The purple dashed lines and solid lines show the perturbation bounds from Theorem 1 and Proposition 1 in Cai et al. (2018) respectively. Since all these perturbation

angle bounds are subspace oriented, they are the same in each cases. As seen in Figure 5.3, all these perturbation bounds are conservative relative to the resampled direction specific perturbation angles. Figure 5.3 also implies the Wedin bound is less conservative than the perturbation bounds from Cai et al. (2018) in the square matrix case.



(a) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC1 of $\mathbf{X}$, where $\hat{r}_x = 2$.

(b) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC2 of $\mathbf{X}$, where $\hat{r}_x = 2$.

(c) Direction specific perturbation angle for $\hat{\mathbf{a}} = (\text{PC1} + \text{PC2})/\sqrt{2}$, where $\hat{r}_x = 2$.

(d) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC1 of $\mathbf{X}$, where $\hat{r}_x = 1$.

Figure 5.3: Shows the direction specific perturbation angles for different vectors in the estimated signal row space of $\mathbf{X}$, a square matrix introduced in Section 3.1.1 of Chapter 3. The x-axis represents angles and the y-axis represents the values of empirical distribution of resampled direction specific perturbation angles. The vertical blue lines segment on the upper side of each panel show the value of true direction specific perturbation angle. The blue plus signs show the empirical distribution of the bootstrap samples of perturbation angles. The black vertical lines on the lower part of each panel show the principal angles between the signal space and the estimated signal space. The vertical cyan dot-dashed lines show the values of the Wedin bound of the perturbation angle between the signal space and the estimated signal space. The purple dashed line and solid line show the perturbation bounds from Theorem 1 and Proposition 1 in Cai et al. (2018) respectively. Panel 5.3(d) shows the case of rank underestimation, $\hat{r}_x = 1$. In all other panelsthe signal rank of $\mathbf{X}$ is correctly specified, i.e. $\hat{r}_x = 2$. The half value of random angle bound is not shown on this figure since it is larger than all of these angles.

Figure 5.4 is an analogous plot to Figure 5.3 with respect to the non-square matrix $\mathbf{Y}$. In this case, the sampled direction specific perturbation angles still match very well with the true

direction specific perturbation angles. However, the Wedin bound becomes very conservative in the non-square matrix case. The perturbation bound from Theorem 1 of Cai et al. (2018) is better than the Wedin bound but still very conservative. The perturbation bound from Proposition 1 of Cai et al. (2018) is sharp in the sense of estimating the largest principal angle, but it is usually still conservative for direction specific perturbation angles.

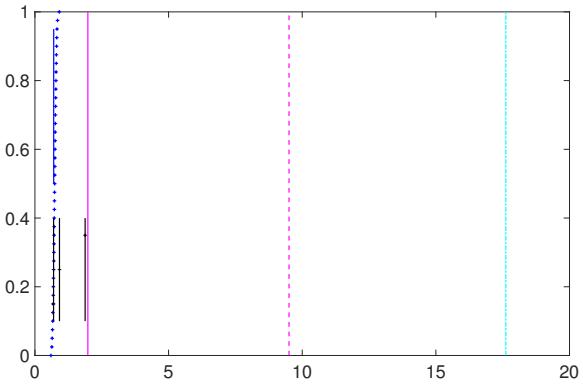
Results under rank misspecification

The estimator $\hat{\sigma}$ in (5.8) works pretty well on the toy datasets. In practice, it is common to overestimate or underestimate σ. As mentioned in Section 5.2.2, overestimating σ could cause underestimating the signal rank; and vice versa. In this paragraph, we will analyze the behavior of the algorithm under rank misspecification.

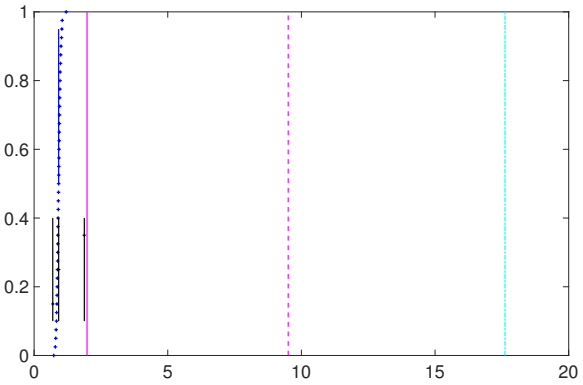
First, we analyze the case of rank underestimation, i.e. overestimation of $\hat{\sigma}$. The values of $\hat{\sigma}_x$ and $\hat{\sigma}_y$ have been manually set to be 11500 and 3.02 respectively, which results in $\hat{r}_x = 1 < r_x = 2, \hat{r}_y = 2 < r_y = 3$. Panel 5.3(d) of Figure 5.3 shows the direction specific perturbation angle for PC1 of $\mathbf{X}$ with under estimated signal rank $\hat{r}_x = 1$ does not change, and is still in the range of the resampled perturbation angles. However, the variance of the distribution of the resampled perturbation angles becomes larger, and thus the 95th percentile estimator becomes more conservative. In this case, the perturbation bounds from Wedin (1972) and Cai et al. (2018) are still effective, while the Wedin bound is more accurate. Panel 5.4(e) and 5.4(f) of Figure 5.4 illustrate the perturbation angles for $\mathbf{Y}$ with underestimated rank $\hat{r}_y = 2$. We have similar conclusions in the case of $\mathbf{Y}$, but the perturbation bounds from Cai et al. (2018) are more accurate than the Wedin bound.

Next the behavior of the algorithm in Section 5.3 under rank overestimation, i.e. $\hat{\sigma}$ under estimation, is analyzed. The values of $\hat{\sigma}_x$ and $\hat{\sigma_y}$ are manually set to be 4800 and 0.996, which lead to $\hat{r}_x = 3$ and $\hat{r}_y = 4$. As seen in Figure 5.5 and Figure 5.6, the perturbation bounds from Wedin (1972) become conservative. Also, the perturbation bounds from Cai et al. (2018) are 90°, and thus they are useless in this case.

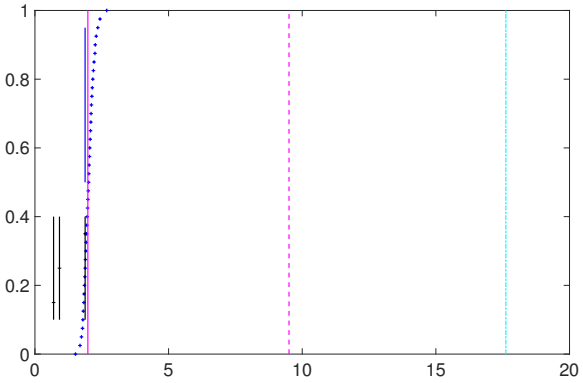
Figure 5.5 shows the estimation of the direction specific perturbation angles for directions in the rank 3 right singular matrices of $\mathbf{X}$. Panel 5.5(a), 5.5(b) and 5.5(c) of Figure 5.5 show the results for the directions of PC1, PC2, and (PC1 + PC2)/$\sqrt{2}$, where the true perturbation angles and the



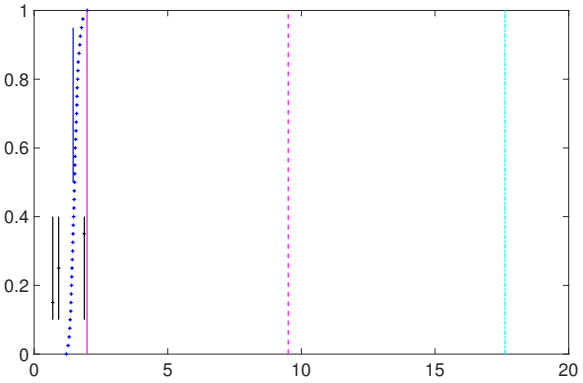
(a) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC1 of $\mathbf{Y}$, where $\hat{r}_y = 3$.



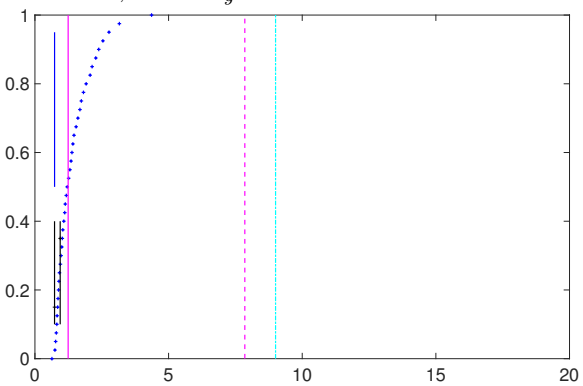
(b) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC2 of $\mathbf{Y}$, where $\hat{r}_y = 3$.



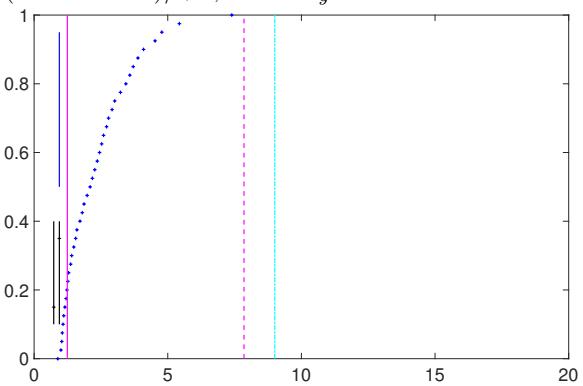
(c) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC3 of $\mathbf{Y}$, where $\hat{r}_y = 3$.



(d) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ (PC1 + PC3)$/\sqrt{2}$, where $\hat{r}_y = 3$.



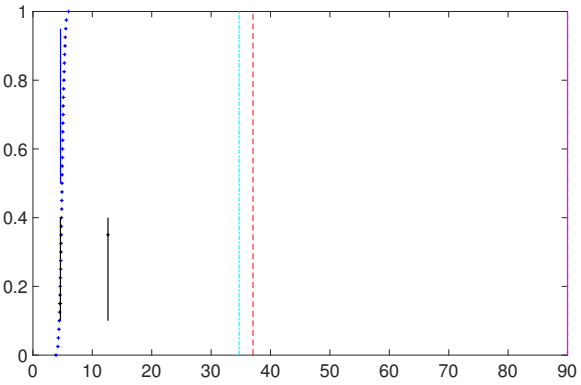
(e) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC1 of $\mathbf{Y}$, where $\hat{r}_y = 2$.



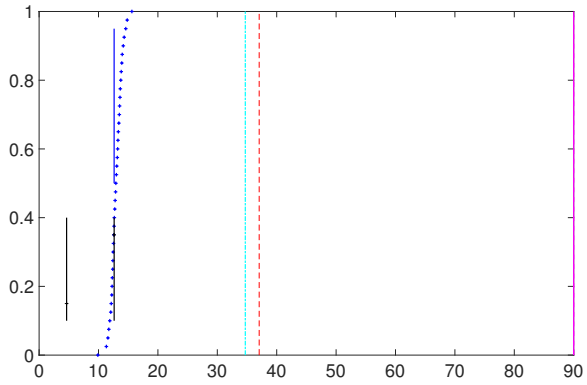
(f) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC2 of $\mathbf{Y}$, where $\hat{r}_y = 2$.

Figure 5.4: Shows the direction specific perturbation angles for different vectors in the estimated signal row space of $\mathbf{Y}$, a non-square matrix introduced in Section 3.1.1 of Chapter 3.In Panel 5.4(a), 5.4(b), 5.4(c) and 5.4(d), the signal rank of $\mathbf{Y}$ is correctly specified, i.e., $\hat{r}_y = 3$. In Panel 5.4(e) and 5.4(f), the signal rank is underestimated, $\hat{r}_y = 2$.

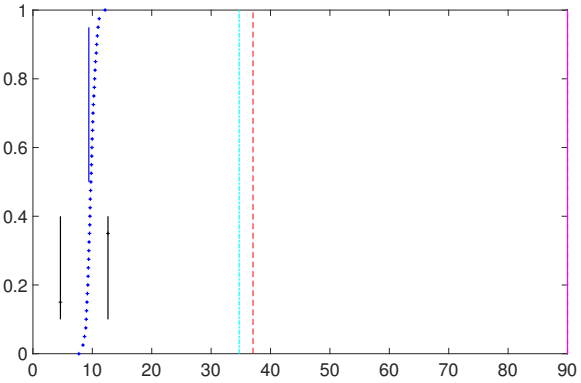
resampled perturbation angles are almost the same as in the corresponding panels of Figure 5.3 under correct rank specification of $\mathbf{X}$. In this case, the true direction specific perturbation angle for PC3 of $\mathbf{X}$ is not defined since the true signal rank of $\mathbf{X}$ is 2. The vertical red dashed lines show the half values of the random angle bounds. As seen in Panel 5.5(d) of Figure 5.5, the resampled perturbation angles are in the range of the random angle bound. This makes sense since they correspond to noise. In particular, they are larger than the half of the value of the random angle bound defined in Section 5.2.4. This also indicates the $\hat{r}$ has been over-estimated. This is the motivation for the reduction of $\hat{\mathbf{V}}$ to $\hat{\mathbf{U}}$ in Section 5.2.4.



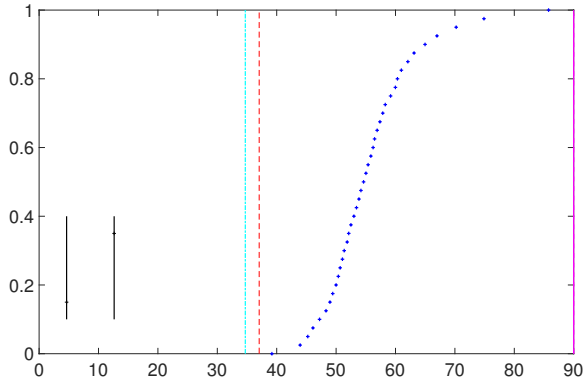
(a) Direction specific perturbation angle for $\hat{\mathbf{a}}$ = PC1 of $\mathbf{X}$, where $\hat{r}_x = 3$.



(b) Direction specific perturbation angle for $\hat{\mathbf{a}}$ = PC2 of $\mathbf{X}$, where $\hat{r}_x = 3$.



(c) Direction specific perturbation angle for $\hat{\mathbf{a}}$ = (PC1 + PC2)/$\sqrt{2}$, where $\hat{r}_x = 3$.

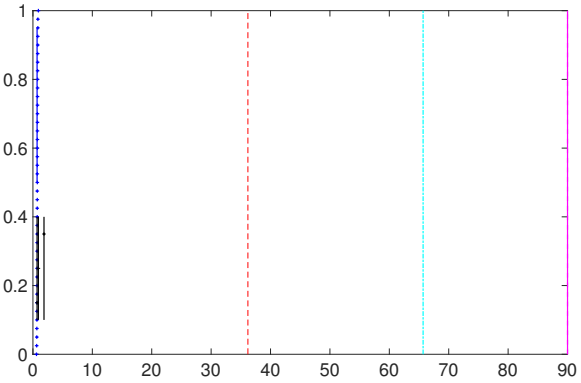


(d) Direction specific perturbation angle for $\hat{\mathbf{a}}$ = PC3 of $\mathbf{X}$, where $\hat{r}_x = 3$. In this case, the true perturbation angle is not defined.

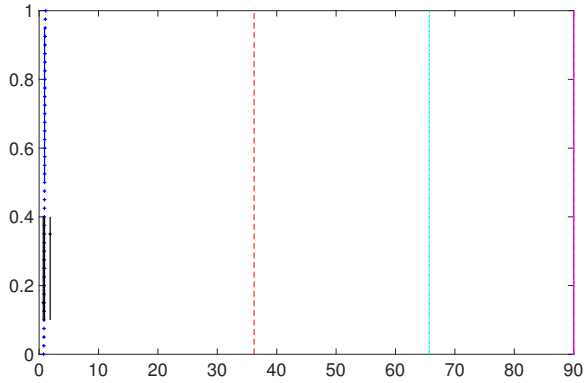
Figure 5.5: Shows the direction specific perturbation angles for different vectors in the estimated signal row space of $\mathbf{X}$ when the signal rank, $\hat{r}_x = 3$, is over estimated. The vertical red dashed lines show the half of the values of random angle bound. The perturbation bounds from Cai et al. (2018) are 90°, which are not useful in this case.

Figure 5.6, which is analogous to Figure 5.5, shows the results for $\mathbf{Y}$ in the case of rank overestimation. Panel 5.6(a), 5.6(b), and 5.6(c) show that the results of directions in the rank 3

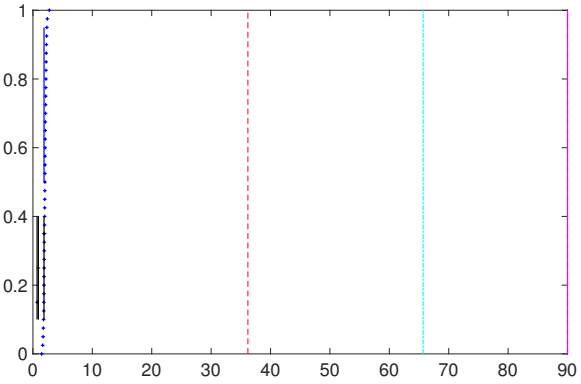
right singular matrices of $\mathbf{Y}$ are similar to the corresponding results in Figure 5.4. Panel 5.6(d) shows that the resampled perturbation angles for PC4 are larger than the half value of the random angle bound.



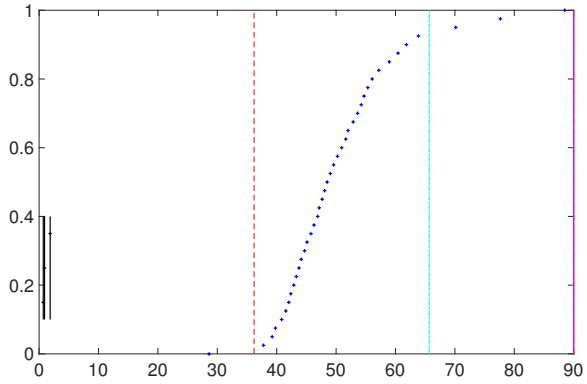
(a) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC1 of $\mathbf{Y}$, where $\hat{r}_y = 4$.



(b) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC2 of $\mathbf{Y}$, where $\hat{r}_y = 4$.



(c) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC3 of $\mathbf{Y}$, where $\hat{r}_y = 4$.



(d) Direction specific perturbation angle for $\hat{\mathbf{a}} =$ PC4 of $\mathbf{Y}$, where $\hat{r}_y = 4$. In this case, the true perturbation angle is not defined.

Figure 5.6: Shows the direction specific perturbation angles for different vectors in the estimated signal row space of $\mathbf{Y}$ when the signal rank, $\hat{r}_y = 4$, is over estimated. The bounds in Cai et al. (2018) is not useful in this case.

5.4.2 Gaussian vs non-Gaussian

In this section, we study new toy examples inspired by the toy example in Shabalin and Nobel (2013) under both Gaussian and non-Gaussian noise with the settings of the standard model (5.5). For the first data matrix $\mathbf{X}_1$,

$$\mathbf{X}_1 = \mathbf{A}_1 + \frac{1}{\sqrt{500}} Z_1 \in \mathbb{R}^{500 \times 500}$$

where $\mathbf{A}_1 \in \mathbb{R}^{500\times 500}$ is a rank 50 signal matrix with non-zero singular values, $5.0, 4.9, 4.8, \ldots, 0.2, 0.1$, and $\mathbf{Z}_1$ is a standard Gaussian random matrix.

For the second data matrix $\mathbf{Y}_1$,

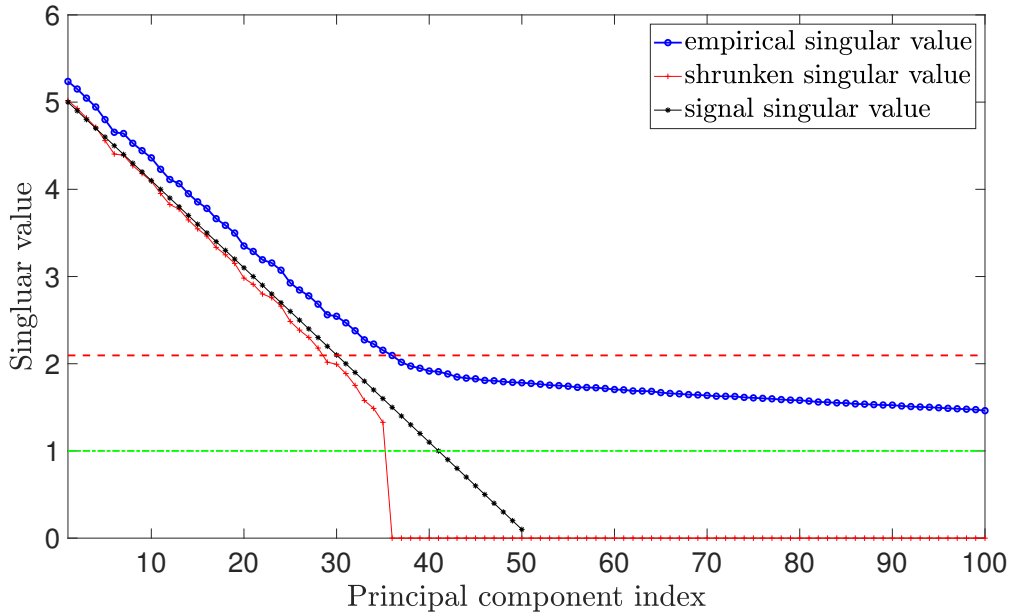
$$\mathbf{Y}_1 = \mathbf{A}_2 + \frac{1}{\sqrt{5000}} Z_2 \in \mathbb{R}^{5000\times 500}$$

where $\mathbf{A}_2 \in \mathbb{R}^{5000\times 500}$ is another rank 50 signal matrix with non-zero singular values, $5.0, 4.9, 4.8, \ldots, 0.2, 0.1$, and $\mathbf{Z}_2$ is also a standard Gaussian random matrix.

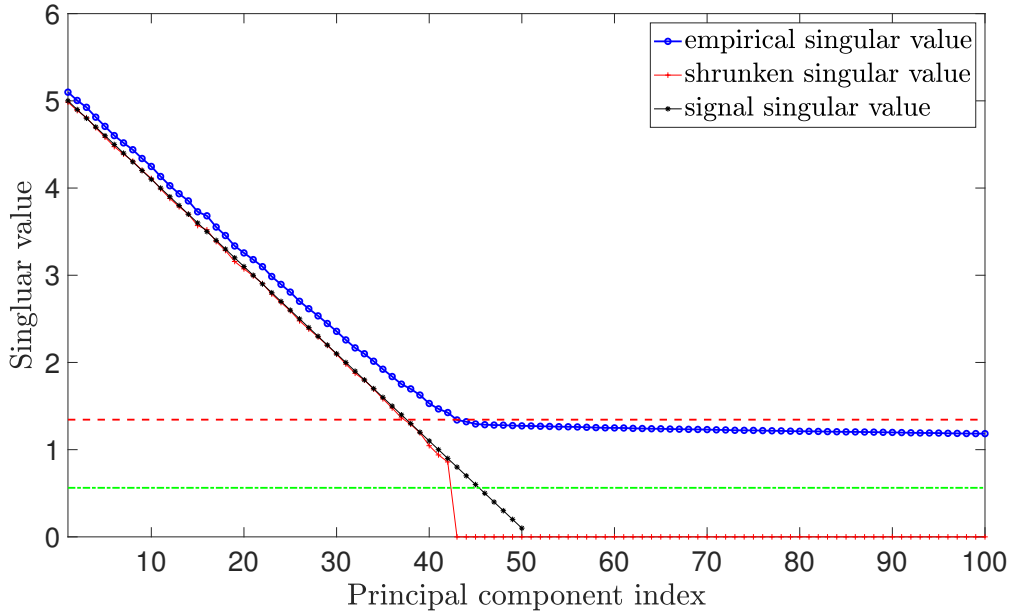
The third and fourth data matrices $\mathbf{X}_2 = \mathbf{A}_1 + \frac{1}{\sqrt{500}} Z_3 \in \mathbb{R}^{500\times 500}$ and $\mathbf{Y}_2 = \mathbf{A}_2 + \frac{1}{\sqrt{5000}} Z_4 \in \mathbb{R}^{5000\times 500}$ have the same signal matrices as $\mathbf{X}_1$ and $\mathbf{Y}_1$ respectively, where the entries of $\mathbf{Z}_3$ and $\mathbf{Z}_4$ are i.i.d. scaled Student t distribution, $\sqrt{3/5} \quad t_5$, with degree of freedom 5 and variance 1. Notice that the distributions of $\mathbf{Z}_3$ and $\mathbf{Z}_4$ are not orthogonally invariant.

Figure 5.7 shows the scree plots of $\mathbf{X}_1$ and $\mathbf{Y}_1$ for the first 200 Principal components on the upper and lower panels respectively. The blue circles show the empirical singular values from data matrices. The black stars show the corresponding singular values of signal matrices. The discrepancies between the blue circles and the black stars clearly illustrate the inflation of the empirical singular values relative to the underlying signal singular values. The red pluses show the shrunken singular values. The horizontal red dashed lines show the empirical singular value thresholds, $\hat{\sigma}(\sqrt{d} + \sqrt{n})$, in (5.10), which lead to the rank estimations $\hat{r}_{x1} = 35, \hat{r}_{y1} = 42$. The horizontal green dot-dashed lines show the cutoff, $\beta^{1/4}$, for the signal singular values to be distinguishable from the noise. All signal matrix principal components with singular value below the green dashed line are not recoverable. In theory, only first 40 and 45 principal components in $\mathbf{A}_1$ and $\mathbf{A}_2$ are distinguishable from the noise. In practice, the signal matrix principal components near the green dashed line are also shrunken to 0 values since their empirical singular values are below the red dashed lines. For both square and non-square Gaussian matrix $\mathbf{X}_1$ and $\mathbf{Y}_1$, the shrunken singular values of the empirical principal components with singular value above the red dashed line match very well with the true signal singular values.

Figure 5.8 shows the analogous scree plots for $\mathbf{X}_2$ and $\mathbf{Y}_2$, which are data matrices with Student t noise. It indicates that the theoretical recoverable rank for $\mathbf{X}_2$ and $\mathbf{Y}_2$ are also 40 and 45, and the estimated rank from the optimal shrinkage estimators 5.9 are $\hat{r}_{x2} = 36$, and $\hat{r}_{y2} = 43$ respectively.



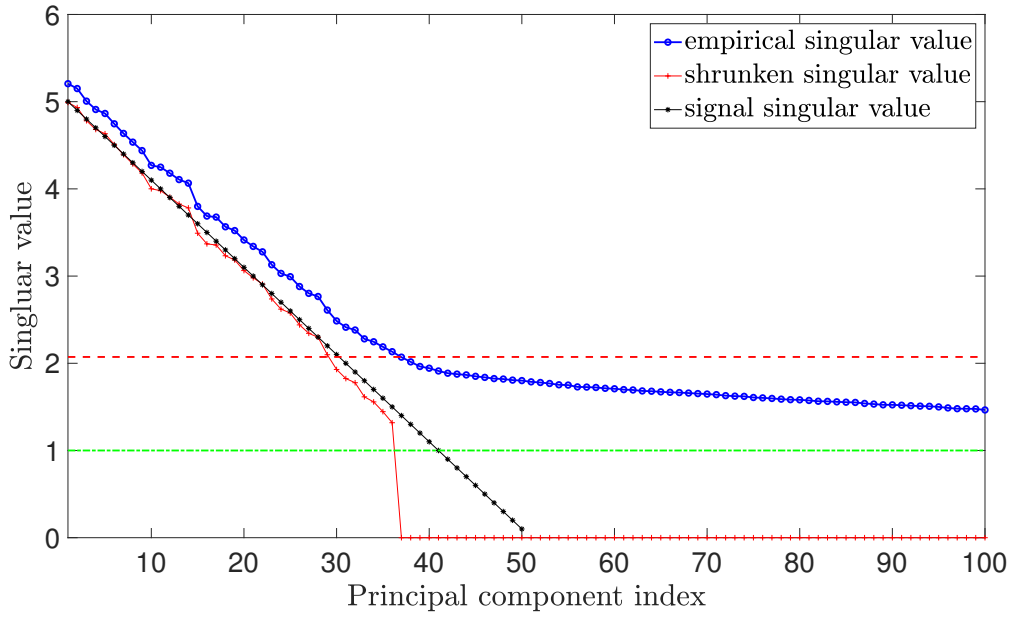
(a) Square Gaussian noise matrix $\mathbf{X}_1$.



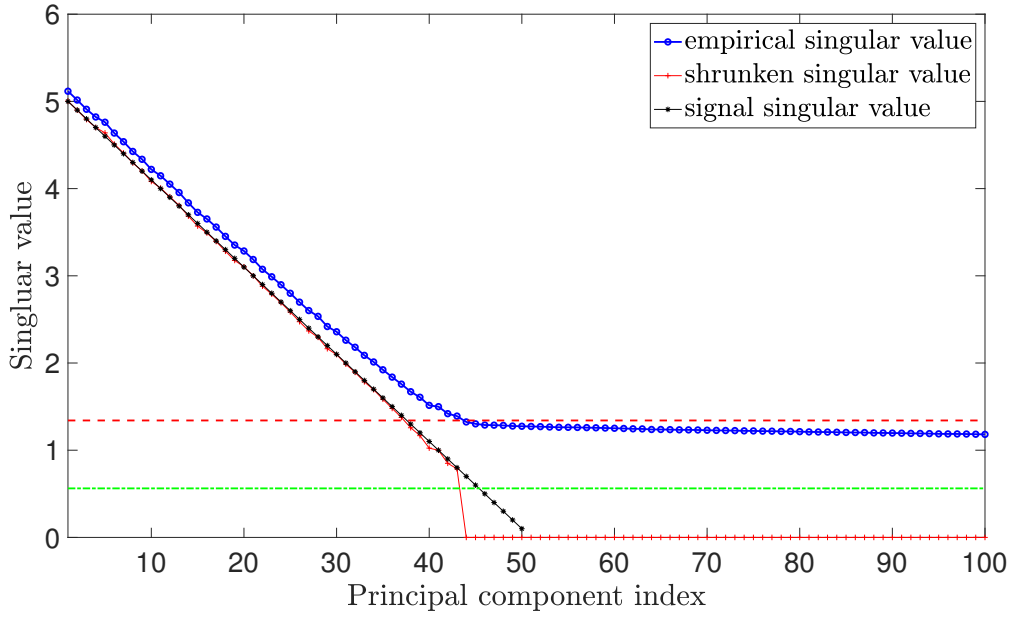
(b) Non-square Gaussian noise matrix $\mathbf{Y}_1$.

Figure 5.7: Scree plots of singular values for data matrices with Gaussian noise. The x-axis represents the principal component index, and the y-axis shows the singular values. The black stars show the values of the true signal singular values. The blue circles and red pluses show the values of the empirical and shrunken singular values respectively. The horizontal red dashed lines show the empirical singular value thresholds. The horizontal green dashed lines show the value $\beta^{1/4}$ in each matrix, which is the cutoff for distinguishable signal singular values. This shows the singular value shrinkage function in (5.7) works well in the Gaussian noise case.

As seen in both Panel 5.8(a) and 5.8(b) of Figure 5.8, the shrunken singular values match very well with the underlying signal singular values. This illustrates that the optimal shrinkage function in (5.7) gives good estimates of signal singular value under non-Gaussian noise case.



(a) Square Student t noise matrix $\mathbf{X}_2$.



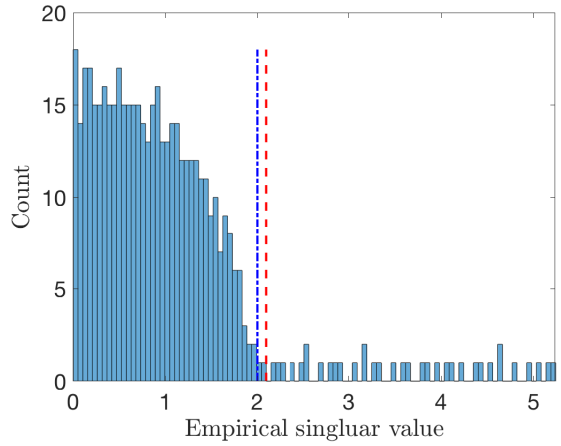
(b) Non-square Student t noise matrix $\mathbf{Y}_2$.

Figure 5.8: Scree plot of singular values for data matrices with Student t noise. This shows the singular value shrinkage function in (5.7) works well in the non-Gaussian noise case.

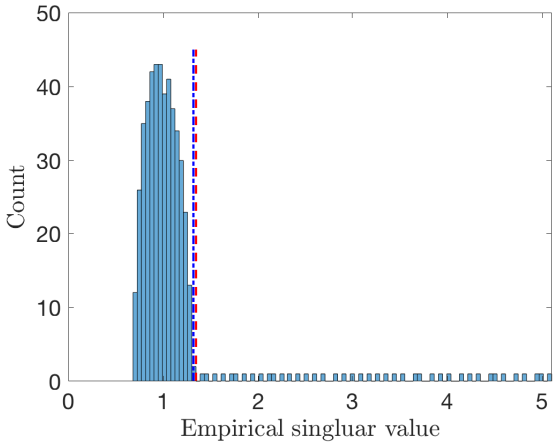
Figure 5.9 shows the histograms of the empirical singular values of all four data matrices. The x-axis represents the empirical singular values and the y-axis shows the counts of each bin. The vertical green dot-dashed lines show the theoretical values of the bulk edge, $1 + \sqrt{\beta}$. On the left of the vertical green dot-dashed line, we can see the histograms of the singular values are in the shape of a (deformed) quarter circle. For the non-square matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$, the deformed quarter circle is more compact. The vertical red dashed lines show the values of the estimated singular value thresholds, which are slightly larger than the actual values of the bulk edge. Figure 5.9 indicates that the singular values of both Gaussian and non-Gaussian noise matrices follow the asymptotic results in Section 5.2.3.

Figure 5.10 shows the direction specific perturbation angles for the principal components in the estimated signal row spaces of $\mathbf{X}_1$ and $\mathbf{Y}_1$ in the upper and lower panels respectively. The x-axis represents the component index. The y-axis is the perturbation angles. The black dots show the values of the true perturbation angles. The blue and green dots show the values of 95th and 5th percentiles of the resampled direction specific perturbation angles for each principal component. Figure 5.7 shows that the true perturbation angles for each principal component are almost between the 5th and 95th percentiles of the resampled perturbation angle. This implies our estimation algorithm in Section 5.3 works well for Gaussian datasets. Moreover, the perturbation angle tends to increase as the component index increases. This is consistent with our analysis of perturbation angle in Section 5.2.4. The perturbation bound from Cai et al. (2018) is 90°in this toy example, which is not informative. This is due the fact that there is no gap between signal and noise singular values. The original Wedin bound from Wedin (1972) is also not informative for the same reason. The modified Wedin bound in Feng et al. (2017) is applicable and shown as the horizontal dashed cyan lines in Figure 5.10. The Wedin bound is uniform for all principal components and conservative as expected. The horizontal red dashed lines show the half values of the random direction bound. It indicates only the rank 31 and 42 right singular matrices of $\mathbf{X}_1$ and $\mathbf{Y}_1$ are useful for the direction specific perturbation angle estimation.

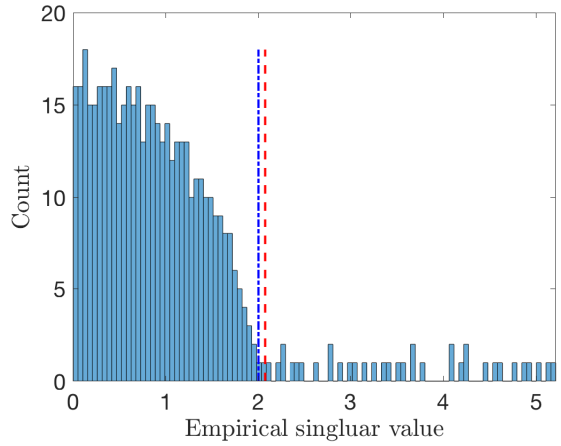
Figure 5.11, which is analogous to Figure 5.10, shows the direction specific perturbation angles for the principal components in the estimated signal row spaces of $\mathbf{X}_2$ and $\mathbf{Y}_2$. Similar to the Gaussian noise case, the true direction specific angles of first 25 PCs in $\mathbf{X}_2$ and first 30 PCs in $\mathbf{Y}_2$ are mostly in the range of 5th and 95th percentiles of the resampled angles in both square and



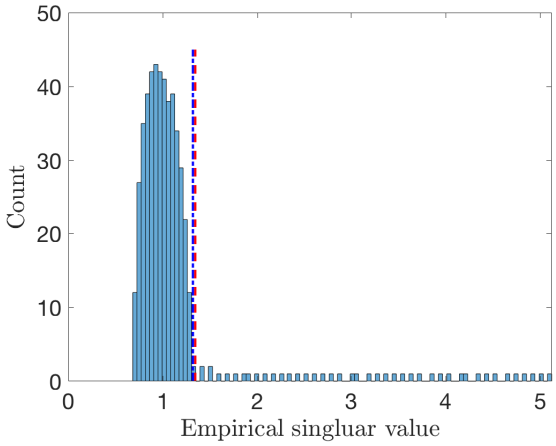
(a) Square Gaussian noise matrix $\mathbf{X}_1$.



(b) Non-square Gaussian noise matrix $\mathbf{Y}_1$.

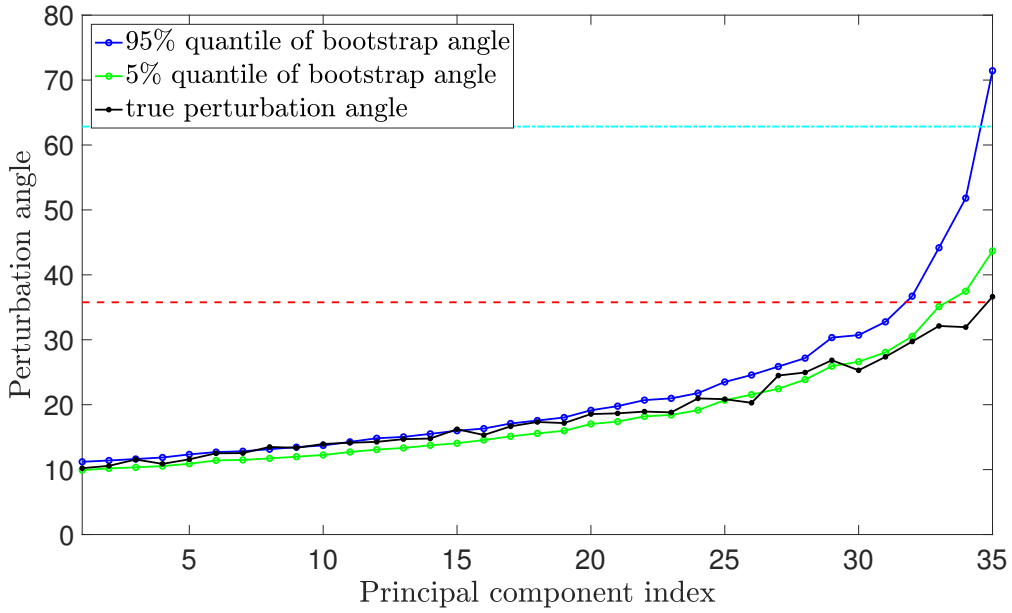


(c) Square Student t noise matrix $\mathbf{X}_2$.

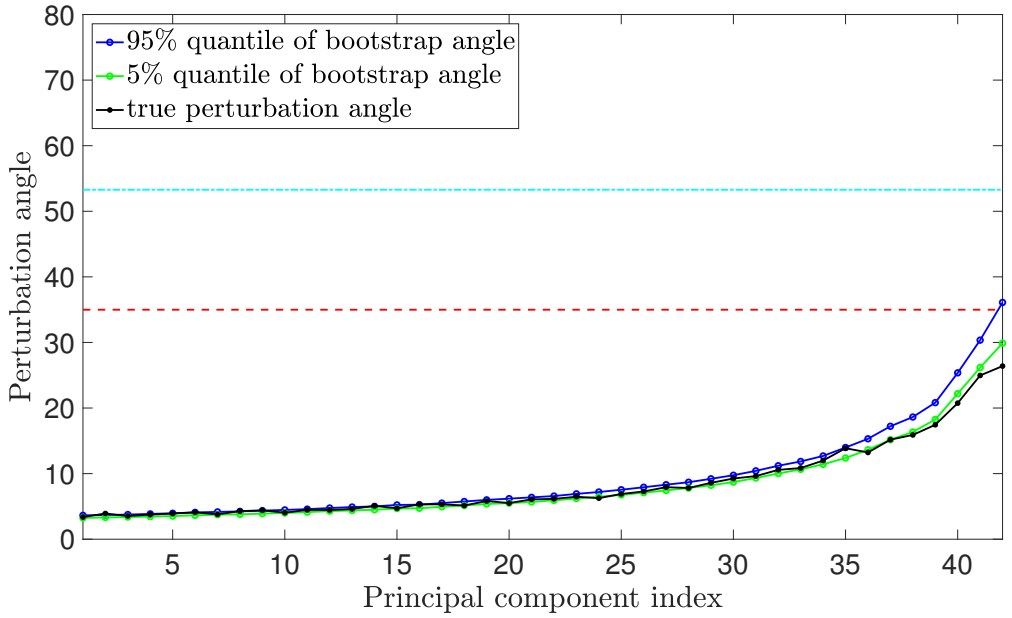


(d) Non-square Student t noise matrix $\mathbf{Y}_2$.

Figure 5.9: Histogram of empirical singular values. The x-axis shows the empirical singular value and the y-axis shows the count of each bin. The vertical red dashed lines show the empirical singular value thresholds. The vertical green dashed lines show the theoretical bulk edge, $1 + \sqrt{\beta}$, in each matrix. This figure shows that the distributions of the empirical singular values from both Gaussian and non-Gaussian noise matrices follow the generalized quarter circle law.



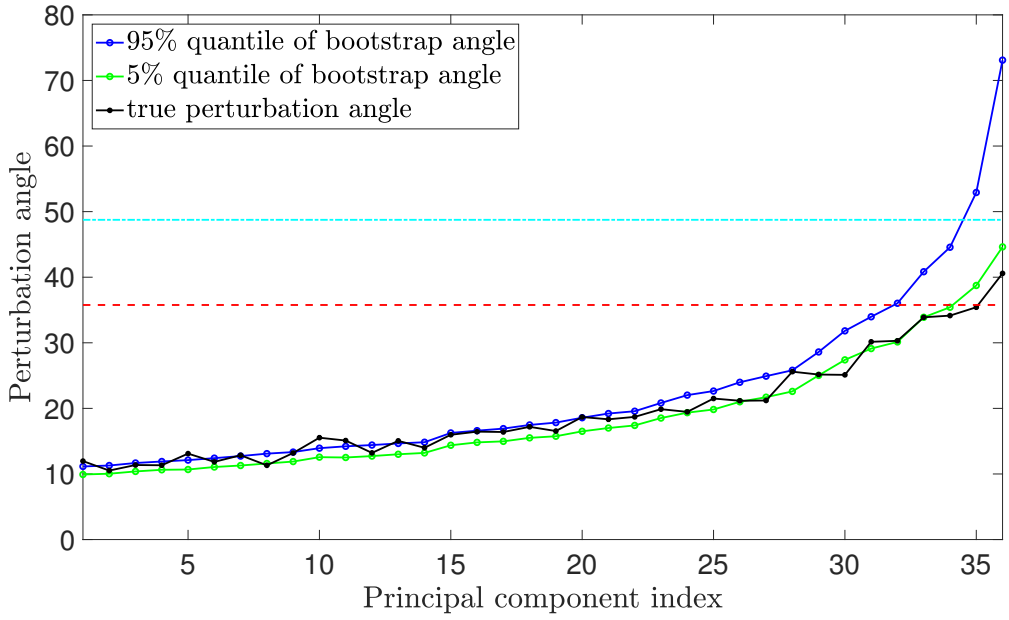
(a) Square Gaussian matrix $\mathbf{X}_1$.



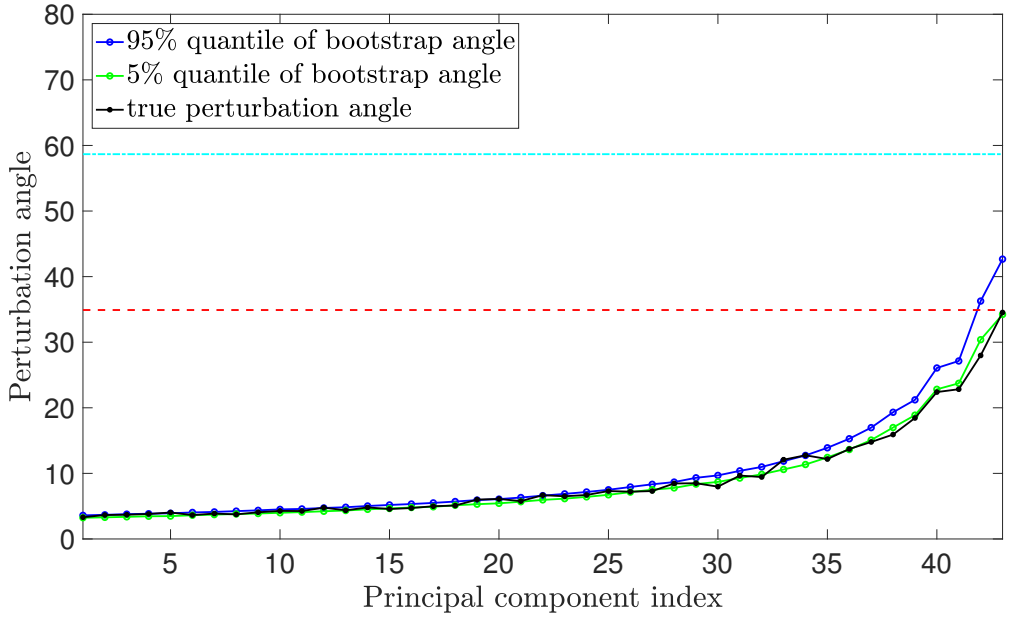
(b) Non-square Gaussian matrix $\mathbf{Y}_1$.

Figure 5.10: Direction specific perturbation angles for empirical principal components for data matrices with Gaussian noise. The black dots show the values of true perturbation angle. The blue and green dots show the values of the 95th and 5th percentiles of the resampled perturbation angles. The horizontal dashed cyan lines show the values of the Wedin bound. The horizontal red dashed lines show the half values of the random direction bound. This figure indicates the direction specific estimation algorithm in Section 5.3 works well for the Gaussian noise case.

non-square matrices. For the PCs with small variance, the true perturbation angles are below 5th percentile of resampled angles. After adjusting by the random angle bound, the adjusted ranks of $\mathbf{X}_2$ and $\mathbf{Y}_2$ are 31 and 41 respectively.



(a) Square Student t matrix $\mathbf{X}_2$.



(b) Non-square Student t matrix $\mathbf{Y}_2$.

Figure 5.11: Direction specific perturbation angles for empirical principal components for data matrices with Student t noise.

5.4.3 AJIVE directions perturbation analysis

In this subsection, the perturbation analysis framework is applied to each joint basis vector from AJIVE algorithm on toy dataset in Section 3.1.1, with correct input ranks $\hat{r}_x = 2, \hat{r}_y = 3$. For each basis vector, the angle between itself and each data block is bounded as in (5.20) by applying the algorithm in Section 5.3.

The results of direction specific perturbation analysis are shown in Table 5.1. The first column lists each data block. The second and third columns summarize the angle ranges for the first and second joint basis vectors. As seen in Table 5.1, the true values of θ are always in the range of the estimated lower and upper bounds. For each basis vector, the projection angles $\hat{\theta}$ are the same for each data block. This is due to the fact that the AJIVE basis vector are the average of the principal vectors in each data block. Thus the basis vector lies in the middle of the two estimated signal spaces. However, the lengths of each angle range on the two data blocks are different since each vector has different perturbation angle nn each data block. For the first joint basis vector, $\mathbf{v}_{\mathbf{M},1}$, the lower bounds of its angle between row$(\hat{\mathbf{A}})_1$ and row$(\hat{\mathbf{A}}_2)$, the estimated row spaces of $\mathbf{X}$ and $\mathbf{Y}$, are 0°and 3.22°respectively, which indicates $\mathbf{v}_1$ could be in row$(\mathbf{A}_1)$ but outside of row$(\mathbf{A}_2)$. The corresponding upper bounds on row$(\hat{\mathbf{A}})_1$ and row$(\hat{\mathbf{A}}_2)$ are 17.02°and 7.58°, which implies its perturbation on $\hat{\mathbf{A}}_1$ is larger than that on $\hat{\mathbf{A}}_2$. Combining these two results, we can see that the estimation of the first joint basis vector is not accurate. This is because the AJIVE optimization problem searches the direction merely minimizing the angles among each estimated signal space without taking into account the perturbation effects. Moreover, the perturbation analysis in the AJIVE algorithm is also too conservative to judge which of the right singular vectors in (3.6) are belonging to the joint space.

Table 5.1: Perturbation analysis for AJIVE directions on the toy dataset.

	$\mathbf{v}_{\mathbf{M},1}$				$\mathbf{v}_{\mathbf{M},2}$			
	lower	$\hat{\theta}$	upper	θ	lower	$\hat{\theta}$	upper	θ
$\mathbf{X}$	0°	5.5°	17.0°	6.0°	10.7°	23.6°	31.4°	25.1°
$\mathbf{Y}$	3.2°	5.5°	7.6°	5.5°	21.3°	23.6°	24.6°	23.4°

Adjusted AJIVE directions

A simple approach to adjust AJIVE directions by considering the perturbation effects is to construct $\bar{\mathbf{M}}$ by multiplying each $\mathbf{V}$ with $\cos^2 \phi_{\bar{r}_k}$, i.e.,

$$\bar{\mathbf{M}} \triangleq \begin{bmatrix} \tilde{\mathbf{V}}_1^\top * \cos^2 \hat{\phi}_{\bar{r}_1} \\ \vdots \\ \tilde{\mathbf{V}}_K^\top * \cos^2 \hat{\phi}_{\bar{r}_K} \end{bmatrix} = \mathbf{U}_{\bar{\mathbf{M}}} \mathbf{\Sigma}_{\bar{\mathbf{M}}} \mathbf{V}_{\bar{\mathbf{M}}}^\top. \tag{5.22}$$

Then the column vectors in $\mathbf{V}_{\bar{\mathbf{M}}}$ will be more close to the row$(\mathbf{A}_k)$s with smaller max perturbation angle, $\hat{\phi}_{\bar{r}_k}$. Applying the direction specific perturbation analysis on the adjusted AJIVE directions, i.e., column vectors of $\mathbf{V}_{\bar{\mathbf{M}}}$, the ranges of the angle between each direction and each data block are shown in Table 5.2. As expected, $\mathbf{v}_{\bar{\mathbf{M}},1}$ is adjusted to be more close to row$(\hat{\mathbf{A}}_2)$ and far away from row$(\hat{\mathbf{A}}_1)$. Notice that $\mathbf{v}_{\bar{\mathbf{M}},1}$ is more close to both row$(\mathbf{A}_1)$ and row$(\mathbf{A}_2)$ than $\mathbf{v}_{\mathbf{M},1}$. This shows taking into account the perturbation effects can improve AJIVE estimation.

Table 5.2: Perturbation analysis for the adjusted AJIVE directions of the toy dataset.

	$\mathbf{v}_{\bar{\mathbf{M}},1}$				$\mathbf{v}_{\bar{\mathbf{M}},2}$			
	lower	$\hat{\theta}$	upper	θ	lower	$\hat{\theta}$	upper	θ
$\mathbf{X}$	0°	5.9°	19.1°	5.7°	10.7°	25.6°	34.6°	27.1°
$\mathbf{Y}$	2.7°	5.1°	7.2°	5.1°	19.1°	21.5°	22.5°	21.4°

However, the lower bound of the angle between $\mathbf{v}_{\mathbf{M},1}$ and row$(\mathbf{A}_2)$ is still greater than 0, which indicates that $\mathbf{v}_{\bar{\mathbf{M}},1}$ is not a joint direction by our definition. This motivates us to construct a new optimization problem to take into account the direction specific perturbation angles in a more sophisticated way. Under the direction specific perturbation analysis framework, the joint space basis vector lies in the region,

$$R1 = \{\mathbf{v} \in \mathbb{R}^n | \mathbf{v}^\top \hat{\mathbf{V}}_1 \hat{\mathbf{V}}_1^\top \mathbf{v} \geq \cos^2 \hat{\phi}_{\bar{r}_1} \|\mathbf{v}\|^2, \mathbf{v}^\top \hat{\mathbf{V}}_2 \hat{\mathbf{V}}_2^\top \mathbf{v} \geq \cos^2 \hat{\phi}_{\bar{r}_2} \|\mathbf{v}\|^2\}. \tag{5.23}$$

For this particular toy example, we can obtain a feasible solution of the 1-d joint space in (5.23) from the first pair of the principal vectors, which can be obtained from the left singular vectors of $\mathbf{M}$ (see Section 4.4). Denote the first pair of the principal vectors as $\mathbf{p}_1 \in \hat{\mathbf{V}}_1$ and $\mathbf{q}_1 \in \hat{\mathbf{V}}_2$

respectively. It is easy to verify that the region $R2$,

$$R2 = \{\mathbf{v} \in \mathbb{R}^n | \operatorname{Corr}(\mathbf{v}, \mathbf{p}_1) \geq \cos \hat{\phi}_{\bar{r}_1}, \operatorname{Corr}(\mathbf{v}, \mathbf{q}_1) \geq \cos \hat{\phi}_{\bar{r}_2}\}, \tag{5.24}$$

is inside of region $R1$. In particular, the line segment,

$$L1 = \{\mathbf{v} = \alpha \mathbf{p}_1 + (1-\alpha)\mathbf{q}_1 \in \mathbb{R}^n | \operatorname{Corr}(\mathbf{v}, \mathbf{p}_1) \geq \cos \hat{\phi}_{\bar{r}_1}, \operatorname{Corr}(\mathbf{v}, \mathbf{q}_1) \geq \cos \hat{\phi}_{\bar{r}_2}, \alpha \in [0,1]\} \tag{5.25}$$

is also part of region $R2$. Moreover, $L1$ is empty if and only if $R2$ is empty. The middle point of $L1$ is used as the estimate of the joint basis vector. Denote this direction as $\mathbf{v}_{L1}$. Remark that the joint basis vector from AJIVE algorithm, $\mathbf{v}_{\mathbf{M},1}$, is in the direction of $\frac{1}{2}\mathbf{p}_1 + \frac{1}{2}\mathbf{q}_1$. However, $\frac{1}{2}\mathbf{p}_1 + \frac{1}{2}\mathbf{q}_1$ is not necessarily inside of $R1$ as well as $L1$.

Table 5.3: Perturbation analysis for the middle point of $L1$.

	$\mathbf{v}_{L1}$			
	lower	$\hat{\theta}$	upper	θ
$\mathbf{X}$	0°	9.8°	23.0°	3.0°
$\mathbf{Y}$	0°	1.2°	3.3°	1.9°

As shown in Table 5.3, $\mathbf{v}_{L1}$ is a joint direction under the direction specific perturbation framework. Indeed, $\mathbf{v}_{L1}$ is much closer to both row($\mathbf{A}_1$) and row($\mathbf{A}_2$) than $\mathbf{v}_{\mathbf{M},1}$ and $\mathbf{v}_{\bar{\mathbf{M}},1}$. Notice that $\mathbf{v}_{L1}$ is much closer to row($\hat{\mathbf{A}}_2$) than row($\hat{\mathbf{A}}_1$). This is due to the fact that the principal vectors on row($\hat{\mathbf{A}}_2$) are more reliable than those on row($\hat{\mathbf{A}}_1$). The more general cases of finding multi-block joint directions and partially shared joint directions are interesting open questions for future research.

5.5 AJIVE directions perturbation analysis on TCGA dataset

This section discusses the application of the direction specific perturbation analysis on the TCGA dataset in Section 3.4.1. Applying the optimal singular value shrinkage estimators in (5.9) on the four data blocks, the estimated ranks of GE, CN, RPPA and Mutation are 210, 271, 58 and 209 respectively. Adjusted by the random angle bound, the updated estimated ranks of each data block are 144, 226, 46 and 142 respectively. The maximum perturbation angles on the low

rank approximations of each data block are 20.2°, 15.0°, 30.8°and 20.8°. Table 5.4 summarizes the estimated angle ranges of each AJIVE direction based on low rank approximations of each data block. It indicates that $\mathbf{v}_{\mathbf{M},1}$ is the only four-block joint direction and $\mathbf{v}_{\mathbf{M},2}$ is a three-block joint direction, which excludes the Mutation data block. This is consistent with the analysis in Section 3.4.1. Moreover, we also identified other partially shared joint structure for the rest AJIVE directions: $\mathbf{v}_{\mathbf{M},2}$ is another three-block joint direction excluding Mutation; $\mathbf{v}_{\mathbf{M},4}, \mathbf{v}_{\mathbf{M},5}$ and $\mathbf{v}_{\mathbf{M},6}$ are two-block joint direction between CN and RPPA. However, carefully looking at the angle ranges of $\mathbf{v}_{\mathbf{M},4}, \mathbf{v}_{\mathbf{M},5}$ and $\mathbf{v}_{\mathbf{M},6}$ on each data block, the projection angles and the angle upper bounds on CN are much smaller than those on RPPA. This is due to the fact that the max perturbation angle on RPPA are much larger than CN. Thus it is not very reliable to conclude these directions are in the row space of RPPA. In fact, it indicates these directions are more likely to be adjusted to the joint directions between GE and CN.

Table 5.4: Perturbation analysis on the AJIVE directions of the TCGA data set in Section 3.4.1

	$\mathbf{v}_{\mathbf{M},1}$			$\mathbf{v}_{\mathbf{M},2}$			$\mathbf{v}_{\mathbf{M},3}$		
	lower	$\hat{\theta}$	upper	lower	$\hat{\theta}$	upper	lower	$\hat{\theta}$	upper
GE	0°	8.3°	13.0°	0°	12.2°	18.1°	0°	16.7°	25.5°
CN	0°	11.0°	14.2°	0°	14.7°	19.1°	0°	14.3°	20.6°
RPPA	0°	12.5°	18.6°	0°	18.9°	28.5°	0°	22.5°	34.5°
Mutation	0°	14.9°	21.7°	2.1°	22.9°	34.8°	9.0°	29.8°	46.7°
	$\mathbf{v}_{\mathbf{M},4}$			$\mathbf{v}_{\mathbf{M},5}$			$\mathbf{v}_{\mathbf{M},6}$		
	lower	$\hat{\theta}$	upper	lower	$\hat{\theta}$	upper	lower	$\hat{\theta}$	upper
GE	0°	17.5°	26.9°	0°	20.2°	31.3°	0°	19.6°	29.6°
CN	2.5°	17.6°	23.9°	3.7°	18.7°	26.4°	6.7°	21.7°	29.5°
RPPA	0°	27.8°	43.7°	0°	28.0°	41.2°	0°	29.3°	47.2°
Mutation	11.1°	32.0°	48.8°	10.9°	31.8°	48.0°	12.2°	33.0°	49.9°

Applying the adjusted AJIVE directions in Section 5.4.3, the updated angle ranges are presented in Table 5.5. As expected, the adjusted AJIVE directions are closer to CN, which has the smallest perturbation angle, and further away from RPPA, which has the largest perturbation angle. Now $\mathbf{v}_{\bar{\mathbf{M}},4}, \mathbf{v}_{\bar{\mathbf{M}},5}$ and $\mathbf{v}_{\bar{\mathbf{M}},6}$ become two-block joint directions between GE and CN, which is more reasonable since the corresponding projection angles and angle upper bounds are also smaller.

Table 5.5: Perturbation analysis on the adjusted AJIVE directions of the TCGA dataset.

	$\mathbf{v}_{\bar{\mathbf{M}},1}$			$\mathbf{v}_{\bar{\mathbf{M}},2}$			$\mathbf{v}_{\bar{\mathbf{M}},3}$		
	lower	$\hat{\theta}$	upper	lower	$\hat{\theta}$	upper	lower	$\hat{\theta}$	upper
GE	0°	8.3°	13.0°	0°	12.2°	18.2°	0°	16.3°	25.3°
CN	0°	9.9°	12.9°	0°	13.2°	17.4°	0°	12.9°	18.9°
RPPA	0°	14.5°	20.9°	0°	21.7°	31.7°	0°	26.8°	39.4°
Mutation	0°	14.2°	20.9°	1.0°	21.8°	33.5°	6.7°	27.6°	44.2°
	$\mathbf{v}_{\bar{\mathbf{M}},4}$			$\mathbf{v}_{\bar{\mathbf{M}},5}$			$\mathbf{v}_{\bar{\mathbf{M}},6}$		
	lower	$\hat{\theta}$	upper	lower	$\hat{\theta}$	upper	lower	$\hat{\theta}$	upper
GE	0°	16.7°	26.9°	0°	18.5°	29.2°	0°	17.4°	28.7°
CN	0°	14.9°	21.0°	0°	15.0°	21.4°	0.4°	15.4°	22.1°
RPPA	4.4°	35.2°	52.4°	5.5°	36.3°	50.6°	14.9°	45.7°	68.6°
Mutation	7.0°	27.8°	44.0°	7.5°	28.3°	43.9°	3.4°	24.3°	38.9°

BIBLIOGRAPHY

Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary reviews: computational statistics*, 5(2):149–179.

Ana, L. and Jain, A. K. (2003). Robust data clustering. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–128. IEEE.

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.

Beck, J. M., Young, V. B., and Huffnagle, G. B. (2012). The microbiome of the lung. *Translational Research*, 160(4):258–266.

Benaych-Georges, F. and Nadakuditi, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135.

Björck, Å. and Golub, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of computation*, 27(123):579–594.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Bousbia, S., Papazian, L., Saux, P., Forel, J. M., Auufray, J.-P., Martin, C., Raoult, D., and La Scola, B. (2012). Repertoire of intensive care unit pneumonia microbiota. *PLOS One*, 7:e32486.

Bray, J. R. and Curtis, J. T. (1957). An ordination of upland forest communities of southern wisconsin. *Ecological Monographs*, 27:325–349.

Cabanski, C. R., Qi, Y., Yin, X., Bair, E., Hayward, M. C., Fan, C., Li, J., Wilkerson, M. D., Marron, J. S., Perou, C. M., et al. (2010). Swiss made: Standardized within class sum of squares to evaluate methodologies and dataset elements. *PloS one*, 5(3):e9905.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982.

Cai, T. T., Zhang, A., et al. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27.

Chen, J., Chen, W., Zhao, N., Wu, M. C., and Schaid, D. J. (2016). Small sample kernel association tests for human genetic and microbiome association studies. *Genetic epidemiology*, 40(1):5–19.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519.

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, 72(7):5069–5072.

Dickson, R. P., Erb-Downward, J. R., Prescott, H. C., Martinez, F. J., Curtis, J. L., Lama, V. N., and Huffnagle, G. B. (2014). Cell-associated bacteria in the human lung microbiome. *Microbiome*, 2:28.

Ding, T. and Schloss, P. D. (2014). Dynamics and associations of microbial community types across the human body. *Nature*, 509:357–360.

Dozier, R. B. and Silverstein, J. W. (2007). On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices. *Journal of Multivariate Analysis*, 98(4):678–694.

Draper, B., Kirby, M., Marks, J., Marrinan, T., and Peterson, C. (2014). A flag representation for finite collections of subspaces of mixed dimensions. *Linear Algebra and its Applications*, 451:15–32.

Fan, J., Wang, W., and Zhong, Y. (2016). An l eigenvector perturbation bound and its application to robust covariance estimation. *arXiv preprint arXiv:1603.03516*.

Feng, Q., Jiang, M., Hannig, J., and Marron, J. (2017). Angle-based joint and individual variation explained. *Journal of multivariate analysis*.

Gavish, M. and Donoho, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053.

Gavish, M. and Donoho, D. L. (2017). Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152.

Geman, S. (1980). A limit theorem for the norm of random matrices. *The Annals of Probability*, pages 252–261.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Hanafi, M. and Kiers, H. A. (2006). Analysis of k sets of data, with differential emphasis on agreement between and within sets. *Computational Statistics & Data Analysis*, 51(3):1491–1508.

Hanafi, M., Kohler, A., and Qannari, E.-M. (2011). Connections between multiple co-inertia analysis and consensus principal component analysis. *Chemometrics and intelligent laboratory systems*, 106(1):37–40.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Hennig, C. (2015). fpc: Flexible procedures for clustering. r package version 2.1-10.

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLOS One*, 7:e30126.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Horst, P. (1961). Relations among m sets of measures. *Psychometrika*, 26(2):129–149.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Huckemann, S., Hotz, T., and Munk, A. (2010). Intrinsic shape analysis: Geodesic pca for riemannian manifolds modulo isometric lie group actions. *Statistica Sinica*, pages 1–58.

Jere, S., Dauwels, J., Asif, M. T., Vie, N. M., Cichocki, A., and Jaillet, P. (2014). Extracting commuting patterns in railway networks through matrix decompositions. In *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, pages 541–546. IEEE.

Jordan, C. (1875). Essai sur la géométrie à n dimensions. *Bulletin de la Société mathématique de France*, 3:103–174.

Jung, S., Dryden, I. L., and Marron, J. S. (2012). Analysis of principal nested spheres. *Biometrika*, 99(3):551–568.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, pages 433–451.

Koenig, S. M. and Truwit, J. D. (2006). Ventilator-associated pneumonia: diagnosis, treatment, and prevention. *Clinical microbiology reviews*, 19(4):637–657.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t–distributions and their applications*. Cambridge University Press.

Kühnle, O. (2011). *Integration of multiple high-throughput data-types in cancer research*. PhD thesis, Ludwig Maximilian University of Munich.

Kuligowski, J., Pérez-Guaita, D., Sánchez-Illana, Á., León-González, Z., de la Guardia, M., Vento, M., Lock, E. F., and Quintás, G. (2015). Analysis of multi-source metabolomic data using joint and individual variation explained (jive). *Analyst*.

La Rosa, P. S., Shands, B., Deych, E., Zhou, Y., Sodergren, E., Weinstock, G., and Shannon, W. D. (2012). Statistical object data analysis of taxonomic trees from human microbiome data. *PLOS One*, 7:doi:10.1371/journal.pone.0048996.

Lê Cao, K.-A., Martin, P. G., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics*, 10(1):34.

Lee, M. H. (2007). *Continuum direction vectors in high dimensional low sample size data*. PhD thesis, University of North Carolina at Chapel Hill.

Lee, S. (2016). High-dimension, low sample size asymptotics of canonical correlation analysis. *arXiv preprint arXiv:1609.02992*.

Li, H. (2014). Microbiome, metagenomics and high-dimensional compositional data analysis. In *Banff International Research Station for Mathematical Innovation and Discovery (BIRS) Workshop Lecture Videos*. Banff International Research Station for Mathematical Innovation and Discovery.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616.

Lock, E. F., Hoadley, K. A., Marron, J., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523.

Löfstedt, T., Hoffman, D., and Trygg, J. (2013). Global, local and unique decompositions in onpls for multiblock data analysis. *Analytica chimica acta*, 791:13–24.

Lozupone, C. and Knight, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235.

Lu, X., Marron, J. S., and Haaland, P. (2014). Object oriented data analysis of cell images. *Journal of the American Statistical Association*, 109:548–559.

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.

Marron, J. S. and Alonso, A. M. (2014a). Overview of object oriented data analysis. *Biometrical Journal*, 56:732–753.

Marron, J. S. and Alonso, A. M. (2014b). Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753.

Matsen, F. A. and Evans, S. N. (2013). Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PloS one*, 8(3):e56859.

Miao, J. and Ben-Israel, A. (1992). On principal angles between subspaces in R^n. *Linear algebra and its applications*, 171:81–98.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250.

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. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

Nielsen, A. A. (2002). Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE transactions on image processing*, 11(3):293–305.

NIH HMP Working Group (2009). The nih human microbiome project. *Genome Research*, 19:2317–2323.

O'Connell, M. J. and Lock, E. F. (2016). R. JIVE for exploration of multi-source molecular data. *Bioinformatics*, 32(18):2877 – 2879.

O'Rourke, S., Vu, V., and Wang, K. (2013). Random perturbation of low rank matrices: Improving classical bounds. *arXiv preprint arXiv:1311.2657*.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. In *BMC proceedings*, volume 1, page S119. BioMed Central.

Parkhomenko, E., Tritchler, D., Beyene, J., et al. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34.

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

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*, volume 77. Springer, New York, 1 edition.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, 2 edition.

Ray, P., Zheng, L., Lucas, J., and Carin, L. (2014). Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, 30(10):1370–1376.

Samarov, D. V. (2009). *The analysis and advanced extensions of canonical correlation analysis*. PhD thesis, University of North Carolina at Chapel Hill.

Schouteden, M., Van Deun, K., Pattyn, S., and Van Mechelen, I. (2013). Sca with rotation to distinguish common and distinctive information in linked data. *Behavior research methods*, 45(3):822–833.

Schouteden, M., Van Deun, K., Wilderjans, T. F., and Van Mechelen, I. (2014). Performing DISCO-SCA to search for distinctive and common information in linked data. *Behavior Research Methods*, 46(2):576–587.

Segal, L. N., Alekseyenko, A. V., Clemente, J. C., Kulkarni, R., B, W., Chen, H., Berger, K. I., Goldring, R. M., Rom, W. N., Blaser, M. J., and Weiden, M. D. (2014). Enrichment of lung microbiome with supraglottic taxa is associated with increased pulmonary inflation. *Microbiome*, 2:21.

Segata, N., Börnigen, D., Morgan, X. C., and Huttenhower, C. (2013). Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, 4.

Sen, S. K., Foskey, M., Marron, J. S., and Styner, M. A. (2008). Support vector machine for data on manifolds: An application to image analysis. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1195–1198. IEEE.

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., Zhu, H., and Marron, J. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica*, 26(4):1747.

Smilde, A. K., Westerhuis, J. A., and de Jong, S. (2003). A framework for sequential multiblock component methods. *Journal of chemometrics*, 17(6):323–337.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, pages 33–40.

Stewart, G. and Sun, J.-g. (1990). *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press.

Trygg, J. and Wold, S. (2003). O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of chemometrics*, 17:53–64.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. *Journal of econometrics*, 4(2):147–166.

Waaijenborg, S., de Witt Hamer, P. V., Zwinderman, A. H., et al. (2008). Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(1):3.

Wang, H. and Marron, J. S. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35:1849–1873.

Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

Wei, S., Lee, C., Wichers, L., and Marron, J. (2016). Direction-projection-permutation for high-dimensional hypothesis tests. *Journal of Computational and Graphical Statistics*, 25(2):549–569.

Westerhuis, J. A., Kourti, T., and MacGregor, J. F. (1998). Analysis of multiblock and hierarchical pca and pls models. *Journal of chemometrics*, 12(5):301–321.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515 – 534.

Wold, H. (1975). Path models with latent variables: The nipals approach. *Quantitative sociology: International perspectives on mathematical statistical model building*.

Wold, H. (1985). Partial least squares. *Encyclopedia of statistical sciences*.

Wold, S., Geladi, P., Esbensen, K., and Öhman, J. (1987). Multi-way principal components-and pls-analysis. *Journal of chemometrics*, 1(1):41–56.

Wold, S., Kettaneh, N., and Tjessem, K. (1996). Hierarchical multiblock pls and pc models for easier model interpretation and as an alternative to variable selection. *Journal of chemometrics*, 10(5-6):463–482.

Wu, C., Chen, J., Kim, J., and Pan, W. (2016). An adaptive association test for microbiome data. *Genome medicine*, 8(1):1.

Yamasaki, K., Kawanami, T., Yatera, K., Fukuda, K., Noguchi, S., Nagata, S., Nishida, C., Kido, T., Ishimoto, H., Taniguchi, H., and Mukae, H. (2013). Significance of anaerobes and oral bacteria in community-acquired pneumonia. *PLOS One*, 8:10.1371/journal.pone.0063103.

Yang, Z. and Michailidis, G. (2015). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1 – 8.

Yu, Q., Risk, B. B., Zhang, K., and Marron, J. (2017). Jive integration of imaging and behavioral data. *NeuroImage*, 152:38–49.

Zhang, Y., Zhou, G., Jin, J., Wang, X., and Cichocki, A. (2015). Ssvep recognition using common feature analysis in brain–computer interface. *Journal of neuroscience methods*, 244:8–15.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015). Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *The American Journal of Human Genetics*, 96(5):797–807.

Zhou, G., Cichocki, A., Zhang, Y., and Mandic, D. P. (2016). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE transactions on neural networks and learning systems*, 27(11):2426–2439.