INTEGRATIVE ANALYSIS OF VARIATION STRUCTURE IN HIGH-DIMENSIONAL MULTI-BLOCK DATA

by

Sungwon Lee

BA, Yonsei University, 2003

MA, State University of New York at Buffalo, 2010

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich Graduate School of Arts and Science in

partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH GRADUATE SCHOOL OF ARTS AND SCIENCE

This dissertation was presented

by

Sungwon Lee

It was defended on

October 19th 2016

and approved by

Sungkyu Jung, Statistics Department

Satish Iyengar, Statistics Department

Kehui Chen, Statistics Department

George Tseng, Biostatistics Department

Dissertation Director: Sungkyu Jung, Statistics Department

INTEGRATIVE ANALYSIS OF VARIATION STRUCTURE IN HIGH-DIMENSIONAL MULTI-BLOCK DATA

Sungwon Lee, PhD

University of Pittsburgh, 2016

The multi-block data stand for the data situation where multiple data sets possibly from different platforms are measured on common subjects. This data type is ubiquitous in modern sciences. Moreover, data become increasingly high-dimensional. For example, in genetic studies, it is common to evaluate gene expression, microRNA and DNA methylation levels on a single tissue sample and, thanks to the advancing microarray technology, scientists examine thousands of genes in a single experiment. Separate analyses of individual data sets will not capture critical association relations among them that could encode valuable information for better understanding of the target subjects. Currently, there is a strong need for new statistical methods of analyzing high-dimensional multi-block data in an integrative and unified way.

This dissertation consists of three parts whose shared theme is to identify meaningful variations in multi-block data that account for the complex associations between component data sets. The found variations are then utilized for various statistical purposes: characterizing data in a precise and interpretable way; estimating weights in calculating scores of data that give maximal correlation; identifying the dynamics of how ancillary data affect variations over multi-block data; serving as an effective dimension reduction for classification. In the first part, we propose a non-linear extension of functional principal component analysis to effectively catch major variabilities in functional data exhibiting both amplitude and phase variations by taking into account the associations between those two variations. The second topic is an asymptotic study of the canonical correlation analysis where dimension grows and sample size remains fixed. In the third part, we devise a supervised multi-block data factorization scheme that decomposes the primary data sets with guidance from auxiliary data sets. Estimated layers of the resulting decomposition provide detailed information on variation structures and supervision effects. The advantages of an integrative analysis

of multi-block data will be demonstrated by simulation studies and real data applications such as pediatric growth curve, lip motion and gene expression-microRNA data analyses.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Sungkyu. I feel very lucky to have you as my advisor during my doctoral study in statistics at University of Pittsburgh. Without your dedicated and patient supervising, this thesis could not exist at this moment. I still remember the time back early 2013 when you handed the first project over to me saying it is an easy one. To be honest with you now, it was really hard to me at that time. Since then, you have guided me through numerous critical situations to an independently-functioning statistician with sometimes encouraging words, sometimes constructive criticisms, sometimes fruitful discussions and always enlightening advice. I am deeply indebted to you for your being the inspiration and motivation throughout those arduous years of my graduate study.

Dear committee members, Satish, Kehui and George. I am expressing my deep thanks to all of you for squeezing your busy time to read my manuscript, deliver insightful comments at my defense and convey heartful congratulations when I defended. The moment shall be hovering in my memory for life.

Finally, my parents. Simply, I owe my everything to your constant love and support. I would like to pay my highest gratitude to your always standing by me in joy and in sorrow. Knowing that there is somebody somewhere in this planet who fully believes in me is the most essential source in taking courage to keep moving forward for what I want to do.

In retrospect, my Ph.D. years were precious, rewarding and shaped me to be mature both academically and personally.

1.0 INTRODUCTION

1.1 BACKGROUND, MOTIVATION AND PROBLEMS

In this dissertation, we addresses development, theoretical study and implementation of statistical methods for the integrated analysis of multi-block data. We refer to multi-block data as a collection of multiple sets of variables measured on a same set of subjects where the different sets represent different aspects of subject. The multi-block data are otherwise referred to as multi-view or multi-source data. As prominent examples where multi-block data can arise, consider the following multi-block data situations,

- High-throughput biomedical data: With recent proliferation of biomedical technologies, scientists can now obtain diverse types of measurements on a given set of sample tissues such as gene expression level, genotype information, DNA methylation rate and microRNA number.
- Financial data: Banks, in an effort to assess financial solvency of potential business debtors, collect their revenue, profit and liability trend data of the past years.
- Weather data: To better predict local weather, weather stations in a region deploy equipments that measure various atmospheric conditions such as temperature, humidity, barometric pressure and wind speed over a certain time grid.

Multi-block data are increasingly emerging as more new technologies are introduced, and data processing becomes steadily cheaper. The expanding amount of available multi-block data necessitates the development of systematic statistical methodologies accommodating multi-block nature of those data.

A distinctive feature of multi-block data is that there possibly exist dependencies among variables in constituent data sets as they belong to a same set of objects. The dependence structure between multiple data sets can be utilized to infer interesting patterns of the population of samples that would not be found with separate analyses of individual data sets. On the other hand,

dependence structure within variables in a specific data set could impart unique and useful information. In this case, removing the between-data-sets dependence structure and working with net within-data-set dependency structure will lead to a more effective and clearer inference. These considerations motivate us to set a broad goal of statistical research: Define, measure, validate and utilize dependence structure between and within component sets of multi-block data.

Dependence among variables is often summarized by a direction along which data exhibit meaningful variations. For example, if observations with three numeric variables show large variation at the direction of $[1, -2, 0]^T$, then we can say that the first and second variables are associated in a way that, as the first increases by 1, the second tends to decrease by -2. There are largely two ways of defining directions in multi-block data which make sense of dependence structure between multiple data sets. Let $\mathbf{X}_i$, for $i = 1, 2, \ldots, m$, be a $n \times p_i$ matrix containing measurements for the p_i variables of the ith data sets on a common set of n objects. The first approach seeks a direction vector ξ in the $p_1 + p_2 + \cdots + p_m$ dimensional row space of the concatenated matrix $\mathbf{X}$,

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m], \tag{1.1}$$

such that the variation of scores (of projection of $\mathbf{X}$ onto ξ) is maximized under a certain regularity. Accordingly, the $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ parts of the resulting direction ξ can be viewed as associated directions between the data sets, and variation of $\mathbf{X}$ along ξ (e.g., variance of $\mathbf{X}^T\xi$) can be thought of as a joint variation across the data sets. A commonly employed tool for the first approach is singular value decomposition (SVD) or its variants applied to $\mathbf{X}$. The second approach, handling each data set $\mathbf{X}_i$ seperately, seeks associated directions for each pair of $(\mathbf{X}_i, \mathbf{X}_j)$, $i \neq j$. A direction ψ_{X_i} in the row space of $\mathbf{X}_i$ and another direction ψ_{X_j} in the row space of $\mathbf{X}_j$ are found such that a strength of association between two sets of scores (of projection of $\mathbf{X}_i$ and $\mathbf{X}_j$ onto ψ_{X_i} and ψ_{X_j}, respectively) measured by a certain quantity such as Pearson correlation coefficient is maximized. Canonical correlation analysis (CCA) [20] and its variants are commonly used for the second approach. The first is an extension of PCA for understanding the dependence of multivariate data while the latter is an extension of the examination of correlation matrix. The latter can only capture dependencies between pairs of data blocks. We will consider both approaches.

Challenges that make multi-block data analysis more complicated arise when i) one or more data sets are high-dimensional (possibly higher dimension than a given sample size), ii) one or more data sets are manifold-valued, and iii) one set assumes a distinct role among data sets. As we will encounter some combinations of these situations, we briefly address the challenges associated with

these situations.

High-dimensional data are frequently encountered in modern scientific data, in which a large number (hundreds or thousands) of variables are measured for each object. Furthermore, the number of variables often exceeds the number of sample objects, resulting in High Dimension, Low Sample Size (HDLSS or the large p, small n) situation [15]. For example, due to advances in modern medical scanning technology, the resolution of obtained images becomes higher. However, the sample size stays low because of the high cost in obtaining medical images such as MRI. Standard statistical tools typically assume that the sample size is larger than the number of variables. When applied to HDLSS data, many of them suffer from a "too lean information to infer" situation. For example, when CCA is applied to HDLSS data, it is well-known that there exist infinitely many pairs of empirical canonical weight vectors with a perfect canonical correlation coefficient of 1.

Non-standard data arise where the observations do not have natural vector representations (e.g., algebraic data), or the sample space naturally form a curved manifold (e.g., directional data). Many of standard statistical methods that benefits from Euclidean geometry are not directly applicable to manifold-valued data, and if so, they often fail to provide legitimate statistics. As a schematic example, consider data on the unit sphere. Taking the arithmatic mean of two points a and b places the resulting statistic inside the sphere. A legitimate "mean" would be a mid-point between a and b *on the sphere*. Hence, when we have one or more non-standard data sets in multi-block data, a special care needs to be called for in analyzing them.

An example where one of the data sets in multi-block data assumes a distinct role from those of the rest is the case where one data set has a role of supervision over the joint variations of the rest of data sets. To elaborate the supervision effect intuitively, consider a motivating example of high-throughput biomedial data. Let $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$ contain gene expression levels, genotype information and DNA methylation rates, respectively. Oftentimes, we obtain an additional data set $\mathbf{Y}$ on the same set of tissues that inherently relates to the underlying joint variations of the multi-block data. Suppose that we have disease subtype information stored in $\mathbf{Y}$ for all samples. Conceptually, different disease subtypes may explain a large portion of genetic variations in individual data sets. In other words, the data set $\mathbf{Y}$, here called supervision, potentially drives the joint variations across data sets in the multi-block data. In this case, rather than looking for joint variations of a concatenated matrix of $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ and $\mathbf{Y}$ in (1.1), it is more natural and desirable to treat the two sets of data, $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ and $\mathbf{Y}$, separately and work with a concatenated matrix of $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$ to look for joint variations driven by the subtype information of $\mathbf{Y}$.

With background and motivations described so far, we propose to address the following three specific research questions,

- When component data sets in multi-block data are functional (of uncountably infinite dimensionality so automatically HDLSS data), and one of them form a sphere-valued data, how to deal with the situation?
- What is the behavior of CCA in the HDLSS situation?
- When a supervision data set is available,, how to identify joint variations across data sets in multi-block data driven by the supervision?

Chapter 2, 3, and 4 answer the first, second and third questions, respectively.

1.2 SUMMARY AND CONTRIBUTIONS

Chapter 2 proposes two exploratory data analysis techniques of probing the internal structure of functional data containing amplitude and phase variations. A commonly-employed framework for analyzing those data is to take away the phase variation by a function alignment method and then to use the aligned functions for subsequent analysis. Accordingly, the resulting statistics characterize the behaviors of the amplitude variation only. We propose two methods to bring all of the amplitude variation, the phase variation and their association structures into account. The first method effectively reveals major modes of variation in the original form of functions. In this method, called functional combined principal component analysis (FCPCA), amplitude and phase variations are joined into a single random function using the weight that provides the maximal explaining power of observed functions. On the other hand, the second method, called functional combined canonical correlation analysis (FCCCA) presents the combined effect of highly correlated pairs of amplitude and phase variations, also in the original form. Appropriate statistical models for these methods assume that the functional data are decomposed into two sets of functional data, one for the amplitude and the other for the phase variation. This decomposition is a natural set up for multi-block data analysis. It turns out that the phase part of the functional data naturally sits on the unit sphere in function space. To deal with this non-standard situation, these manifold-valued data are approximated by mapping to a linear space (tangent to the unit sphere) so that functional PCA and CCA can be employed to extract joint variations and correlated directions. As

	Angle		
	$\theta = 0°$	$0° < \theta < 90°$	$\theta = 90°$
$\alpha > 1$	Consistent	Inconsistent	Strongly inconsistent
$\alpha < 1$	Strongly inconsistent	Strongly inconsistent	Strongly inconsistent

Table 1. Summary on consistency of the first sample canonical weight vector.

we will see later in Chapter 2, our methods effectively capture interesting features of data structure and visualize results in an interpretable manner.

In Chapter 3, an asymptotic behavior of CCA is studied when dimension d grows and the sample size n is fixed (i.e., under the HDLSS situation). In particular, we are interested in the conditions for which CCA works or fails in the HDLSS situation. Let $X^{(d)}, Y^{(d)} \sim N_d(0, \Sigma)$ be two normally distributed random vectors. To ease the interpretation, we consider a special case where the population canonical weight vectors $\psi_X^{(d)}$ and $\psi_Y^{(d)}$ for $X^{(d)}$ and $Y^{(d)}$ are set to be,

$$\psi_X^{(d)} = \cos\theta_X \xi_{X1}^{(d)} + \sin\theta_X \xi_{X2}^{(d)}, \ \psi_Y^{(d)} = \cos\theta_Y \xi_{Y1}^{(d)} + \sin\theta_Y \xi_{Y2}^{(d)},$$

where, for $\alpha > 0$, $\xi_{X1}^{(d)}$ and $\xi_{X2}^{(d)}$ are the population principal component (PC) directions of X with corresponding PC variances d^α and 1, and $\xi_{Y1}^{(d)}$ and $\xi_{Y2}^{(d)}$ are population (PC) directions of Y with their PC variances d^α and 1. Note that the angle between $\psi_X^{(d)}$ and $\xi_{X1}^{(d)}$ is θ_X and that the angle between $\psi_Y^{(d)}$ and $\xi_{Y1}^{(d)}$ is θ_Y. The success and failure of CCA can be described by the consistency of the sample canonical weight vector $\hat{\psi}_X^{(d)}$ and $\hat{\psi}_Y^{(d)}$ with their population counterpart $\psi_X^{(d)}$ and $\psi_Y^{(d)}$ under the limiting operation of $d \to \infty$ and n fixed. Using the angle as a measure of consistency, we say that $\hat{\psi}_X^{(d)}$ (similarly $\hat{\psi}_Y^{(d)}$) is,

- Consistent with $\psi_X^{(d)}$ if angle$(\hat{\psi}_X^{(d)}, \psi_X^{(d)}) \xrightarrow{P} 0$ as $d \to \infty$,
- Inconsistent with $\psi_X^{(d)}$ if angle$(\hat{\psi}_X^{(d)}, \psi_X^{(d)}) \xrightarrow{P} a$, for $0 \le a \le \pi/2$, as $d \to \infty$,
- Strongly inconsistent with $\psi_X^{(d)}$ if angle$(\hat{\psi}_X^{(d)}, \psi_X^{(d)}) \xrightarrow{P} \pi/2$ as $d \to \infty$.

Strong inconsistency implies that the estimate $\hat{\psi}_X^{(d)}$ and $\hat{\psi}_Y^{(d)}$ become completely oblivious of their population counterparts and become arbitrary quantities, as indicated in the fact that $\pi/2$ is indeed the largest angle possible between two vectors. It turns out that the convergence of $\hat{\psi}_X^{(d)}$ and $\hat{\psi}_Y^{(d)}$ depends heavily on the size of the variance d^α of the population PC directions $\xi_{X1}^{(d)}$ and $\xi_{Y1}^{(d)}$. That

is, the estimates $\hat{\psi}_X^{(d)}$ and $\hat{\psi}_Y^{(d)}$ tend to converge to the PC population directions $\xi_{X1}^{(d)}$ and $\xi_{Y1}^{(d)}$ when their PC variance d^α is large enough ($\alpha > 1$). The critical conditions governing the consistency or inconsistency of empirical canonical weight vectors are summarized in Table 1. The sample canonical weight vector $\hat{\psi}_X^{(d)}$ (similarly $\hat{\psi}_Y^{(d)}$) is,

- Consistent with $\psi_X^{(d)}$ if $\alpha > 1$ and angle$(\psi_X^{(d)}, \xi_{X1}^{(d)}) \xrightarrow{P} 0$ as $d \to \infty$,
- Inconsistent with $\psi_X^{(d)}$ if $\alpha > 1$ and angle$(\psi_X^{(d)}, \xi_{X1}^{(d)}) \xrightarrow{P} \theta_X$, for $0 < \theta_X < \pi/2$, as $d \to \infty$,
- Strongly inconsistent with $\psi_X^{(d)}$ if $\alpha < 1$ or if $\alpha > 1$ and angle$(\psi_X^{(d)}, \xi_{X1}^{(d)}) \xrightarrow{P} \pi/2$ as $d \to \infty$.

The mathematical mechanism behind these statistical phenomena is the main topic of Chapter 3.

Chapter 4 proposes a framework for a systematic decomposition of variation in multi-block data when supervision information is available. In particular, we aim to identify joint variations across multiple data sets and individual variations specific to each data set that are partially or fully driven by supervision effects. To achieve this aim, we first attempt to extend JIVE (Joint and Individual Variation Explained) proposed in [35]. JIVE decomposes a multi-block data into the joint and individual variations, but supervision effect was not considered. Formally, JIVE decomposes multi-block data $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m)$ into a sum of three components: a low-rank matrix $\mathbf{J}$ capturing the joint structure across data sets, low-rank matrices $\mathbf{A}_i$ capturing the individual structure specific to each data set, and a noise matrix $\mathbf{E}$,

$$\begin{cases} \mathbf{X}_i = \mathbf{J}_i + \mathbf{A}_i + \mathbf{E}_i, \ i = 1, 2, \ldots, m, \\ \mathbf{J} = [\mathbf{J}_1, \mathbf{J}_2, \ldots, \mathbf{J}_m] = \mathbf{U}\mathbf{V}^T, \\ \mathbf{A}_i = \mathbf{U}_i \mathbf{V}_i^T, \end{cases}$$

where columns of $\mathbf{V}$ and $\mathbf{V}_i$'s contain loading vectors, and $\mathbf{U}$ and $\mathbf{U}_i$'s contain scores. Note that loading vectors in $\mathbf{V}$ contribute to variables across data sets $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ and so, with their scores in $\mathbf{U}$, constitute the joint variation structure. Similarly $\mathbf{U}_i \mathbf{V}_i^T$ represents the individual variation structure specific to the ith data set $\mathbf{X}_i$.

We extend the JIVE model by assuming that a supervision data set $\mathbf{Y}$ affects the joint variation structure. By adopting the model proposed in [32], called Supervised Singular Value Decomposition (SupSVD), we model $\mathbf{J}$ as,

$$\mathbf{J} = \mathbf{U}\mathbf{V}^T = (\mathbf{Y}\mathbf{B} + \mathbf{F})\mathbf{V}^T,$$

where $\mathbf{Y}$ contains the supervision information, $\mathbf{B}$ is responsible for conversion of $\mathbf{Y}$ into scores with respect to loading vectors in $\mathbf{V}$, and $\mathbf{F}$ is a random matrix. Intuitively, the first part of $\mathbf{J}$, $\mathbf{Y}\mathbf{B}\mathbf{V}^T$,

captures the variation that is driven by supervision $\mathbf{Y}$, and the second part $\mathbf{F}\mathbf{V}^T$ captures the variation that is irrelevant of $\mathbf{Y}$. We further extend the model by assuming that not only joint but also individual variation structures $\mathbf{A}_i$'s are possibly driven by their own supervisions $\mathbf{Y}_i$'s, where $\mathbf{A}_i$ is expressed as,

$$\mathbf{A}_i = \mathbf{U}_i\mathbf{V}_i^T = (\mathbf{Y}_i\mathbf{B}_i + \mathbf{F}_i)\mathbf{V}_i^T, \ i = 1, 2, \ldots, m.$$

Estimations of parameters for these two models can be performed efficiently by iteratively applying the modified EM algorithm iteratively. Unfortunately, the naive combination of JIVE and SupSVD exhibits several drawbacks. Among those, we point out the following.

- When we have more than two data sets, it may be reasonable to assume that every subset of them has its own variation structure, which leads to three different types of variation structures: the individual variation in a single data set, the full joint variation over whole data sets and the partial joint variation over multiple data sets but whole. The previous methods are incapable of modeling the partial joint variations.
- As the number of sets increases, the workload in estimating the ranks for individual and joint variations increases substantially.
- Supervision sets $\mathbf{Y}$ and $\mathbf{Y}_i$'s need to be pre-selected.

To address these issues, we propose a fully automated data-driven framework for the integrated analysis of multi-block data, termed Generalized Supervised Joint and Individual Variation Explained (G-SupJIVE). This framework assumes the following model,

$$\begin{cases} \mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m], \\ \mathbf{Y}^C = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_k], \\ \mathbf{X} = (\mathbf{Y}\mathbf{B} + \mathbf{F})\mathbf{V}^T + \mathbf{E}. \end{cases} \tag{1.2}$$

where $\mathbf{Y}^{\mathbf{C}}$ is a matrix that collects candidate supervisions. Model (1.2), unlike the previous model, does not separate joint and individual variations a priori but is flexible to model all three types of variations including the partial joint variation. We desire estimates of variation directions (columns of $\mathbf{V}$) to be interpretable (either the individual, partial joint, or fully joint structure) and also desire an adaptive selection of supervision from the candidate set $\mathbf{Y}^{\mathbf{C}}$. It turns out that imposing group-wise sparsity condition on columns of $\mathbf{B}$ and $\mathbf{V}$ in its estimating procedure fulfills our aim. To this end, we propose a sequential estimating procedure, where each step is a penalized likelihood maximization problem, with the group lasso penalty [55].

The advantages of the proposed method over other competing ones,

- Does not need to sepatate individual, partial joint and full joint variations a priori: algorithm sequentially estimates a variation direction without differentiating its type.
- Does not require pre-calculation of the rank for each variation; algorithm automatically stops when the rank of the multi-block data set $\mathbf{X}$ is exhausted.
- Incorporates multiple supervision sets; algorithm automatically choose a supervision set that drives a specific variation and estimates its supervision effect.

are demonstrated by a simulation study and a real-world application to Glioblastome Multiforme (GBM) cancer tumor data.

2.0 COMBINED ANALYSIS OF AMPLITUDE AND PHASE VARIATIONS IN FUNCTIONAL DATA.

2.1 INTRODUCTION

Functional data [44], observations measured over line and so best described as curve, known as functional data, are frequently encountered in modern sciences. When functional data consist of repeated measurements of some activity or development over time taken from a group of subjects, they often show a similar pattern of progression with two major variations, one being the amplitude variation and another the phase variation. As an example, human growth curves, which record height measurements of subjects over ages, share common events such as pubertal growth spurt and maturity. However, different curves develop those events with different magnitudes of height, which represents the amplitude variation, and at different temporal paces, which stands for the phase variation [10].

When the phase variation resides in functional data, a naïve application of the functional versions of standard statistical tools such as pointwise mean/variance, functional principal component analysis (FPCA) and functional canonical correlation analysis (FCCA) tend to yield misleading or inadequate results [11]. For a motivating example, growth velocity curves of 39 boys from Berkeley growth study [45] are plotted in Figure 1.(a). Visual inspection reveals a dominant trend where the boys who reach the phase of pubertal growth spurt (corresponding to the main peaks of curves) late have the tendency to show smaller maximum growth rate. In other words, there exists a clear association between phase and growth rate. An application of FPCA does not fully capture the important major association. This is exemplified in Figure 1.(b), illustrating the first mode of variation resulting from FPCA. The resulting mode, which indicates the resurgence of peaks of the boys with late pubertal growth spurt phase, is not capturing the dominant mode of variation in this example.

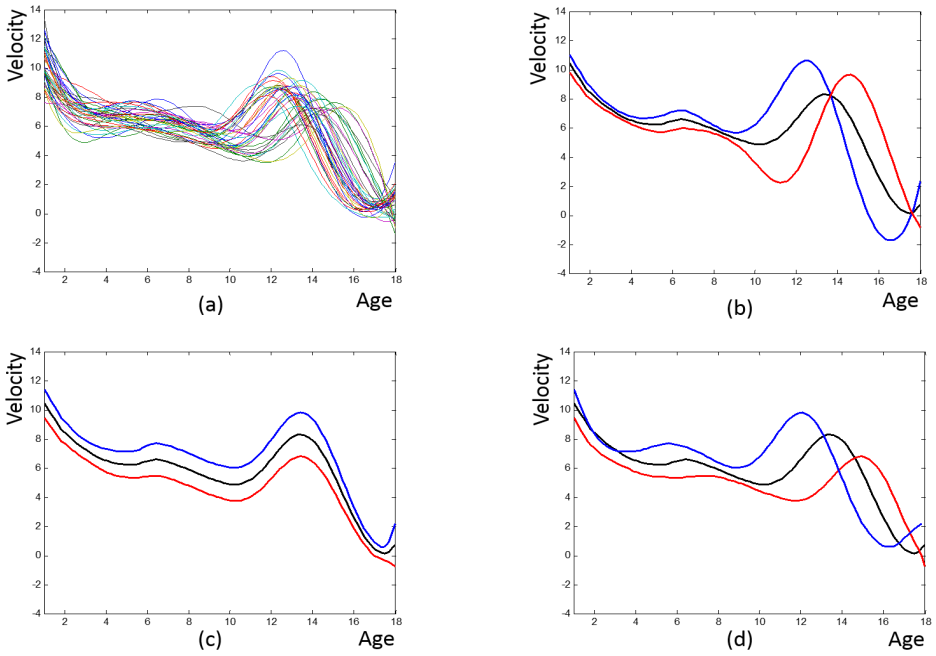

Figure 1. (a) 39 growth velocity curves. (b) Three functions describing a first mode of variation from FPCA to the original 39 curves. (c) Three functions describing a first mode of variation from FPCA to the aligned 39 curves. (d) Three functions describing a first mode variation from FCPCA to the original 39 curves.

To alleviate the adverse effect of the phase variation on statistical analysis of functional data, various methods of functional registration have been introduced. Landmark registration [25, 09, 03, 04] removes the phase variation by transforming the domain of each curve so that salient shape features of curves such as peaks and valleys are synchronized. Since landmarks are not always identifiable in all curves, flexible methods have been developed that find time-warping functions $\gamma(t)$ based on minimizing the distances among aligned curves $f(\gamma(t))$ in a suitable metric. Time shift model [50] uses $\gamma(t) = t + s$, where s is a random shift in time, and the studies [13, 21, 26, 34, 43, 52, 51] thereafter work with more general warping functions $\gamma(t)$ assumed to be continuous and non-linear. Most of the registration methods choose the usual L_2 metric for comparing functions. The unsatisfactory alignment results due to the asymmetry of the metric, i.e., $\|(f_1(t) - f_2)(t)\|_2 \neq \|(f_1(\gamma(t)) - f_2(\gamma(t))\|_2$ for two function f_1 and f_2, is discussed in [37]. Recently, registration by the use of the Fisher-Rao metric [51] is proposed that attains symmetry. Performances of their method are demonstrated in comparison study of [27].

A common practice in analyzing functional data is to first align curves to remove the phase

variation and then carry out standard functional data analysis techniques to the aligned functions. This practice is based on the belief that the phase variations are mostly immaterial. Accordingly, the resulting statistics deliver little information on the phase variation as illustrated in Figure 1.(c) while we see that the phase variation constitutes an integral part of variation in Figure 1.(a). Recently, several research works start to incorporate the phase variation into their frameworks for segmentation of periodic signals [27], clustering [49], functional regression [12, 14], classification of functional data [53] and manifold learning [06]. Our work attempts to reveal how the amplitude variation, the phase variation and their association structure combine to give rise to meaningful variabilities in the original function space.

In this chapter, we propose two exploratory data analysis techniques for investigating characteristics of the internal structure of functional data carrying amplitude and phase variations: Functional Combined Principal Component Analysis (FCPCA) and Functional Combined Canonical Correlation Analysis (FCCCA). FCPCA effectively reveals major modes of variation in the original form of functions as in Figure 1.(a), where amplitude and phase variations are combined to form the curves. The resulting first mode of variation of FCPCA plotted in Figure 1.(d) as compared to those of FPCA on the original and aligned data (see Figure 1.(b) and Figure 1.(c), respectively), clearly capture the dominant trend of the 39 growth velocity curves. FCCCA presents the combined effect of highly correlated pairs of amplitude and phase variations found by FCCA also in the original form of functions.

To this aim, we combine the aligned functions together with the warping functions used in the registration. The aligned functions represent the amplitude variation and warping functions represent the phase variation. Two major challenges need to be dealt with. First, the warping functions constitute a non-linear manifold in a function space, which complicates the statistical analysis of functional data. We circumvent the issue of non-linearity by borrowing a tool from differential geometry; a linear subspace best approximating the space of warping functions will be used. The second challenge is a need to address an association structure between amplitude and phase variations. For FCPCA, we model a single random function that joins the aligned functions and the warping functions. FPCA is then performed for the joined functions. The weights in joining the two sets of functions are adaptively chosen so that the resulting principal components (PCs) achieve the maximal explaining power of the observed functions in a suitable metric. On the other hand, FCCCA handles the aligned and warping functions separately and uses FCCA to deal with their associations. Subsequently, the resulting statistics from FCPCA and FCCCA

are transformed into a original form of functions that exhibits amplitude and phase variations for interpretation and visualization.

The rest of this chapter is organized as follows. Section 2.2 formulates the underlying models that generate the functional data based on which FCPCA and FCCCA are defined. Section 2.3 explains the estimation procedures of the parameters of FCPCA and FCCCA. The performance of the estimators is shown in Section 2.4 via a simulation study. In Section 2.5, we apply FCPCA and FCCCA to a synthetic data set, Berkeley growth data set [45] and a lip motion data set [42] to explore their major data structures. Section 2.6 contains technical details 2.8.

2.2 MODEL

2.2.1 Decomposition into two variations

We consider a smooth random function f that inherently contains amplitude and phase variations. The random function f is composed of two random functions y and γ,

$$f(t) = y \circ \gamma(t) = y(\gamma(t)), \ t \in [0,1]. \tag{2.1}$$

We restrict the domain of f to be exactly $[0,1]$ without losing generality. The random function y, which accounts for the amplitude variation, is a smooth square-integrable function on $[0,1]$, i.e.,

$$y \in L_2[0,1] = \{h : [0,1] \mapsto R \mid E\|h\|^2 = E\left(\int_0^1 h^2(t)dt\right) < \infty\}.$$

The random function γ is a time-warping function that introduces the phase variation to y,

$$\gamma \in \Gamma = \{h : [0,1] \mapsto [0,1] \mid h(0) = 0, h(1) = 1, h'(t) > 0 \ \forall \ t \in (0,1)\} \subset L_2[0,1],$$

where h' is a derivative of h. Each of the constraints on γ has its due justification: 1) constraint of $\gamma(0) = 0$ and $\gamma(1) = 1$ implies that the phase variation of f only occurs over the open interval $(0,1)$ with starting and ending times remaining fixed, 2) constraint of $\gamma'(t) > 0$ does not allow f to travel back into the past, which, in turn, makes γ invertible so that $y = f(\gamma^{-1})$. We assume that mean of the random time-warping function γ corresponds to an identity function $\gamma_{id}(t) = t$ (i.e., no phase variation) as in [34, 51]. This assumption rigorously defines the phase variation as deviation of γ from an identity. Note that Γ is the set of cumulative distribution functions of absolutely continuous random variables on $[0,1]$.

2.2.2 Simplifying the geometry of Γ

Working with γ directly is not desirable due to the non-linearity of Γ. Let $\gamma_1, \gamma_2 \in \Gamma$ and a scaler $C \in R$. Then in general $\gamma_1 + \gamma_2 \notin \Gamma$ and $C\gamma \notin \Gamma$. To address this problem, we adopt the geometric approach laid out [51, 36] by transforming Γ to a linear space so that the standard statistical operations such as cross-sectional mean and covariance can be used.

2.2.2.1 Mapping to unit sphere The level of difficulty in dealing with γ is eased with transforming γ by the mapping,

$$\Theta : \gamma \mapsto \frac{\gamma'}{\sqrt{|\ \gamma'\ |}} = \sqrt{\gamma'}, \tag{2.2}$$

where $|\ \bullet\ |$ stands for an absolute value. A significant benefit of taking the representation in (2.2) is that the complicated structure of Γ is simplified to a much simpler structure. Since $\gamma(0) = 0$, and $\gamma(1) = 1$,

$$\left\| \sqrt{\gamma'} \right\|_2^2 = \int_0^1 \left[\sqrt{\gamma'(t)} \right]^2 dt = \int_0^1 \gamma'(t) dt = \gamma(1) - \gamma(0) = 1 - 0 = 1,$$

where $\| \bullet \|_2$ is a usual L_2 norm. This implies that, for any $\gamma \in \Gamma$, $\sqrt{\gamma'}$ lies in a unit sphere in $L_2[0, 1]$,

$$\sqrt{\gamma'} \in S = \{h(t) \in L_2[0, 1] \mid \|h\|_2 = 1\}.$$

Specifically, the image of $\Theta(\gamma)$ is the positive hyper-orthant of S.

2.2.2.2 Mapping to tangent space While the geometry of S is indeed simpler than Γ, it still is a non-linear manifold. However, the unit sphere is now easier to approximate by a linear subspace in $L_2[0, 1]$. A tangent space of S at a point $\mu \in S$, denoted by $T_\mu S$, is the collection of functions in $L_2[0, 1]$ orthogonal to μ,

$$T_\mu S = \{h(t) \in L_2[0, 1] \mid \langle h, \mu \rangle = 0\},$$

where $\langle \bullet, \bullet \rangle$ is an usual inner product in $L_2[0, 1]$. Functions in S will be approximated by functions in $T_\mu S$. Figure 2.(b) schematically illustrates $T_\mu S$ and the approximation of S-valued functions by functions in $T_\mu S$. Imagine the hyperplane T in $L_2[0, 1]$ tangent to S at μ. The tangent space $T_\mu S$ is the parallel translation of the hyperplane T such that the tangent point μ meets the origin, being a subspace in $L_2[0, 1]$. Points on the tangent space $T_\mu S$, when shifted back to T, are used to

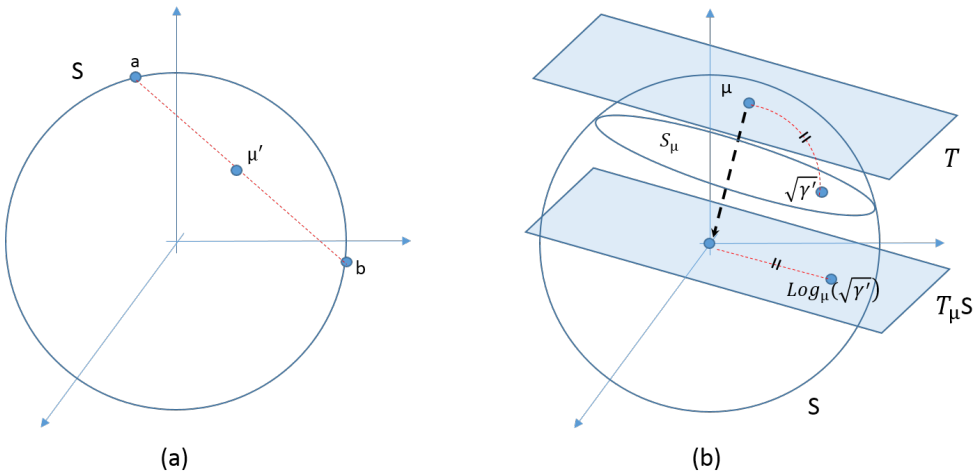

Figure 2. (a) The pointwise mean μ' of two functions $a, b \in S$ does not lie on S. (b) Mapping of $\sqrt{\gamma'} \in S$ to a tangent space $T_\mu S$ by the log map and its inverse mapping of $\exp_\mu(\mathrm{Log}_\mu(\sqrt{\gamma'})) = \sqrt{\gamma'}$ by the exponential map.

locally approximate functions in a subset $S_\mu \subset S$ near μ, described in 2.(b). Specifically, we use the log map $\mathrm{Log}_\mu : S_\mu \mapsto T_\mu S$,

$$\mathrm{Log}_\mu : \sqrt{\gamma'} \mapsto \frac{d_g(\sqrt{\gamma'}, \mu)}{\sin(d_g(\sqrt{\gamma'}, \mu))}(\sqrt{\gamma'} - \cos(d_g(\sqrt{\gamma'}, \mu))\mu), \tag{2.3}$$

where $d_g(\sqrt{\gamma'}, \mu) = \arccos(\langle\sqrt{\gamma'}, \mu\rangle)$. The $d_g(\sqrt{\gamma'}, \mu)$ measures the distance between $\sqrt{\gamma'}$ and μ by the length of the shortest arc on S that joins $\sqrt{\gamma'}$ and μ. Compared to other possible mappings from S to $T_\mu S$, the log map has an appealing feature: preservation of the (geodesic) distance between μ and $\mathrm{Log}_\mu(\sqrt{\gamma'})$ and the direction in which $\mathrm{Log}_\mu(\sqrt{\gamma'})$ shoots from μ. The log map provides a good approximation especially for a manifold of a relatively simple structure like a unit sphere S.

A sensible choice is a point on S corresponding to the mean of γ in the mapping (2.2). Since the mean of γ is assumed to be an identity function γ_{id}, a tangent space $T_\mu S$ is placed at a constant function $\mu \equiv 1$. The image of Log_μ is denoted by $B_\pi = \{\mathrm{Log}_\mu(\sqrt{\gamma'}) \in T_\mu S \mid \|\mathrm{Log}_\mu(\sqrt{\gamma'})\|_2 < \pi\}$.

2.2.3 Construction of f

The inverse of the log map defined on B_π is the exponential map, defined by,

$$\mathrm{Exp}_\mu : \mathrm{Log}_\mu(\sqrt{\gamma'}) \mapsto \mathrm{Log}_\mu(\sqrt{\gamma'})\frac{\sin(\|\mathrm{Log}_\mu(\sqrt{\gamma'})\|_2)}{\|\mathrm{Log}_\mu(\sqrt{\gamma'})\|_2} + \cos(\|\mathrm{Log}_\mu(\sqrt{\gamma'})\|_2)\mu. \tag{2.4}$$

The exponential map, acting as the inverse of the log map as illustrated in Figure 2.(b), is useful to construct $\gamma(t)$ from a function in $T_\mu S$. Let $x = \text{Log}_\mu(\sqrt{\gamma'}) \in B_\pi \in T_\mu S$. Since $\gamma(0)$=0, γ can be uniquely constructed from x using the following,

$$\int_0^t \text{Exp}_\mu^2(x)(s)ds = \int_0^t \left[\sqrt{\gamma\prime(s)}\right]^2 ds = \int_0^t \gamma\prime(s)ds = \gamma(t) - \gamma(0) = \gamma(t), \ t \in [0,1].$$

Then, Equation (1) can be rewritten as,

$$f(t) = y(\Theta^{-1} \circ Exp_\mu(x)) = y\left(\int_0^t \text{Exp}_\mu^2(x)(s)ds\right), \ t \in [0,1], \tag{2.5}$$

which implies that the random function f is constructed from the two random functions $y \in L_2[0,1]$ and $x \in T_\mu S$.

2.2.4 Models of FCPCA and FCCCA

FCPCA and FCCCA are defined with the two random functions $y \in L_2[0,1]$ and $x \in T_\mu S$. Note that both $L_2[0,1]$ and $T_\mu S$ are linear spaces. We assume $E(x) = 0$ so that $E(x)$ is identified with $E(\gamma) = \gamma_{id}$ when mapped back into a warping function by $\Theta^{-1} \circ \text{Exp}_\mu(E(x))$ (see (2.2) and (2.4) for Θ^{-1} and Exp_μ).

2.2.4.1 FCPCA model To capture the joint variability between y and x, we define a random function g^C on the extended domain $[0,2]$ by,

$$g^C(t) = \begin{cases} y(t) & t \in [0,1), \\ Cx(t-1) & t \in [1,2], \end{cases} \tag{2.6}$$

where $C > 0$. The exclusion of the end point $\{1\}$ of the domain $[0,1]$ of y in the construction of g^C does not lose any information of y since a point is measure zero and, if necessary, $y(1)$ can always be recovered using $\lim_{t\to 1} y(t)$ (since y is a smooth function). Note that $g^C \in L_2[0,2]$ since y and Cx are defined on disjoint subintervals of [0,1], $y \in L_2[0,1]$ and $x \in T_\mu S \subset L_2[0,1]$. The scaling parameter C is introduced to adjust scaling imbalance between y and x due to a unit change in y.

The eigendecomposition of the covariance function Σ_{g^C} of g^C gives,

$$\Sigma_{g^C}(s,t) = \sum_{i=0}^{\infty} \lambda_i^C \xi_i^C(s)\xi_i^C(t), s, t \in [0,2],$$

where the superscript C is used to make clear the dependency of λ_i's and ξ_i's on C, λ_i^C's denote eigenvalues of Σ_{g^C} with $\lambda_1^C \geq \lambda_2^C \geq \cdots \geq 0$, and ξ_i^C's are their corresponding eigenfunctions with $\|\xi_i^C\|_2 = 1$ and $\langle \xi_i^C, \xi_j^C \rangle = 0$ for $i \neq j$. Then by the Karhunen-Loéve decomposition,

$$g^C(t) = \mu(t) + \sum_{i=1}^{\infty} z_i^C \xi_i^C(t), t \in [0, 2], \tag{2.7}$$

where $E(z_i^C) = 0, E((z_i^C)^2) = \lambda_i^C$, $E(z_i^C z_j^C) = 0$ for $i \neq j$. Note that $\mu = E(g^C)$ does not depend on C since y is irrelevant of C and $E(x) = 0$. Equation (2.7) gives the representations of y and x,

$$\begin{aligned} y^C(t) &= \mu(t) + \sum_{i=1}^{\infty} z_i^C \xi_i^C(t), t \in [0, 1) \\ x^C(t) &= \sum_{i=1}^{\infty} \frac{z_i^C}{C} \xi_i^C(t+1), t \in [0, 1] \end{aligned} \tag{2.8}$$

In (2.8), the associated variations between y and x are paired in the eigenfunctions ξ_i^C.

The role of the scaling parameter $C \in R$ in (2.6) becomes clear from (2.7). As opposed to the unit-less x, values of y depend on the unit in which measurements are made. The scaling parameter C can depend on the change of the unit of Y. Since scaling y up/down by C is equivalent to scaling x down/up by C, the scaling parameter is introduced to the x part of g^C to keep the original unit of y. The eigenfunctions $\{\xi_i^C\}_{i=1}^{\infty}$ and their eigenvalues $\{\lambda_i^C\}_{i=1}^{\infty}$ may vary for different choices of C; for a small C, the first a few eigenfunctions ξ_i^C are found to explain more of the variation of the y part of the random function g^C and, for a large C, the leading eigenfuncions reflect more of the variation of the x part. In other word, there are infinite choices for a set of $\{\xi_i^C\}_{i=1}^{\infty}$ depending on C, leading to an identifiability issue. To our aim of succinctly representing the combined variation of y and x in the original function space, we define C to be dependent on the original random function f in the following.

Let m be a positive integer. From (2.5) and (2.8), we define $A_m^C(f)$ as the approximation of f by the first m eigenfunctions,

$$A_m^C(f)(t) = y_m^C \left(\int_0^t \mathrm{Exp}_\mu^2(x_m^C(s)) ds \right), t \in [0, 1), \tag{2.9}$$

where

$$\begin{aligned} y_m^C(f)(t) &= \mu(t) + \sum_{i=1}^{m} z_i^C \xi_i^C(t), t \in [0, 1), \\ x_m^C(f)(t) &= \sum_{i=1}^{m} \frac{z_i^C}{C} \xi_i^C(t+1), t \in [0, 1]. \end{aligned} \tag{2.10}$$

The scaling parameter C is chose so that, for m' of one's choice, the construction of f by the use of the first m' eigenfunctions $\{\xi_i^C\}_{i=1}^{m'}$ best approximates f, i.e.,

$$\begin{aligned} C &= \text{argmin}_{C\in R} E(d^2(A_{m'}^C(f), A_\infty^C(f))) \\ &= \text{argmin}_{C\in R} E(d^2(A_{m'}^C(f), f)), \end{aligned} \tag{2.11}$$

where d is the usual distance function on $L_2[0,1]$. There are other choices for d such as L_1 and earth mover's distances [47]. We choose L_2 distance for fast computations and mathematical convenience.

With C determined, the original random function f is constructed by plugging (2.8) into (2.5). FCPCA reveals the ith mode of variation of f by setting $z_i^C = \pm c\sqrt{\lambda_i^C}$ for $c \in R$ and $z_j^C = 0$ for all $j \neq i$ in (2.8) and constructing a function using (2.5), where the proportion of the total variability explained by the ith mode of variation is calculated by $\lambda_i^C / \sum_{j=1}^{\infty} \lambda_i^C$.

2.2.4.2 FCCCA model In FCCCA, we are interested in pairs of correlated variations between y and x. Denote by $\rho_P(\psi_y, \psi_x)$ the correlation coefficent between $\langle \psi_y, y\rangle$ and $\langle \psi_x, x\rangle$ as a function of $\psi_y \in L_2[0,1]$ and $\psi_x \in T_\mu S$,

$$\rho_P(\psi_y, \psi_x) = Cov(\langle \psi_y, y\rangle, \langle \psi_x, x\rangle), \tag{2.12}$$

subject to $Var(\langle \psi_y, y\rangle) = Var(\langle \psi_x, x\rangle) = 1$. We find a first pair of functions (ψ_{y1}, ψ_{x1}) that maximizes ρ and subsequently look for pairs of functions $\{(\psi_{yi}, \psi_{xi})\}_{i=2}^{\infty}$ that maximize ρ under the conditions:

$$\begin{aligned} Var(\langle \psi_{yi}, y\rangle) &= Var(\langle \psi_{xi}, x\rangle) = 1, \\ Cov(\langle \psi_{yi}, y\rangle, \langle \psi_{yj}, y\rangle) &= Cov(\langle \psi_{xi}, x\rangle, \langle \psi_{xj}, x\rangle) \\ &= Cov(\langle \psi_{yi}, y\rangle, \langle \psi_{xj}, x\rangle) \\ &= Cov(\langle \psi_{xi}, x\rangle, \langle \psi_{yj}, y\rangle) \\ &= 0, \text{ for } j = 1, 2, .., i-1 \text{ for each } i \end{aligned} \tag{2.13}$$

The ith component ψ_{yi} (or ψ_{xi}) is called the ith canonical weight function of y (respectively x) and the correlation coefficient ρ_i evaluated at the ith two weight functions is called the ith canonical correlation coefficient.

Let $\mu_y = E(y)$. The effects on the means μ_y and $\mu_x = 0$ of the ith canonical weight functions ψ_{yi} and ψ_{xi} respectively are,

$$\begin{aligned} P_y(t) &= \mu_y(t) + a\psi_{yi}, \ t \in [0,1], \\ P_x(t) &= b\psi_{xi}, \ t \in [0,1], \end{aligned} \tag{2.14}$$

where $a, b \in R$ are scalars of one's choice (reasonable choice is $a/b = \beta$, where β is a slope from a regression of $\langle \psi_x, x \rangle$ against $\langle \psi_y, y \rangle$). Then, the ith mode of variation by FCCCA is visualized by choosing various values of (a, b) in (2.14) and by plugging P_y and P_x in place of y and x respectively of (2.5).

2.3 ESTIMATION

The following parameters are needed to be estimated to carry out FCPCA and FCCCA: the mean function μ, the pairs of eigenvalue and eigenfunction (λ_i^C, ξ_i^C) and the scaling parameter C for FCPCA and the triples of canonical correlation coefficient and canonical weight function $(\rho_i, \psi_{yi}, \psi_{xi})$ for FCCCA.

2.3.1 Obtaining functional realizations $\hat{f}_i$

Let f_i for $i = 1, 2, .., n$ be the ith realization of an underlying random function f from n independent experiments. The realizations f_i's do not manifest themselves in a direct way. They are measured and recorded at discrete time points and usually blurred with measurement errors. The available data are written as,

$$f_{ij} = f_i(t_{ij}) + \epsilon_{ij}, \ i = 1, 2, .., n, \ j = 1, 2, .., n_i, \ t_{ij} \in [0, 1], \tag{2.15}$$

where f_{ij} is a measurement of the ith realization f_i at the jth time point t_{ij} contaminated by a measurement error ϵ_{ij}. We assume ϵ_{ij} are i.i.d. with $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma_\epsilon^2$, for each (i, j). We also assume that the time points $\{t_{ij}\}_{j=1}^{n_i}$'s are placed dense enough so that smoothing the observations $\{f_{ij}\}_{j=1}^{n_i}$ with a suitable basis function system gives a close approximation of f_i. Denote the approximations to $\{f_i\}_{i=1}^n$ by $\{\hat{f}_i\}_{i=1}^n$.

2.3.2 Obtaining $\hat{y}_i$ and $\hat{x}_i$

By (2.1), each function $\hat{f}_i$ is decomposed into amplitude and phase parts (called aligned and warping functions respectively) by an application of function registration,

$$\hat{f}_i(t) = \hat{y}_i(\hat{\gamma}_i(t)), \ i = 1, 2, .., n, \ t \in [0, 1]. \tag{2.16}$$

We choose the alignment method using Fisher-Rao metric proposed in [51] for its good performance. As explained in Section 2.2.2, the warping functions are transformed into the tangent space $T_\mu S$ by $x_i = \text{Log}_\mu(\Theta(\hat{\gamma}_i))$, where $\mu \equiv 1$.

2.3.3 Estimation of FCPCA

2.3.3.1 Estimation of μ and (λ_i^C, ξ_i^C) Let the scaling parameter C be given. Evaluate the functions $\{\hat{y}_i\}_{i=1}^n$ and $\{\hat{x}_i\}_{i=1}^n$ on a fine and evenly spaced grid $\{t_1, t_2, .., t_k\}$ of $[0, 1]$ to get their vectorized versions $\{\hat{\mathbf{y}}_i\}_{i=1}^n$ and $\{\hat{\mathbf{x}}_i\}_{i=1}^n$. For each ith pair of vectors $(\hat{\mathbf{y}}_i, \hat{\mathbf{x}}_i)$, stack them up to form a single vector $\hat{\mathbf{g}}_i^C$ in such a way that (superscript C is attached for every quantity that depends on C),

$$\hat{\mathbf{g}}_i^C = \begin{bmatrix} \hat{\mathbf{y}}_i \\ C\hat{\mathbf{x}}_i \end{bmatrix}, \quad \hat{\mathbf{y}}_i = [y_i(t_1)\ y_i(t_2)\ \dots\ y_i(t_k)]^T,$$
$$\hat{\mathbf{x}}_i = [x_i(t_1)\ x_i(t_2)\ \dots\ x_i(t_k)]^T.$$

Let $\hat{\boldsymbol{\mu}} = \sum_{i=1}^n \hat{\mathbf{g}}_i^C/n$. The eigendecomposition of the sample covariance matrix $\widehat{\Sigma}_{g^C}$ obtained from $\{\hat{\mathbf{g}}_i^C\}_{i=1}^n$ provides $n-1$ pairs of eigenvalues and eigenvectors $(\hat{\lambda}_i^C, \hat{\boldsymbol{\xi}}_i^C)$'s,

$$\begin{aligned} \widehat{\Sigma}_{g^C} &= \sum_{i=1}^{n} [\hat{\mathbf{g}}_i^C - \hat{\boldsymbol{\mu}}][\hat{\mathbf{g}}_i - \hat{\boldsymbol{\mu}}]^T \\ &= \sum_{i=1}^{n-1} \hat{\lambda}_i^C \hat{\boldsymbol{\xi}}_i^C \left(\hat{\boldsymbol{\xi}}_i^C\right)^T, \end{aligned}$$

where $\hat{\lambda}_1^C \geq \hat{\lambda}_2^C \geq \cdots \geq \hat{\lambda}_{n-1}^C \geq 0$, $\|\hat{\boldsymbol{\xi}}_i^C\|_2 = 1$ and $\langle \hat{\boldsymbol{\xi}}_i^C, \hat{\boldsymbol{\xi}}_j^C \rangle = 0$ for $i \neq j$. Estimates of λ_i^C are $\hat{\lambda}_i^C$. Estimates $\hat{\mu}$ and $\hat{\xi}_i^C$ of μ and ξ_i^C are obtained by interpolating $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\xi}}_i^C$ and normalizing $\hat{\boldsymbol{\xi}}_i^C$.

2.3.3.2 Estimation of C The estimates $\hat{\mu}$, $\{(\hat{\lambda}_i^C, \hat{\xi}_i^C)\}_{i=1}^{n-1}$ and $\{\hat{\mathbf{g}}_i^C\}_{i=1}^n$ are dependent on the value of C. Our strategy in the estimation of C is to use an empirical minimizer of (2.11). For this, interpolate $\hat{\mathbf{g}}_i^C$'s to get their functional versions $\{\hat{g}_i^C\}_{i=1}^n$ and calculate the scores $a_{ij}^C = \langle \hat{g}_i^C, \hat{\xi}_j^C \rangle$ for $i = 1, 2, .., n$ and $j = 1, 2, .., n-1$. Viewing $\{\hat{f}_i\}_{i=1}^n$ as n realizations of f, the empirical version $\hat{A}_m^C(\hat{f}_i)$ of $A_m^C(f)$ in (2.9), which approximates $\hat{f}_i$, is defined by replacing y_m^C and x_m^C in (2.10) with $\hat{y}_m^C$ and $\hat{x}_m^C$, where

$$\hat{y}_m^C(\hat{f}_i)(t) = \hat{\mu}(t) + \sum_{j=1}^{m} a_{ij}^C \hat{\xi}_j^C(t),\ t \in [0, 1),$$

$$\hat{x}_m^C(\hat{f}_i)(t) = \sum_{j=1}^{m} \frac{a_{ij}^C}{C} \hat{\xi}_j^C(t+1),\ t \in [0,1].$$

For some integer m' and a suitable interval I_C,

$$\hat{C} = \underset{C \in I_C}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{\|\hat{A}_{m'}^C(\hat{f}_i) - \hat{f}_i\|_2^2}{n}, \tag{2.17}$$

which implies that the mean function $\hat{\mu}$ and the first m' eigenfunctions $\hat{\xi}_i^{\hat{C}}$ found at $C = \hat{C}$ reconstruct $\{\hat{f}_i\}_{i=1}^n$ the most faithfully.

The minimizer $\hat{C}$ exists for most situations of our simulations and real data analyses, not degenerating to 0 or ∞. Heuristically, this is because, for $C > \hat{C}$, $\{\hat{\xi}_i^C\}_{i=1}^{m'}$ more reflect the variation of x than y's so that the residuals from (2.17) start to increase due to the amplitude of $\hat{f}_i$'s not being recovered from the approximation $\hat{A}_{m'}^C(\hat{f}_i)$'s and, for $C < \hat{C}$, temporal mismatching between $\hat{f}_i$'s and $\hat{A}_{m'}^C(\hat{f}_i)$'s begin to raise the residuals.

2.3.4 Estimation of $(\rho_i, \psi_{yi}, \psi_{xi})$ of FCCCA

It is well know that Naïve maximization of the quantity ρ in (2.12) using the pairs of functions $\{(\hat{y}_i, \hat{x}_i)\}_{i=1}^n$ usually produces pairs of uninterpretable canonical weight functions whose canonical correlation coefficients are very close to one.

Following [29], we regularize canonical weight functions by introducing a roughness penalty term into the constraints of (2.12) and find estimates $\hat{\rho}_1, \hat{\psi}_{y,1}$ and $\hat{\psi}_{x,1}$, which maximize the following quantity,

$$\rho_P(\hat{\psi}_{y1}, \hat{\psi}_{x1}) = \operatorname{argmax}_{\psi_y, \psi_x \in L_2[0,1]} \widehat{Cov}(\langle \psi_y, \hat{y}_i \rangle, \langle \psi_x, \hat{x}_i \rangle) \tag{2.18}$$

subject to $\widehat{Var}(\langle \psi_y, \hat{y}_i \rangle) + \lambda \|D^2 \psi_y\|_2^2 = \widehat{Var}(\langle \psi_x, \hat{x}_i \rangle) + \lambda \|D^2 \psi_y\|_2^2 = 1$, where $\widehat{Cov}$ and $\widehat{Var}$ are sample covariance and variance, D^2 is a second order differential operator and λ is a smoothing parameter. The $\hat{\rho}_1$ is obtained by evaluating (2.18) at $(\hat{\psi}_{y1}, \hat{\psi}_{x1})$. The subsequent triples $\{\hat{\rho}_i, \hat{\psi}_{yi}, \hat{\psi}_{xi}\}_{i=2}^{n-1}$ are found as maximizers of (2.12) subject to the constraints (2.13) with a roughness penalty. Refer to [44, Ch 11.] for the estimation algorithm and generalized cross-validation (GCV) method for determining λ.

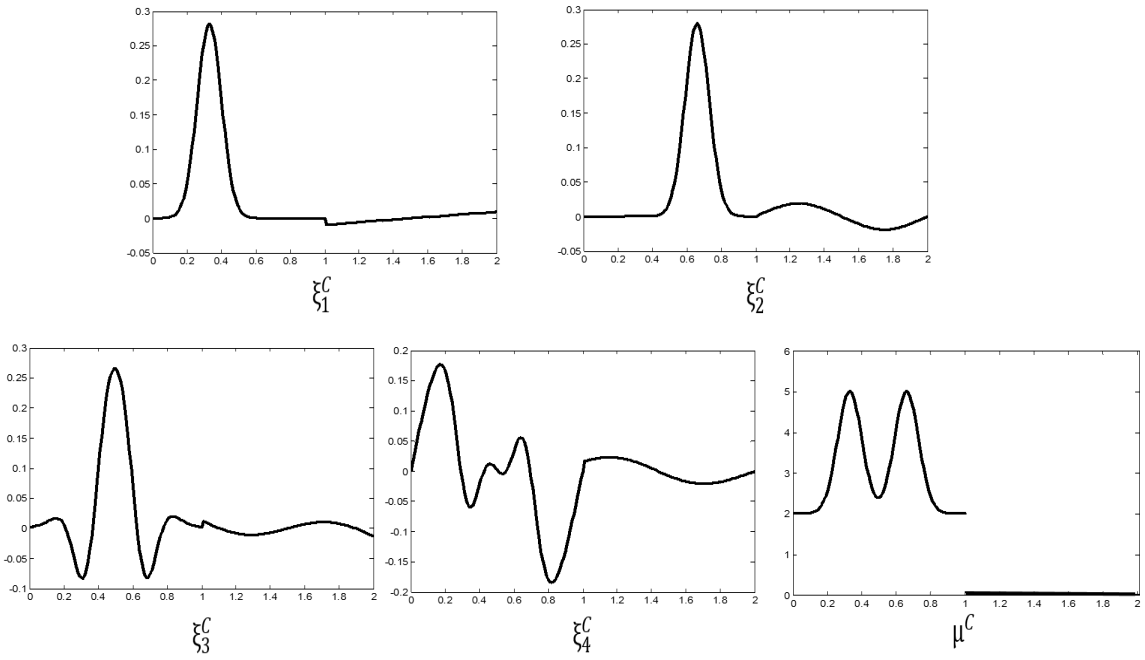

Figure 3. Parameter settings for $\xi_1^C, \xi_2^C, \xi_3^C, \xi_4^C$ and μ^C

.

2.4 SIMULATION STUDY

In this section, we empirically observe the consistency of the estimators of FCPCA and FCCCA. We have tried a range of parameter settings for each of FCPCA and FCCCA and the results are concordant across settings. Below we only choose to present representative cases.

2.4.1 Simulation for estimators in FCPCA

2.4.1.1 Simulation configurations We generate n independent realizations $\{g_i^C\}_{i=1}^n$ of g^C using a finite version of (2.7),

$$g_i^C(t) = \mu^C(t) + \sum_{j=1}^{4} z_{ij}\xi_j^C(t),\ t \in [0,2],$$

where the four eigenfunctions $\{\xi_i^C\}_{i=1}^4$ and the mean function μ^C are shown in Figure 3, and $\{(z_{i1}, z_{i2}, z_{i3}, z_{i4})\}_{i=1}^n$ are sampled from the multivariate normal distribution N_4,

$$\bar{\mathbf{z}}_i = \begin{bmatrix} z_{i1} & z_{i2} & z_{i3} & z_{i4} \end{bmatrix}^T,\ i = 1, 2, \ldots, n,$$

Parameter	Estimate	n=30		n=100	
		Mean	Std	Mean	Std
$\lambda_1^C = 3.5$	$\hat{\lambda}_1^{\hat{C}}$	4.12	0.26	3.81	0.21
$\lambda_2^C = 2.6$	$\hat{\lambda}_2^{\hat{C}}$	2.98	0.37	2.74	0.18
$C = 1$	$\hat{C}$	1.44	0.31	1.28	0.29

Parameter	Estimate	n=30		n=100	
		L_2 Diff Mean	L_2 Diff Std	L_2 Diff Mean	L_2 Diff Std
μ^C	$\hat{\mu}^{\hat{C}}$	2.89	1.27	2.15	0.85
ξ_1^C	$\hat{\xi}_1^{\hat{C}}$	0.49	0.24	0.38	0.36
ξ_1^C	$\hat{\xi}_2^{\hat{C}}$	0.71	0.41	0.34	0.50

Table 2. Results of simulations for consistency of estimators in FCPCA.

$$\bar{\mathbf{z}}_i \sim N_4\left(\underset{4\times 1}{\mathbf{0}},\ \mathbf{\Sigma_z} = \text{diag}(\ [3.5, 2.6, 0.3, 0.1]\ ) \right),$$

where, for a vector $\mathbf{v}$, diag($\mathbf{v}$) denotes a square diagonal matrix with the elements of $\mathbf{v}$ on its diagonal and 0 off of its diagonal. We found that it is almost intractable to track down the true value of a scailing parameter C analytically. The set of parameter eigenfunctions given in 3 is purposely selected so that the estimates $\hat{C}$ from 100 simulations with a sample size of 1000 is distributed around 1 with a narrow spread of 0.031 when the first 2 sample eigenfunctions are chosen to approximate the observed functions.

The functions $\{g_i^C\}_{i=1}^n$ are transformed to $\{f_i\}_{i=1}^n$ by (2.8) and (2.5). We observe f_i at each time point $t_j := (j-1)/101$, for $j = 1, 2, \ldots, 101$ with measurement error $\epsilon_{ij} \sim N(0, 0.1)$. As for a smoothing step for f_{ij}'s, the B-spline basis system of degree 4 with a roughness penalty on second derivative is used. Following [08], knots are placed at evaluation points $\{t_k\}_{k=1}^{101}$ and, following [07], the value of the smoothing parameter λ is determined by the generalized cross-validation method.

2.4.1.2 Simulation results For each sample size $n = 30, 100$, we perform 100 runs of simulations to collect 100 sets of the six estimates $\hat{\mu}^{\hat{C}}, (\hat{\lambda}_1^{\hat{C}}, \hat{\xi}_1^{\hat{C}}), (\hat{\lambda}_2^{\hat{C}}, \hat{\xi}_2^{\hat{C}})$ and $\hat{C}$. Tables 2 report means and standard deviations of scaler estimates and means and standard deviations of L_2 differences

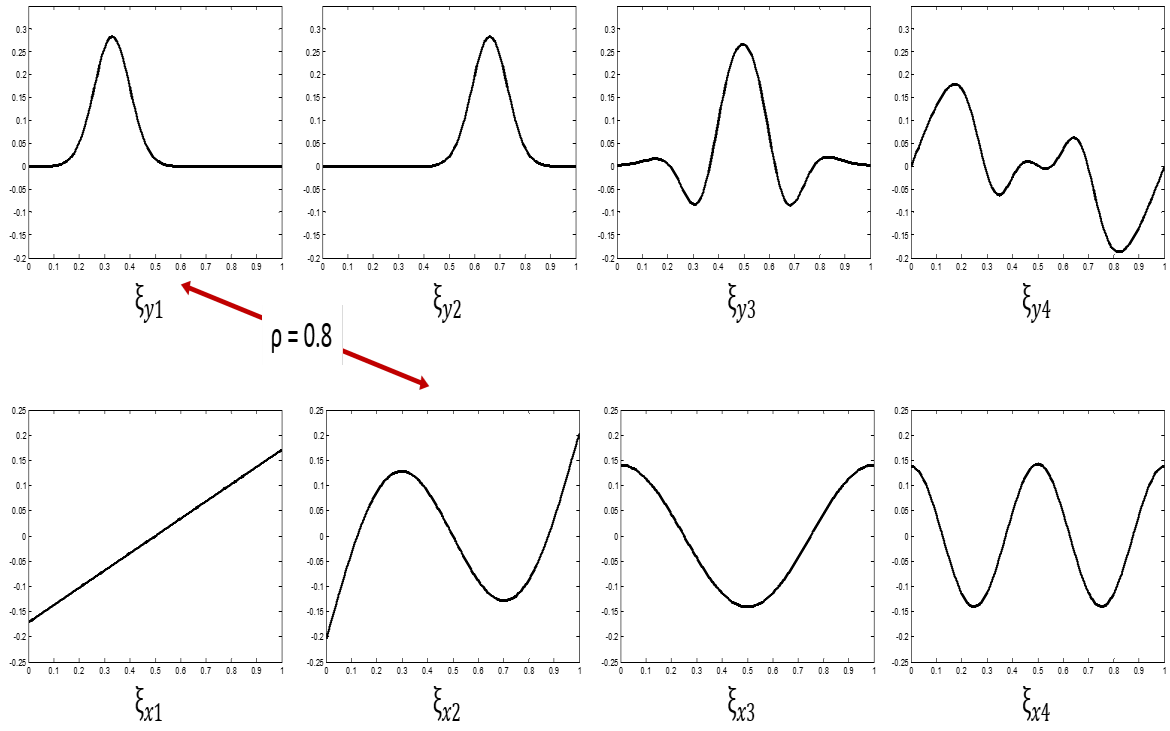

Figure 4. Parameter settings for $\{\xi_{yi}\}_{i=1}^4$ and $\{\xi_{xi}\}_{i=1}^4$, where ξ_{y1} and ξ_{x2} are chosen as canonical weight functions with $\rho = 0.8$.

between the functional estimates and their corresponding parameters.

We see that the estimates approach their population counterparts with narrower spreads as a sample size n increases. One noticeable observation is that the estimate μ^C shows larger discrepancy from the population mean and wider variability than other estimates. This may be because estimation of the mean involves all the other sample eigenfunctions in addition to the first two.

2.4.2 Simulation for estimators in FCCCA

2.4.2.1 Simulation configurations We create n pairs of independent realizations $\{(y_i, x_i)\}_{i=1}^n$ of y and x using,

$$y_i(t) = \mu_y(t) + \sum_{i=1}^{4} u_i \xi_{yi}(t), \ t \in [0,1],$$

$$x_i(t) = \sum_{j=1}^{4} v_j \xi_{xj}(t), \ t \in [0,1],$$

Parameter	Estimate	n=30		n=100	
		Mean	Std	Mean	Std
$\rho_1 = 0.8$	$\hat{\rho}_1$	0.67	0.21	0.76	0.18

Parameter	Estimate	n=30		n=100	
		L_2 Diff Mean	L_2 Diff Std	L_2 Diff Mean	L_2 Diff Std
ψ_{y1}	$\hat{\psi}_{y1}$	0.89	0.31	0.43	0.18
ψ_{x1}	$\hat{\psi}_{x1}$	0.72	0.28	0.55	0.15

Table 3. Results of simulations for consistency of estimators in FCCCA.

where the four eigenfunctions $\{\xi_{yi}\}_{i=1}^4$(respectively $\{\xi_{xi}\}_{i=1}^4$) chosen for y(or x) are depicted in Figure 4 and the mean functions μ_y for y is the first half of μ^C of Figure 4. The first eigenfunction ξ_{y1} of y and the second one ξ_{x2} of x are selected as a first pair of (normalized) canonical weight functions (ψ_{y1}, ψ_{y1}) with its canonical correlation coefficient of 0.8. The random variables $\{(\bar{\mathbf{u}}_i, \bar{\mathbf{v}}_i)\}_{i=1}^n$ are sampled from the multivariate normal distribution N_8,

$$\bar{\mathbf{u}}_i = \begin{bmatrix} u_{i1} & u_{i2} & u_{i3} & u_{i4} \end{bmatrix}^T, \quad \bar{\mathbf{v}}_i = \begin{bmatrix} v_{i1} & v_{i2} & v_{i3} & v_{i4} \end{bmatrix}^T,$$
$$\begin{bmatrix} \bar{\mathbf{u}}_i \\ \bar{\mathbf{v}}_i \end{bmatrix} \sim N_8 \left(\underset{8\times 1}{\mathbf{0}}, \begin{bmatrix} \boldsymbol{\Sigma}_u & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\Sigma}_{uv}^T & \boldsymbol{\Sigma}_v \end{bmatrix} \right), \quad i = 1, 2, \ldots, n,$$
$$\boldsymbol{\Sigma}_u = \text{diag}(\ [5, 3.5, 0.8, 0.7]\ ), \quad \boldsymbol{\Sigma}_v = \text{diag}(\ [0.01, 0.007, 0.0016, 0.0014]\ ),$$

where $\boldsymbol{\Sigma}_{uv}$ is the 4×4 matrix whose elements are all zero except for the first row and second column entry being 0.15 and, for a vector $\mathbf{v}$, diag($\mathbf{v}$) denotes a square diagonal matrix with the elements of $\mathbf{v}$ on its diagonal and 0 off of its diagonal. We have used Corollary 1 to construct the cross-covariance matrix $\boldsymbol{\Sigma}_{uv}$ from the given canonical correlation coefficient and canonical weight vectors. The pairs $\{(y_i, x_i)\}_{i=1}^n$ are transformed into $\{f_i\}_{i=1}^n$ using (2.5). Obtaining observations f_{ij}'s of the form of (2.15) and smoothing the observations follow the similar steps as explained in Section 2.4.1.

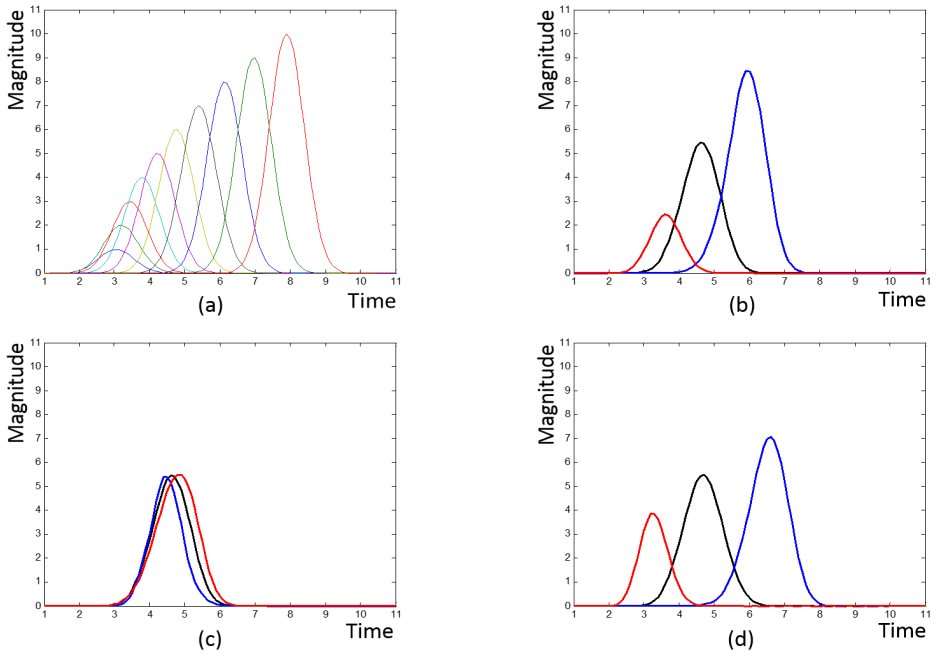

Figure 5. (a) 10 simulated functions. (b) Three functions describing a first mode of variation from FCPCA to the 10 curves. (c) Three functions describing a second mode of variation from FCPCA. (d) Three functions describing a combined effect of the most correlated directions from FCCCA.

2.4.2.2 Simulation results The same settings as those in Section 2.4.1 for sample sizes and the number of runs for each sample size are used to obtain 100 sets of the estimates $(\hat{\rho}_1, \hat{\psi}_{y1}, \hat{\psi}_{x1})$. Tables 3 report means and variances of scaler estimates and means and variances of L_2 differences between functional estimates and their corresponding parameters.

As in the simulation results of FCPCA case, the estimates approach their population counterparts with narrower variabilities as a sample size n increases. However, we are not quite sure of the effect of roughness penalty used in FCCCA on the consistency of the estimates.

2.5 DATA ANALYSIS

2.5.1 Synthetic data

2.5.1.1 Performance of FCPCA and FCCCA We start with a toy example where data structure is made clearly visible to see if FCPCA and FCCCA are capable of capturing it. The 10

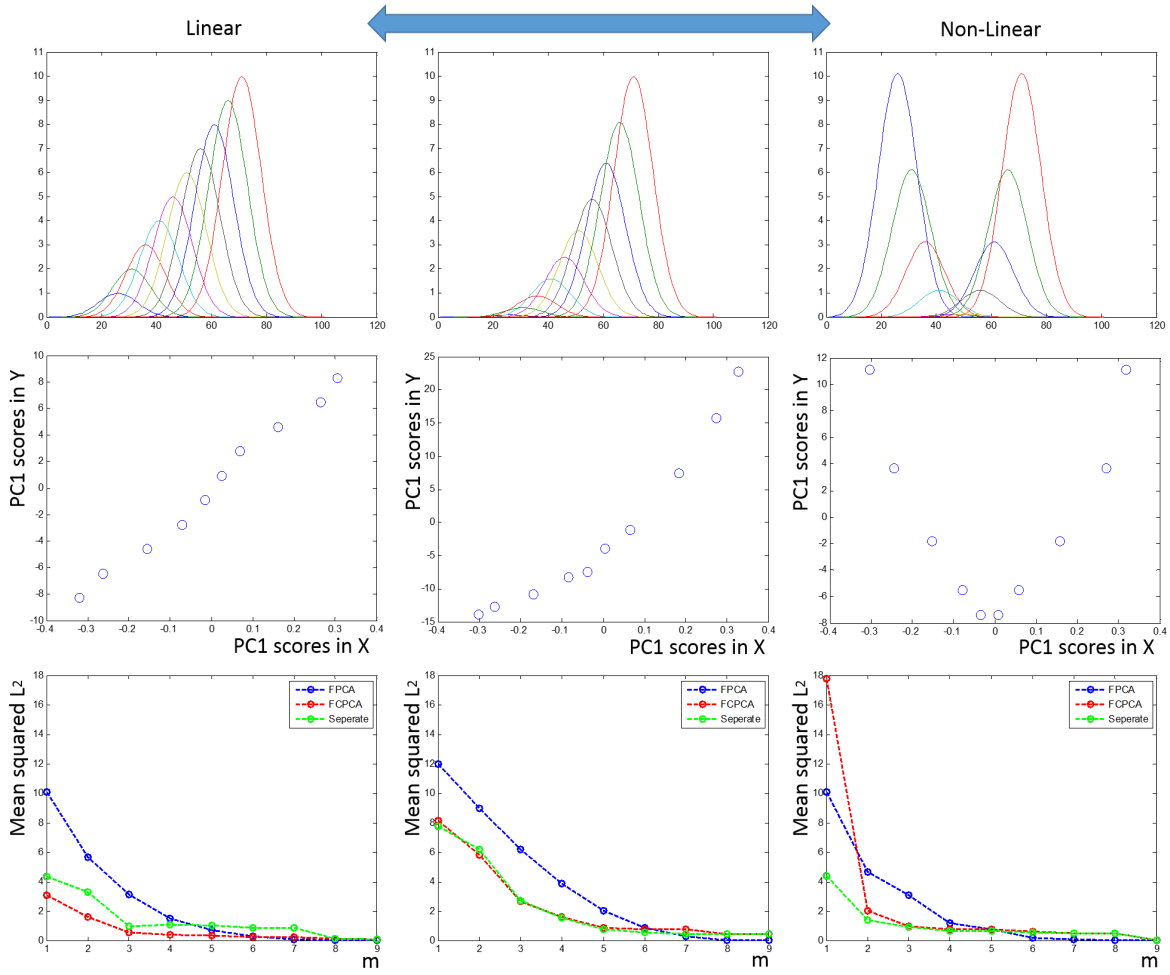

Figure 6. Comparison of reconstruction efficiency of FCPCA with that of FPCA and Separate methods over varying level of non-linearity between amplitude and phase variations.

functions shown in Figure 5.(a) display amplitude and phase variations such that, as a peak rises at a constant rate, the timing of the peak lags with a farther gap in between. Figure 5.(b) describes the first mode of variation obtained from FCPCA, which accounts for 95% of the total variability. Specifically, the black, blue and red curves depict the mean behavior and +/- 2 standard deviations along the direction of the largest variation from the mean. These three functions capture the general trend of the data but only imply the proportionate relationship between the height and timing of a peak; a peak's occurrence is delayed approximately proportionate to its height increment. The second mode of variation described in Figure 5.(c) explains just 4% of total variability yet delivers substantial information on the data structure. That is, to achieve the non-proportionate height

and timing relationship of a peak, the lags of a peak in Figure 5.(b) need to be adjusted by using the type of variation as shown in Figure 5.(c), which shifts the timing of a peak without touching its height.

The combined effect of the most correlated directions in amplitude and phase variations with a magnitude of $\hat{\rho} = 0.79$ resulting from FCCCA is shown in Figure 5.(d), whose interpretation is similar to that of Figure 5.(b).

2.5.1.2 Reconstruction efficiency of FCPCA under non-linear association We briefly discuss the sensitivity of the reconstruction efficiency of FCPCA to non-linearity between amplitude and phase variations. Recall from (2.9) and (2.10) that the observed functions can be reconstructed with a first m principal components (PCs) and their corresponding scores from FCPCA. Three sets of functions f's with varying degree of non-linearity between amplitude and phase variations are shown in the first row of Figure 6. The second row of Figure 6 demonstrates degree of non-linearity between amplitude and phase variations. Each graph in the row is a scatter plot of two sets of scores obtained from the functions f's above it, one for projections of aligned functions (y) onto the first PC of y and the other for projections of phase functions (x) onto the first PC of x. Reconstruction accuracy, measured by mean squared residuals (MSR) of reconstructed functions of FCPCA from f, is compared to those of other two methods across m. The first, FPCA, uses the first m PCs from FPCA to f's for reconstruction. The second, second method, called "Separate", applies FPCA to each of amplitude (y) and phase (x) functions and uses the first m PCs in each group to reconstruct f's using (2.5). In the third row of Figure 6, these three MSRs are compared for increasing values of m.

FCPCA best reconstructs f's when the relationship between amplitude and phase variations is almost linear as shown in the bottom left plot of Figure 6. However, it performs unfavorably compared to FPCA and Separate methods as the non-linearity intensifies. It is due in large part to the nature of FCPCA model that joins associated amplitude and phase variations by a linear method (PCA): recall from (2.8) that if the y part of an eigenfunction ξ_i^C is multiplied by a constant, then so is the x part by the constant, limiting responsiveness of FCPCA to non-linearity. Nonetheless, PCFCA effectively captures major modes of variation as one sees in the following real data examples where data structure becomes complicated.

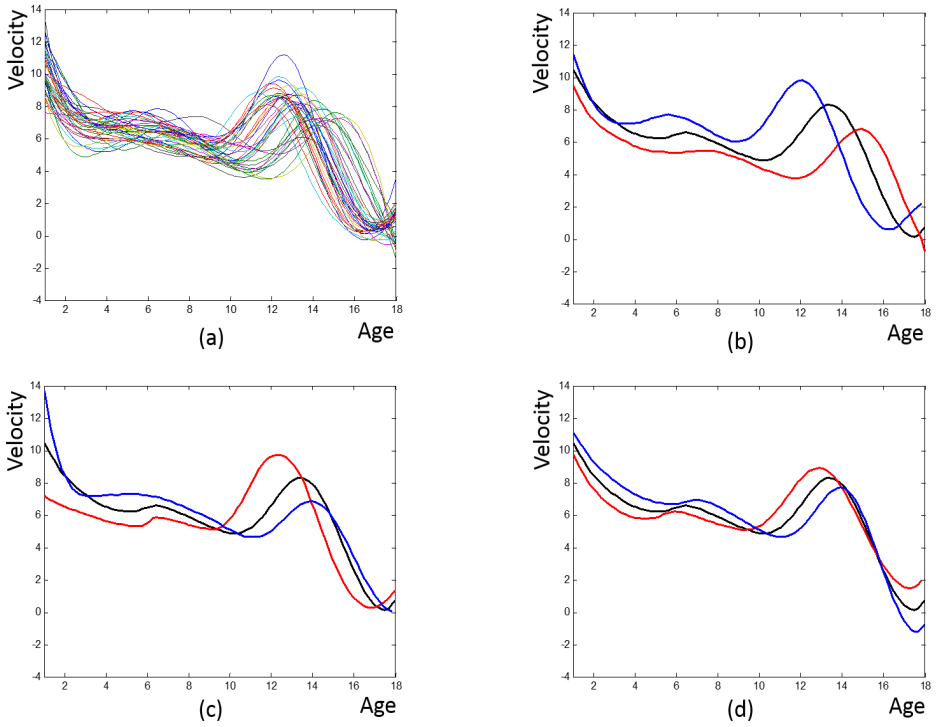

Figure 7. (a) 39 growth velocity curves. (b) Three functions describing a first mode of variation from FCPCA. (c) Three functions describing a second mode of variation from FCPCA. (d) Three functions describing a combined effect of the most correlated directions from FCCCA.

2.5.2 Berkeley Growth data

The Berkeley growth data [45] consist of the height measurements of 39 boys and 54 girls from age 1 to 18 at 31 age points, which are placed every three months until 2, every year until 8, and every six months from 8 to 18 years. We present here the analysis results of boys' data only as those of girls' are similar. To highlight periods of slower and faster growth, the growth velocity curves are found by differentiating the smoothed growth curves of the 39 boys and are included in Figure 7.(a).

It turns out that the first two components of FCPCA are interpretable. The growth type of the largest variation obtained from FCPCA, which accounts for 65% of the total variability, is described in Figure 7.(b), indicating that a boy who grows faster (slower) over the whole period tends to reach his growth development phases such as a pubertal growth spurt earlier (later) than others. On the other hand, the second mode of variation shown in Figure 7.(c), which accounts for 24% of the

total variability, characterizes the growth type that a boy's growth rate and development clock go faster (slower) before around 10 years, and then they turn slower (faster) afterwards.

The FCCCA reveals combined effect of the most correlated directions in amplitude and phase variations with a magnitude of $\hat{\rho} = 0.87$. This finding is different from the associations in the first two FCPCA components; a boy who grows faster (slower) before around 10 years and slower (faster) afterwards is likely to undergo his growth development phases earlier (later) than others over the entire period. The differences in patterns found by FCPCA and FCCCA should not be surprising since FCPCA finds directions in amplitude and phase variations that explain as much variability in the observed functions as possible in the squared L_2 distance sense while FCCCA aims at finding highly correlated directions after standardizing (therefore disregarding) variabilities along those directions.

2.5.3 Lip motion data

The data to be analyzed here is a part of lip motion data used in [42]. The data consist of measurements at 51 equally spaced points in the timeframe from 0 to 340 milliseconds of a vertical position of infrared emitting diodes (ireds) attached at the center of lower lip of a male subject while he speaks a syllable "bob" 20 times. The dynamics of lip motion is well captured by its acceleration. These second derivatives plotted in Figure 8.(a) show a common pattern. Lip movement is first accelerated negatively and then pass through a positive acceleration phase during which the descent of the lower lip is stopped. This lip opening phase is followed by a short period of near zero acceleration when pronunciation of a vowel "o" is at its full force, followed by another strong acceleration upward initiating lip closure and then finally completed by a negative acceleration episode as the lip returns to the closed position.

We used FCPCA and FCCCA to capture the major mode of variation in the data set. The first mode of variation accounting for 78% of the total variability and the combined effect of the most correlated directions with a strength of $\hat{\rho} = 0.83$ obtained from FCPCA and FCCCA are shown in Figure 8.(b) and (c) respectively. They both indicate a speech habit of the speaker. As he articulates the word louder (softer), he tends to speak faster (slower); note that, when someone speaks a word loud (soft), peaks and troughs in an acceleration curve of his lip movement become evident (flattened) as he opens his mouth wide (narrow).

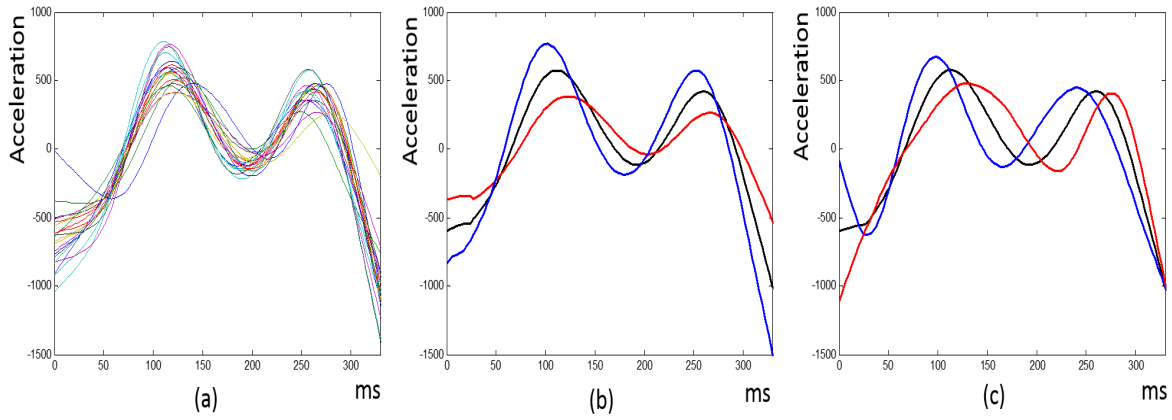

Figure 8. (a) 20 acceleration curves of lip movement. (b) Three functions describing a first mode of variation from FCPCA. (c) Three functions describing a combined effect of the most correlated directions from FCCCA.

2.6 CONCLUSION

This chapter presents a novel framework for exploring the internal structure of functional data varying with amplitude and phase variations. Naïve application of standard functional versions of statistical tools such as FPCA to this type of data sometimes produces unsatisfactory results. The commonly-employed framework of statistical analysis of aligned functions by the use of function registration disregards the phase variation. To overcome the disadvantages, FCPCA and FCCCA investigate main modes of variation and correlated directions of data in the underlying space, where the association structure between amplitude and phase variations can be addressed, and deliver their resulting variations in the original form of observed functions for clear visualization and interpretation purposes.

2.7 DISCUSSION

We mentioned three different metrics L_1, L_2 and earth mover's distance (EMD) [47] in comparing functions. The L_1 and L_2 are among the most commonly-used metrics while EMD is less know to statistics community. As a matter of fact, EMD becomes exactly the same as the Mallows distance when applied to probability distributions [30]. EMD was first introduced as empirical

ways to measure texture and color similarities in computer vision. Intuitively, if the two graphs are interpreted as two different ways of piling up a certain amount of dirt over the region D, EMD is the minimum cost of turning one pile into the other, where the cost is assumed to be amount of dirt that need to be moved times the distance it has to be moved. EMD is known to measure similarities between two graphs in a similar way as how human perception does.

To elaborate the advantages of EMD, take a simple example of three functions given in 9. All of those have a same mass. The function H2 is a horizontal shift of the function H1 by 1.5 whereas H3 is an almost flattened version of H1. To most of human eyes, H1 and H2 look much more similar than H1 and H3 do. However, L_1 and L_2 distances provide the opposite results that H1 and H3 are more similar that H1 and H2 are,

$$L_1(\text{H1, H2}) = 41.4351, \quad L_1(\text{H1, H3}) = 33.4351,$$
$$L_2(\text{H1, H2}) = 5.3403, \quad L_2(\text{H1, H3}) = 3.9477.$$

On the contrary, EMD measures the similarities among them as,

$$\text{EMD(H1, H2)} = 15.47, \ \text{EMD(H1, H2)} = 9.54.$$

The difference in their performances primarily lies in the fact that L_1 and L_2 metrics only consider the vertical gap at each point of the domain while EMD takes into account horizontal distances the masses need to travel from one point to the other for the two graphs to look similar. Although powerful, EMD is only applicable to graphs with same mass and positive values and its calculation is computationally expensive. We devised a new EMD applicable to positive graphs with different masses,

$$\text{EMD}^{\text{new}} = L_1(\text{H1, H2}) + \text{EMD}\left(\frac{\text{H1}}{\|\text{H1}\|_2}, \frac{\text{H2}}{\|\text{H2}\|_2}\right).$$

In simulations using positive graphs with different masses, EMD^{new} shows good performances in assigning small distances for seemingly similar graphs, when L_1 and L_2 do not. However, we need more extensive simulations and theoretical ground before claiming that the metric indeed performs favorably compared to popular L_1 and L_2 distances.

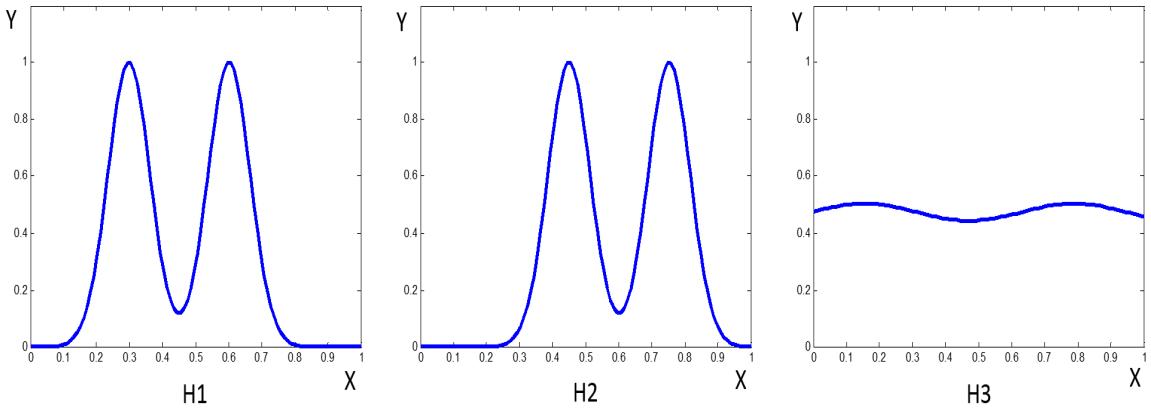

Figure 9. Three functions to be compared with using three different metrics L_1, L_2 and EMD

2.8 TECHNICAL DETAILS

Lemma 1 (Representation of cross-covariance matrix 1.)**.** *Suppose two random functions* $x, y \in L_2[0,1]$ *have respective covariance functions* Σ_x *and* Σ_y*. Let* $\{(\rho_i, \psi_{xi}, \psi_{yi})\}_{i=1}^{\infty}$ *be a sequence of triples of a canonical correlation coefficient and its corresponding canonical weight functions. Then the cross-covariance function* Σ_{xy} *between* x *and* y *is found as follows,*

$$\Sigma_{xy}(s,t) = \Sigma_x \Sigma_{xy}^* \Sigma_y(s,t), s, t \in [0,1],$$

where $\Sigma_{xy}^* = \sum_{i=1}^{\infty} \rho_i \psi_{xi} \psi_{yi}$ *and* $\Sigma\Sigma'(s,t) = \langle \Sigma(s,a), \Sigma'(a,t) \rangle = \int_{[0,1]} \Sigma(s,a)\Sigma'(a,t)da$ *for two functions* Σ *and* Σ'*.*

Proof. Covariance and cross-covariance operators Φ_x, Φ_y and Φ_{xy} as,

$$\Phi_x : L_2[0,1] \mapsto L_2[0,1], \Phi_x(w) = \int_{[0,1]} \Sigma_x(s,t)w(t)dt, w \in L_2[0,1],$$
$$\Phi_y : L_2[0,1] \mapsto L_2[0,1], \Phi_y(w) = \int_{[0,1]} \Sigma_y(s,t)w(t)dt, w \in L_2[0,1],$$
$$\Phi_{xy} : L_2[0,1] \mapsto L_2[0,1], \Phi_{xy}(w) = \int_{[0,1]} \Sigma_{xy}(s,t)w(t)dt, w \in L_2[0,1].$$

Subsequently maximizing ρ in (15) under the constraints in (16) to find the sequence of canonical triples $\{(\rho_i, \psi_{xi}, \psi_{yi})\}_{i=1}^{\infty}$ is equivalent to the eigenanalysis of the cross-correlation operator $R = \Phi_x^{-1/2} \Phi_{xy} \Phi_y^{-1/2}$. The covariance operators Φ_x and Φ_x are invertible under general conditions so

that canonical correlation coefficients and weight functions for x and y are well defined(Theorem 4.8, He et al., 2003). Let ρ_{ri}, ψ^1_{ri} and ψ^2_{ri} be the ith eigenvalue, its corresponding eigenfunction of R^*R and the i eigenfunction of RR^* respectively, where R^* denotes an adjoint operator of R. Then, the ith canonical correlation coefficient and weight functions are found as,

$$\rho_i = \sqrt{\rho_{ri}},\ \psi_{xi} = \Phi_x^{-1/2}(\psi^1_{ri}),\ \psi_{yi} = \Phi_y^{-1/2}(\psi^2_{ri}).$$

By Proposition 6.4 in [19], the kernel of R can be decomposed as,

$$\begin{aligned}\Sigma_x^{-1/2}\Sigma_{xy}\Sigma_y^{-1/2}(s,t) &= \sum_{i=1}^{\infty}\sqrt{\rho_{ri}}\psi^1_{ri}(s)\psi^2_{ri}(t)\\ &= \sum_{i=1}^{\infty}\rho_i\Phi_x^{1/2}(\psi_{xi})(s)\Phi_y^{1/2}(\psi_{yi})(t)\\ &= \Sigma_x^{1/2}\Sigma^*_{xy}\Sigma_y^{1/2}(s,t), s,t\in[0,1].\end{aligned}$$

Therefore we get,

$$\Sigma_{xy} = \Sigma_x\Sigma^*_{xy}\Sigma_y(s,t), s,t\in[0,1].$$

□

Corollary 1 (Representation of cross-covariance matrix 2.)**.** *Consider the eigendecompositions of the covariance functions Σ_x and Σ_y of x and y,*

$$\begin{aligned}\Sigma_x(s,t) &= \sum_{i=1}^{\infty}\lambda^{(d)}_{yi}\xi_{yi}(s)\xi_{yi}(t), s,t\in[0,1]\\ \Sigma_y(s,t) &= \sum_{i=1}^{\infty}\lambda^{(d)}_{xi}\xi_{xi}(s)\xi_{xi}(t), s,t\in[0,1].\end{aligned}$$

Then the cross-covariance function Σ_{xy} of x and y can be represented as follows,

$$\Sigma_{xy}(s,t) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\rho^*_{ijk}\psi^*_{ij}(s)\psi^*_{jk}(t),, s,t\in[0,1],$$

where $\psi^*_{jk} = \langle\psi_{xj},\xi_{xk}\rangle\xi_{xk}, \psi^*_{ij} = \langle\xi_{yi},\psi_{yj}\rangle\xi_{yi}$ and $\rho^*_{ijk} = \lambda^{(d)}_{yi}\rho_j\lambda^{(d)}_{xk}$.

Proof. Using Lemma 1,

$$\begin{aligned}\Sigma_{xy}(s,t) &= \Sigma_y\Sigma^*_{yx}\Sigma_x(s,t)\\ &= \langle\Sigma_y(s,a),\langle\Sigma^*_{yx}(a,b),\Sigma_x(b,t)\rangle\rangle\end{aligned}$$

$$= \left\langle \Sigma_y(s,a), \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \rho_j \lambda_{xk}^{(d)} \psi_{y,j}(a) \langle \psi_{xj}, \xi_{xk} \rangle \xi_{xk}(t) \right\rangle$$

$$= \left\langle \sum_{i=1}^{\infty} \lambda_{yi}^{(d)} \xi_{yi}(s)\xi_{yi}(a), \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \rho_j \lambda_{xk}^{(d)} \psi_{yj}(a)\psi_{jk}^{*} \right\rangle$$

where $\psi_{jk}^{*} = \langle \psi_{xj}, \xi_{xk} \rangle \xi_{xk}$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \lambda_{yi}^{(d)} \rho_j \lambda_{xk}^{(d)} \langle \xi_{yi}(a), \psi_{yj}(a) \rangle \xi_{yi}(s)\psi_{jk}^{*}$$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \rho_{ijk}^{*} \psi_{ij}^{*}(s)\psi_{jk}^{*}(t), , s,t \in [0,1]$$

where $\rho_{ijk}^{*} = \lambda_{yi}^{(d)} \rho_j \lambda_{xk}^{(d)}$ and $\psi_{ij}^{*} = \langle \xi_{yi}, \psi_{yj} \rangle \xi_{yi}$.

□

3.0 HDLSS ASYMPTOTIC ANALYSIS OF CCA

3.1 INTRODUCTION

Canonical correlation analysis (CCA) introduced in [20] is a standard statistical tool to explore the relationship between two sets of random variables. Consider d_X- and d_Y-dimensional random vectors $X^{(d_X)}$ and $Y^{(d_Y)}$,

$$\left(X^{(d_X)}\right)^T = \begin{bmatrix} X_1, & X_2, & \ldots, & X_{d_X} \end{bmatrix}, \left(Y^{(d_Y)}\right)^T = \begin{bmatrix} Y_1, & Y_2, & \ldots, & Y_{d_Y} \end{bmatrix}.$$

CCA first seeks a pair of d_X- and d_Y-dimensional weights vectors $\psi_{X1}^{(d_X)}$ and $\psi_{Y1}^{(d_Y)}$ such that two random variables, one being the linear combination of $X_1, X_2, \ldots, X_{d_X}$ weighted by the elements of $\psi_{X1}^{(d_X)}$ and the other being that of $Y_1, Y_2, \ldots, Y_{d_X}$ weighted by the elements of $\psi_{Y1}^{(d_Y)}$, have a maximal correlation,

$$(\psi_{X1}^{(d_X)}, \psi_{Y1}^{(d_Y)}) = \underset{\text{Var}(\langle \psi_{X1}^{(d_X)}, X^{(d_X)} \rangle) = \text{Var}(\langle \psi_{Y1}^{(d_Y)}, Y^{(d_Y)} \rangle) = 1}{\text{argmax}} \text{Cov}(\langle \psi_{X1}^{(d_X)}, X^{(d_X)} \rangle, \langle \psi_{Y1}^{(d_Y)}, Y^{(d_Y)} \rangle). \quad (3.1)$$

Requiring the norms of the weight vectors $\psi_{X1}^{(d_X)}$ and $\psi_{Y1}^{(d_Y)}$ to be one, the equation (3.1) can be written as an equivalent form of,

$$(\psi_{X1}^{(d_X)}, \psi_{Y1}^{(d_Y)}) = \underset{\|\psi_{X1}^{(d_X)}\|_2 = \|\psi_{Y1}^{(d_Y)}\|_2 = 1}{\text{argmax}} \frac{\text{Cov}(\langle \psi_{X1}^{(d_X)}, X^{(d_X)} \rangle, \langle \psi_{Y1}^{(d_Y)}, Y^{(d_Y)} \rangle)}{\sqrt{\text{Var}(\langle \psi_{X1}^{(d_X)}, X^{(d_X)} \rangle)} \sqrt{\text{Var}(\langle \psi_{Y1}^{(d_Y)}, Y^{(d_Y)} \rangle)}}. \quad (3.2)$$

For convenience, denote the objective function in the right hand side of (3.2) by $\rho_P(\psi^{(d_X)}, \psi^{(d_Y)})$,

$$\rho : R^{d_X} \times R^{d_Y} \mapsto R$$

$$\rho_P(\psi^{(d_X)}, \psi^{(d_Y)}) = \frac{\text{Cov}(\langle \psi^{(d_X)}, X^{(d_X)} \rangle, \langle \psi^{(d_Y)}, Y^{(d_Y)} \rangle)}{\sqrt{\text{Var}(\langle \psi^{(d_X)}, X^{(d_X)} \rangle)} \sqrt{\text{Var}(\langle \psi^{(d_Y)}, Y^{(d_Y)} \rangle)}}.$$

Subsequent weights vectors $\psi_{Xi}^{(d_X)}$ and $\psi_{Yi}^{(d_Y)}$, for $i = 1, 2, \ldots, \min(d_X, d_Y)$, are found by maximizing the objective function $\rho_P(\psi^{(d_X)}, \psi^{(d_Y)})$,

$$(\psi_{Xi}^{(d_X)}, \psi_{Yi}^{(d_Y)}) = \underset{\|\psi_{Xi}^{(d_X)}\|_2 = \|\psi_{Yi}^{(d_Y)}\|_2 = 1}{\text{argmax}} \rho_P(\psi_{Xi}^{(d_X)}, \psi_{Yi}^{(d_Y)}), \ i = 1, 2, \ldots, \min(d_X, d_Y),$$

under the constraint that,

$$\begin{aligned}\mathrm{Cov}(\langle\psi_{Xi}^{(d_X)}, X^{(d_X)}\rangle, \langle\psi_{Xj}^{(d_X)}, X^{(d_X)}\rangle) &= \mathrm{Cov}(\langle\psi_{Yi}^{(d_Y)}, Y^{(d_Y)}\rangle, \langle\psi_{Yj}^{(d_Y)}, Y^{(d_Y)}\rangle) \\ &= \mathrm{Cov}(\langle\psi_{Xi}^{(d_X)}, X^{(d_X)}\rangle, \langle\psi_{Yj}^{(d_Y)}, Y^{(d_Y)}\rangle) \\ &= \mathrm{Cov}(\langle\psi_{Yi}^{(d_Y)}, Y^{(d_Y)}\rangle, \langle\psi_{Xj}^{(d_X)}, X^{(d_X)}\rangle) \\ &= 0,\ i = 1, 2, \ldots, \min(d_X, d_Y),\ j = 1, 2, .., i-1 \text{ for each } i.\end{aligned}$$

The ith pair of weight vectors $\psi_{Xi}^{(d_X)}$ and $\psi_{Yi}^{(d_Y)}$ are usually called the ith pair of canonical weight vectors (or canonical loadings). The correlation ρ evaluated at the ith pair $\psi_{Xi}^{(d_X)}$ and $\psi_{Yi}^{(d_Y)}$, denoted by $\rho_i^{(d_X,d_Y)}$, is called the ith canonical correlation coefficient, that is, $\rho_i^{(d_X,d_Y)} = \rho_P(\psi_{Xi}^{(d_X)}, \psi_{Yi}^{(d_Y)})$.

In practice, we collect two sets of observations of d_X- and d_Y-dimensional random vectors $X^{(d_X)}$ and $Y^{(d_Y)}$ on a common set of samples in a $d_X \times n$ matrix $\mathbf{X}^{(d_X)}$ and a $d_Y \times n$ matrix $\mathbf{Y}^{(d_Y)}$, respectively. We row-center $\mathbf{X}^{(d_X)}$ and $\mathbf{X}^{(d_X)}$ and let $\hat{\mathbf{\Sigma}}_X^{(d_X)}, \hat{\mathbf{\Sigma}}_Y^{(d_Y)}$ and $\hat{\mathbf{\Sigma}}_{XY}^{(d_X,d_Y)}$ be a covariance matrix of $X^{(d_X)}$, a covariance matrix of $Y^{(d_Y)}$ and a cross-covariance matrix of $X^{(d_X)}$ and $Y^{(d_Y)}$,

$$\hat{\mathbf{\Sigma}}_X^{(d_X)} = \frac{1}{n}\mathbf{X}^{(d_X)}\left(\mathbf{X}^{(d_X)}\right)^T,\ \hat{\mathbf{\Sigma}}_Y^{(d_Y)} = \frac{1}{n}\mathbf{Y}^{(d_Y)}\left(\mathbf{Y}^{(d_Y)}\right)^T,\ \hat{\mathbf{\Sigma}}_{XY}^{(d_X,d_Y)} = \frac{1}{n}\mathbf{X}^{(d_X)}\left(\mathbf{Y}^{(d_Y)}\right)^T.$$

For the case where the sample size n is greater than d_X and d_Y, the estimation of sample canonical weight vectors $(\hat{\psi}_{Xi}^{(d_X)}, \hat{\psi}_{Yi}^{(d_Y)})$ and sample canonical correlation coefficients $\hat{\rho}_i^{(d_X,d_Y)}$ are done through singular value decomposition of the matrix $\hat{\mathbf{R}}^{(d_X,d_Y)}$,

$$\begin{aligned}\hat{\mathbf{R}}^{(d_X,d_Y)} &= \left(\hat{\mathbf{\Sigma}}_X^{(d_X)}\right)^{-\frac{1}{2}} \hat{\mathbf{\Sigma}}_{XY}^{(d_X,d_Y)} \left(\hat{\mathbf{\Sigma}}_Y^{(d_Y)}\right)^{-\frac{1}{2}}, \\ \mathrm{SVD}(\hat{\mathbf{R}}^{(d_X,d_Y)}) &= \sum_{i=1}^{\min(d_X,d_Y)} \hat{\lambda}_{Ri}^{(d_X,d_Y)} \hat{\eta}_{RXi}^{(d_X)} \left(\hat{\eta}_{RYi}^{(d_Y)}\right)^T,\end{aligned} \tag{3.3}$$

where $\hat{\lambda}_{Ri}^{(d)}$ is a sample singular value with $\hat{\lambda}_{R1}^{(d)} \geq \hat{\lambda}_{R2}^{(d)} \geq \cdots \geq \hat{\lambda}_{R\min(d_X,d_Y)}^{(d)} \geq 0$, and $(\hat{\eta}_{RXi}^{(d_X)}, \hat{\eta}_{RYi}^{(d_Y)})$ is a pair of left and right sample singular vectors corresponding to $\hat{\lambda}_{Ri}^{(d)}$. Then, the ith sample canonical correlation coefficient $\hat{\rho}_i^{(d)}$ is found to be,

$$\hat{\rho}_i^{(d_X,d_Y)} = \hat{\lambda}_{Ri}^{(d_X,d_Y)}.$$

The ith pair of canonical weight vectors $\hat{\psi}_{Xi}^{(d_X)}$ and $\hat{\psi}_{Yi}^{(d_Y)}$ are obtained by unscaling and normalizing the ith pair of sample singular vectors $\hat{\eta}_{RXi}^{(d_X)}$ and $\hat{\eta}_{RYi}^{(d_Y)}$,

$$\hat{\psi}_{Xi}^{(d_X)} = \frac{\left(\hat{\mathbf{\Sigma}}_X^{(d_X)}\right)^{-\frac{1}{2}} \hat{\eta}_{RXi}^{(d_X)}}{\left\|\left(\hat{\mathbf{\Sigma}}_X^{(d_X)}\right)^{-\frac{1}{2}} \hat{\eta}_{RXi}^{(d_X)}\right\|_2},\ \hat{\psi}_{Yi}^{(d_Y)} = \frac{\left(\hat{\mathbf{\Sigma}}_Y^{(d_Y)}\right)^{-\frac{1}{2}} \hat{\eta}_{RYi}^{(d_Y)}}{\left\|\left(\hat{\mathbf{\Sigma}}_Y^{(d_Y)}\right)^{-\frac{1}{2}} \hat{\eta}_{RYi}^{(d_Y)}\right\|_2}. \tag{3.4}$$

The projection of the data matrix $\mathbf{X}^{(d_X)}$ onto the ith sample canonical weight vector $\hat{\psi}_{Xi}^{(d_X)}$ gives the canonical scores (or canonical variables) of $\mathbf{X}^{(d_X)}$ with respect to $\hat{\psi}_{Xi}^{(d_X)}$ and similarly for $\mathbf{X}^{(d_X)}$. Although powerful, CCA has several disadvantages. first, use of CCA is practically restricted to the case of two sets of data even if there is an attempt to generalize it to more than two sets of data [54]. Second, CCA components are estimable only if the sample size n is greater than d_X and d_Y. It is well know that, when $n < \max(d_X, d_Y)$, one can construct an infinite number of sample canonical weight vector pairs with their correlation of one. Moreover, overfitting is often a problem even when $n > d_X$ and d_Y. Hence, CCA is often considered not reliable in high-dimensional data sets. We, however, will show that, even in the case where sample size n is less than d_X or d_Y, some sample canonical weight vectors is estimable and furthermore consistent under a certain condition.

As high-dimensional data are increasingly common these days, where a large number of variables are measured for each object, there is a strong need to investigate the behavior of estimates resulting from the application of standard statistical tools such as CCA to a high-dimensional case (that is, scalability of those tools). In studies in which dimension d is allowed to go to infinity, three scenarios are typically considered [48],

- Low Dimension High Sample Size (LDHSS): Both dimension d and sample size n go to infinity but n increases much faster than d, which can be summarized as $d/n \to 0$. These problems are similar to conventional asymptotics where $n \to \infty$ with n being fixed.
- High Dimension High Sample Size (HDHSS): In this case, sample size and dimension grow together in the sense that $d/n \to c$ for some constant c. The behavior of eigenvalues of a sample covariance matrix under this high-dimensional situation were studied in [02, 22, 40] primarily using random matrix theories.
- High Dimension Low Sample Size (HDLSS): In this setting, the sample size is fixed and the dimension grows in the sense that $d/n \to \infty$. An important finding in this high-dimensional setting was studied in [01]. They showed that the first eigenvector of the sample covariance matrix converges consistently to its population counterpart in the spiked model, where the leading eigenvalue is considerably larger than the remaining eigenvalues. An interesting geometric structure of HDLSS data were revealed in [16].

In this chapter, we are going to study the asymptotic behavior of the sample canonical weight vectors and canonical correlation coefficients of CCA under the HDLSS setting, where dimension d is allowed to grow with sample size n being fixed.

Literature in the HDLSS asymptotic study of CCA is very limited, while the behavior of PCA components under the similar high-dimensional condition is well-studied in [23, 24]. This might be in part because CCA is not as widely used as PCA, which is almost an indispensable tool for dimension reduction of high-dimensional data prevalent these days, and in part due to the complicated estimation steps involving an inverse operator as in (3.3), which makes the analysis not straightforward. A relevant work is first addressed in [28], where the asymptotic behavior of sample singular vectors and singular values are analyzed under a HDLSS setting. In [48], the similar study of CCA is elaborated on, but their proof should have considered the fact that an infinite sum of quantities converging to zero does not necessarily approach to zero. The HDLSS asymptotic behavior of CCA components in this chapter will be studied in relatively a simple population structure and serves as a groundwork for further analysis.

3.2 ASSUMPTIONS AND DEFINITIONS

Without loss of generality for the case where the dimensions of two random vectors $X^{(d_X)}$ and $Y^{(d_Y)}$ grow in a sense that $d_X/d_Y \to 1$, we set $d_X = d_Y$ and consider two random vectors $X^{(d)}$ and $Y^{(d)}$ of a same dimension with mean zero. We assume that covariance structure of $X^{(d)}$ and $Y^{(d)}$ follows a simple spiked model as in [01], where the leading eigenvalues of their covariance matrix is considerably larger than the rest. In specific, let $\mathbf{\Sigma}_X^{(d)}$ and $\mathbf{\Sigma}_Y^{(d)}$ be the covariance matrices of $X^{(d)}$ and $Y^{(d)}$. Then, a spiked model can be easily understood via eigendecomposition of $\mathbf{\Sigma}_X^{(d)}$ and $\mathbf{\Sigma}_Y^{(d)}$,

$$\mathbf{\Sigma}_X^{(d)} = \sum_{i=1}^{d} \lambda_{Xi}^{(d)} \xi_{Xi}^{(d)} \left(\xi_{Xi}^{(d)}\right)^T, \ \mathbf{\Sigma}_Y^{(d)} = \sum_{j=1}^{d} \lambda_{Yj}^{(d)} \xi_{Yj}^{(d)} \left(\xi_{Yj}^{(d)}\right)^T, \tag{3.5}$$

where $\lambda_{Xi}^{(d)}$ is an population eigenvalue (or population PC variance) with $\lambda_{X1}^{(d)} \geq \lambda_{X2}^{(d)} \geq \cdots \geq \lambda_{Xd}^{(d)} \geq 0$, $\xi_{Xi}^{(d)}$ is an population eigenvector (or population PC direction) with $\|\xi_{Xi}^{(d)}\|_2 = 1$ and $\langle \xi_{Xi}^{(d)}, \xi_{Xj}^{(d)} \rangle = 0$ for $i \neq j$ and similarly for $\lambda_{Yj}^{(d)}$ and $\xi_{Yj}^{(d)}$. Here, we set,

$$\begin{aligned} \lambda_{X1}^{(d)} &= \sigma_X^2 d^\alpha \text{ and } \lambda_{Xi}^{(d)} = \tau_X^2 \text{ for } i = 2, 3, \ldots, d, \\ \lambda_{Y1}^{(d)} &= \sigma_Y^2 d^\alpha \text{ and } \lambda_{Yj}^{(d)} = \tau_Y^2 \text{ for } j = 2, 3, \ldots, d, \end{aligned} \tag{3.6}$$

where one sees that the leading eigenvalues $\lambda_{X1}^{(d)}$ and $\lambda_{Y1}^{(d)}$ become dominating the rest as $d \to \infty$. We now set up the population canonical components. We assume that the two random vector is related by a pair of canonical weight vectors with its canonical correlation coefficient of ρ. The

population canonical weight vector $\psi_X^{(d)}$ in the $X^{(d)}$ part is a linear combination of two eigenvectors $\xi_{X1}^{(d)}$ and $\xi_{X2}^{(d)}$ without loss of generality ($\xi_{X2}^{(d)}$ can be replaced with $\xi_{Xi}^{(d)}$ for any i) and similarly for the other population canonical weight vector $\psi_Y^{(d)}$ in the $Y^{(d)}$ part,

$$\psi_X^{(d)} = \cos\theta_X \xi_{X1}^{(d)} + \sin\theta_X \xi_{X2}^{(d)},\ \psi_Y^{(d)} = \cos\theta_Y \xi_{Y1}^{(d)} + \sin\theta_Y \xi_{Y2}^{(d)}. \tag{3.7}$$

Note that the angle between $\psi_X^{(d)}$ and $\xi_{X1}^{(d)}$ is θ_X and that the angle between $\psi_Y^{(d)}$ and $\xi_{Y1}^{(d)}$ is θ_Y as $\langle \psi_X^{(d)}, \xi_{X1}^{(d)} \rangle = \cos\theta_X$ and $\langle \psi_Y^{(d)}, \xi_{Y1}^{(d)} \rangle = \cos\theta_Y$. At this point, we apply the change of basis to the spaces of $X^{(d)}$ and $Y^{(d)}$ so that the eigenvectors $\{\xi_{Xi}^{(d)}\}_{i=1}^d$ and $\{\xi_{Yi}^{(d)}\}_{j=1}^d$ are represented by the standard basis $\{e_i^{(d)}\}_{k=1}^d$. Then, the canonical weight vectors $(\psi_X^{(d)}, \psi_Y^{(d)})$ given in (3.7) is rewritten as,

$$\psi_X^{(d)} = \cos\theta_X e_1^{(d)} + \sin\theta_X e_2^{(d)},\ \psi_Y^{(d)} = \cos\theta_Y e_1^{(d)} + \sin\theta_Y e_2^{(d)},$$

and the covariance structures given in (3.5) and (3.6) are described as,

$$\underset{d\times d}{\mathbf{\Sigma}_X^{(d)}} = \operatorname{diag}(\sigma_X^2 d^\alpha,\ \tau_X^2,\ \tau_X^2,\ \ldots,\ \tau_X^2),\ \underset{d\times d}{\mathbf{\Sigma}_Y^{(d)}} = \operatorname{diag}(\sigma_Y^2 d^\alpha,\ \tau_Y^2,\ \tau_Y^2,\ \ldots,\ \tau_Y^2), \tag{3.8}$$

where diag($\bullet$) is a square matrix with entries of $\bullet$ in the main diagonal and 0 off of it. With these population covariance structures and canonical components, the multivariate version of the corallory 1 gives the cross-covariance structure of $X^{(d)}$ and $Y^{(d)}$ as follows,

$$\mathbf{\Sigma}_{XY}^{(d)} = \begin{bmatrix} \frac{\rho\sigma_X^2\sigma_Y^2 d^{2\alpha}\cos\theta_X\cos\theta_Y}{AB} & \frac{\rho\sigma_X^2 d^\alpha \tau_Y^2\cos\theta_X\sin\theta_Y}{AB} & \underset{1\times(d-2)}{\mathbf{0}} \\ \frac{\rho\tau_X^2\sigma_Y^2 d^\alpha\sin\theta_X\cos\theta_Y}{AB} & \frac{\rho\tau_X^2\tau_Y^2\sin\theta_X\sin\theta_Y}{AB} & \underset{1\times(d-2)}{\mathbf{0}} \\ \underset{(d-2)\times 1}{\mathbf{0}} & \underset{(d-2)\times 1}{\mathbf{0}} & \underset{(d-2)\times(d-2)}{\mathbf{0}} \end{bmatrix}, \tag{3.9}$$

where

$$A = \sqrt{\sigma_X^2 d^\alpha\cos^2\theta_X + \tau_X^2\sin^2\theta_X},\ B = \sqrt{\sigma_Y^2 d^\alpha\cos^2\theta_Y + \tau_Y^2\sin^2\theta_Y}.$$

Then the covariance and cross-covariance structure of $X^{(d)}$ and $Y^{(d)}$ is succinctly described by the covariance structure of the concatenated random vector $T^{(2d)}$,

$$T^{(2d)} = \begin{bmatrix} X^{(d)} \\ Y^{(d)} \end{bmatrix},\ \mathbf{\Sigma}_T^{(2d)} = \begin{bmatrix} \mathbf{\Sigma}_X^{(d)} & \mathbf{\Sigma}_{XY}^{(d)} \\ \left(\mathbf{\Sigma}_{XY}^{(d)}\right)^T & \mathbf{\Sigma}_Y^{(d)} \end{bmatrix}. \tag{3.10}$$

To make the analysis a bit easy, we are going to work with a different representation of $X^{(d)}$ and $Y^{(d)}$. Let $Z^{(d)}$ be the $2d$-dimensional standard normal random vector. Then, $T^{(2d)}$ can be expressed

as,

$$T^{(2d)} = \begin{bmatrix} X^{(d)} \\ Y^{(d)} \end{bmatrix} = \left(\mathbf{\Sigma}_T^{(2d)}\right)^{\frac{1}{2}} Z^{(2d)},\ Z^{(2d)} \sim N\left(\underset{2d\times 1}{0}, \underset{2d\times 2d}{\mathbf{I}}\right). \tag{3.11}$$

We state some definitions used in the estimation. Since the dimensionality d is much larger than the sample size n in the HDLSS setting, the estimation step (3.3) of canonical components is problematic as the sample covariance matrices $\hat{\mathbf{\Sigma}}_X^{(d)}$ and $\hat{\mathbf{\Sigma}}_Y^{(d)}$ are singular. There are two ways to handle this singularity situation. The first one is to add a minute perturbation of $\epsilon\mathbf{I}$ for a small $\epsilon > 0$ to $\hat{\mathbf{\Sigma}}_X^{(d)}$ and $\hat{\mathbf{\Sigma}}_Y^{(d)}$ and the second is to use a pseudoinverse such as Moore-Penrose pseudoinverse. We use the pseudoinverse obtained from the eigendecomposition of the sample covariance matrices,

$$\hat{\mathbf{\Sigma}}_X^{(d)} = \sum_{i=1}^{n} \hat{\lambda}_{Xi}^{(d)} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T,\ \hat{\mathbf{\Sigma}}_Y^{(d)} = \sum_{j=1}^{n} \hat{\lambda}_{Yj}^{(d)} \hat{\xi}_{Yj}^{(d)} \left(\hat{\xi}_{Yj}^{(d)}\right)^T, \tag{3.12}$$

where $\hat{\lambda}_{Xi}^{(d)}$ is an sample eigenvalue (or sample PC variance) with $\hat{\lambda}_{X1}^{(d)} \geq \hat{\lambda}_{X2}^{(d)} \geq \cdots \geq \hat{\lambda}_{Xd}^{(n)} \geq 0$, $\hat{\xi}_{Xi}^{(d)}$ is an sample eigenvector (or sample PC direction) with $\|\hat{\xi}_{Xi}^{(d)}\|_2 = 1$ and $\langle \hat{\xi}_{Xi}^{(d)}, \hat{\xi}_{Xj}^{(d)} \rangle = 0$ for $i \neq j$ and similarly for $\hat{\lambda}_{Yj}^{(d)}$ and $\hat{\xi}_{Yj}^{(d)}$. The pseudoinverse we employ is defined as,

$$\left(\hat{\mathbf{\Sigma}}_X^{(d)}\right)^{-1} = \sum_{i=1}^{n} \left(\hat{\lambda}_{Xi}^{(d)}\right)^{-1} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T,\ \left(\hat{\mathbf{\Sigma}}_Y^{(d)}\right)^{-1} = \sum_{j=1}^{d} \left(\hat{\lambda}_{Yj}^{(d)}\right)^{-1} \hat{\xi}_{Yj}^{(d)} \left(\hat{\xi}_{Yj}^{(d)}\right)^T. \tag{3.13}$$

Then, the sample canonical correlation coefficient $\hat{\rho}_i^{(d)}$ is found as an ith sample singular value from the SVD of the matrix $\hat{R}^{(d)}$ defined in (3.4). The sample canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and $\hat{\psi}_{Yi}^{(d)}$ corresponding to $\hat{\rho}_i^{(d)}$ are obtained from (3.4) using the pseudoinverses (3.13).

The success and failure of CCA can be described by the consistency of the sample canonical weight vectors $\hat{\psi}_X^{(d)}$ and $\hat{\psi}_Y^{(d)}$ with their population counterpart $\psi_X^{(d)}$ and $\psi_Y^{(d)}$ under the limiting operation of $d \to \infty$ and n fixed. Using the angle as a measure of consistency, we say that $\hat{\psi}_X^{(d)}$ (similarly $\hat{\psi}_Y^{(d)}$) is,

- Consistent with $\psi_X^{(d)}$ if angle$(\hat{\psi}_X^{(d)}, \psi_X^{(d)}) \to 0$ as $d \to \infty$,
- Inconsistent with $\psi_X^{(d)}$ if angle$(\hat{\psi}_X^{(d)}, \psi_X^{(d)}) \to a$, for $0 < a < \pi/2$, as $d \to \infty$,
- Strongly inonsistent with $\psi_X^{(d)}$ if angle$(\hat{\psi}_X^{(d)}, \psi_X^{(d)}) \to \pi/2$ as $d \to \infty$.

Strong inconsistency implies that the estimate $\hat{\psi}_X^{(d)}$ and $\hat{\psi}_Y^{(d)}$ become completely oblivious of its population structure and reduce to arbitrary quantities, as indicated in the fact that $pi/2$ is indeed a largest angle possible between two vectors.

3.3 MAIN THEOREM AND INTERPRETATION

3.3.1 Main theorem

Let $X^{(d)}$ and $Y^{(d)}$ be the d-dimensional random vectors from the multivariate Gaussian distributions with mean 0 and the simple spiked covariance matrices $\mathbf{\Sigma}_X^{(d)}$ and $\mathbf{\Sigma}_Y^{(d)}$ described in (3.5) and (3.6). With the population canonical correlation coefficient ρ for $0 \leq \rho \leq 1$, define the population canonical weight vectors $\psi_X^{(d)}$ and $\psi_Y^{(d)}$ as,

$$\psi_X^{(d)} = \cos\theta_X \xi_{X1}^{(d)} + \sin\theta_X \xi_{X2}^{(d)}, \ \psi_Y^{(d)} = \cos\theta_Y \xi_{Y1}^{(d)} + \sin\theta_Y \xi_{Y2}^{(d)}$$

so that the angle between $\psi_X^{(d)}$ and $\xi_{X1}^{(d)}$ is θ_X, and the angle between $\psi_Y^{(d)}$ and $\xi_{Y1}^{(d)}$ is θ_Y. Then, the cross-covariance matrix $\mathbf{\Sigma}_{XY}^{(d)}$ of $X^{(d)}$ and $Y^{(d)}$ is found as in (3.9). The two random variables $X^{(d)}$ and $Y^{(d)}$ can be written in a equivalent form,

$$\begin{bmatrix} X^{(d)} \\ Y^{(d)} \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_X^{(d)} & \mathbf{\Sigma}_{XY}^{(d)} \\ \left(\mathbf{\Sigma}_{XY}^{(d)}\right)^T & \mathbf{\Sigma}_Y^{(d)} \end{bmatrix} Z^{(2d)}, \tag{3.14}$$

where $Z^{(2d)}$ is a $2d$-dimensional standard normal random vector. The data matrix whose columns consist of n i.i.d. samples from the distribution 3.14 is written as,

$$\begin{bmatrix} \mathbf{X}^{(d)} \\ \mathbf{Y}^{(d)} \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_X^{(d)} & \mathbf{\Sigma}_{XY}^{(d)} \\ \left(\mathbf{\Sigma}_{XY}^{(d)}\right)^T & \mathbf{\Sigma}_Y^{(d)} \end{bmatrix} \mathbf{Z}^{(2d)}, \tag{3.15}$$

where the columns of $\mathbf{Z}^{(2d)}$ consist of n i.i.d. samples from $2d$-dimensional standard normal distribution. Denote by z_1 and z_2 the first and $(d+1)$th rows of $\mathbf{Z}^{(2d)}$ corresponding to the first rows of $\mathbf{X}^{(d)}$ and $\mathbf{Y}^{(d)}$ respectively. Then, as $d \to \infty$ with the sample size n being fixed, the limiting behaviors of the sample canonical correlation coefficient $\hat{\rho}_i^{(d)}$ and its corresponding sample canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and $\hat{\psi}_{Yi}^{(d)}$ obtained from the data 3.15 are as follows,

Theorem 1 (Main result of the HDLSS asymptotic analysis of CCA.)**.** *(i)* $\alpha > 1$

$$\begin{aligned} &\text{angle}\left(\hat{\psi}_{X1}^{(d)}, \psi_X^{(d)}\right) \xrightarrow[d\to\infty]{P} \theta_X, \ \text{angle}\left(\hat{\psi}_{Y1}^{(d_Y)}, \psi_Y^{(d)}\right) \xrightarrow[d\to\infty]{P} \theta_Y, \ \hat{\rho}_1^{(d)} \xrightarrow[d\to\infty]{P} \frac{\langle m_1, m_2\rangle}{\|m_1\|_2 \|m_2\|_2}, \\ &\text{angle}\left(\hat{\psi}_{Xi}^{(d)}, \psi_X^{(d)}\right) \xrightarrow[d\to\infty]{P} 0, \ \text{angle}\left(\hat{\psi}_{Yi}^{(d)}, \psi_Y^{(d)}\right) \xrightarrow[d\to\infty]{P} 0, \ \hat{\rho}_i^{(d)} \xrightarrow[d\to\infty]{P} 0, \ i = 2, 3, \ldots, n, \end{aligned}$$

where

$$m_1 = (\sqrt{C_1} A_1^2 + \sqrt{C_2} B_1^2) z_1 + (\sqrt{C_1} A_1 A_2 + \sqrt{C_2} B_1 B_2) z_2,$$

$$m_2 = (\sqrt{C_1}A_1A_2 + \sqrt{C_2}B_1B_2)z_1 + (\sqrt{C_1}A_1^2 + \sqrt{C_2}B_1^2)z_2,$$

where

$$\begin{aligned}
&z_1, z_2 \overset{i.i.d.}{\sim} N\left(\underset{n\times 1}{0}, \underset{n\times n}{\mathbf{I}}\right), \\
&C_1 = \frac{\sigma_X^2 + \sigma_Y^2 + \sqrt{\left(\sigma_X^2\right)^2 - 2\sigma_X^2\sigma_Y^2 + 4\sigma_X^2\sigma_Y^2\rho^2 + \left(\sigma_Y^2\right)^2}}{2}, \\
&C_2 = \frac{\sigma_X^2 + \sigma_Y^2 - \sqrt{\left(\sigma_X^2\right)^2 - 2\sigma_X^2\sigma_Y^2 + 4\sigma_X^2\sigma_Y^2\rho^2 + \left(\sigma_Y^2\right)^2}}{2}, \\
&A_1 = \frac{C_1 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\Big/\sqrt{\left(\frac{C_1 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\right)^2 + 1},\ A_2 = 1\Big/\sqrt{\left(\frac{C_1 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\right)^2 + 1}, \\
&B_1 = \frac{C_2 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\Big/\sqrt{\left(\frac{C_2 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\right)^2 + 1},\ B_2 = 1\Big/\sqrt{\left(\frac{C_2 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\right)^2 + 1}.
\end{aligned}$$

(ii) $\alpha < 1$

$$\text{angle}\left(\hat{\psi}_{Xi}^{(d)}, \psi_X^{(d)}\right) \xrightarrow[d\to\infty]{P} 0,\ \text{angle}\left(\hat{\psi}_{Yi}^{(d)}, \psi_Y^{(d)}\right) \xrightarrow[d\to\infty]{P} 0,\ \hat{\rho}_i^{(d)} \xrightarrow[d\to\infty]{P} 1,\ i = 1, 2, \ldots, n.$$

3.3.2 Interpretation

Theorem 1 implies that where $\hat{\psi}_{X1}^{(d)}$ and $\hat{\psi}_{Y1}^{(d)}$ converge to depend heavily on the size of the variance d^α of the population eigenvector $\xi_{X1}^{(d)}$ and $\xi_{Y1}^{(d)}$. That is, the estimates $\hat{\psi}_{X1}^{(d)}$ and $\hat{\psi}_{Y1}^{(d)}$ tend to converge to the eigenvectors $\xi_{X1}^{(d)}$ and $\xi_{Y1}^{(d)}$ when their eigenvalues $\sigma_X^2 d^\alpha$ and $\sigma_Y^2 d^\alpha$ become strong enough ($\alpha > 1$) as $d \to \infty$. Briefly, we summarize results. The sample canonical weight vector $\hat{\psi}_{X1}^{(d)}$ (similarly $\hat{\psi}_{Y1}^{(d)}$) is,

- Consistent with $\psi_X^{(d)}$ if $\alpha > 1$ and angle$(\psi_X^{(d)}, \xi_{X1}^{(d)}) = 0$ as $d \to \infty$,
- Inconsistent with $\psi_X^{(d)}$ if $\alpha > 1$ and angle$(\psi_X^{(d)}, \xi_{X1}^{(d)}) = \theta_X$, for $0 < \theta_X < \pi/2$, as $d \to \infty$,
- Strongly inconsistent with $\psi_X^{(d)}$ if $\alpha < 1$ or if $\alpha > 1$ and angle$(\psi_X^{(d)}, \xi_{X1}^{(d)}) = \pi/2$ as $d \to \infty$.

The asymptotic behavior of the sample canonical correlation coefficient $\hat{\rho}_1^{(d)}$ is not straightforward to imagine. Let's take a simple example where $\sigma_X^2 = 1, \sigma_X^2 = 1, \tau_X^2 = 1$ and $\tau_Y^2 = 1$ in the spiked covariance structure in (3.5) and (3.6). In this case, referring to Theorem 1, the sample canonical correlation coefficient $\hat{\rho}_1^{(d)}$ converges in probability to the following random quantity,

$$\hat{\rho}_1^{(d)} \xrightarrow[d\to\infty]{P} \frac{\langle m_1, m_2\rangle}{\|m_1\|_2\|m_2\|_2},$$

where

$$m_1 = \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right) z_1 + \left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right) z_2,$$
$$m_2 = \left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right) z_1 + \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right) z_2.$$

Note that z_1 and z_2 are samples from n-dimensional multivariate standard normal distribution. It can be easily verified that each element m_{1i} of m_1 (similarly for m_{2i} of m_2) follows a standard normal distribution,

$$m_{1,i} = \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right) z_{1i} + \left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right) z_{2i} \sim N(0,1),$$
$$m_{2,i} = \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right) z_{1i} + \left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right) z_{2i} \sim N(0,1),$$

which leads to,

$$\|m_1\|_2 \sim \sqrt{\chi^2_n},\ \|m_2\|_2 \sim \sqrt{\chi^2_n},$$

where χ^2_n denotes the chi-square distribution with degree of freedom of n. Since the numerator part $\langle m_1, m_2 \rangle$ is not a degenerate random quantity, one sees that $\hat{\rho}_1^{(d)}$ does not converge to a trivial random variable such as 1.

Now increase the sample size n to see which value the sample canonical correlation coefficient $\hat{\rho}_1^{(d)}$ converges to. By the law of large numbers and noting that the elements $m_{1,i}$ and $m_{2,i}$ are from i.i.d. standard normal distribution,

$$\frac{\|m_1\|_2^2}{n} = \sum_{i=1}^{n} \frac{m_{1i}^2}{n} \xrightarrow[n\to\infty]{P} 1,\ \frac{\|m_2\|_2^2}{n} = \sum_{j=1}^{n} \frac{m_{2j}^2}{n} \xrightarrow[n\to\infty]{P} 1.$$

Furthermore, noting that m_1 and m_2 are i.i.d. samples,

$$\begin{aligned}
\frac{\langle m_1, m_2 \rangle}{n} &= \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right)\left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right) \sum_{i=1}^{n} \frac{z_{1i}^2}{n} \\
&+ \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right)\left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right) \sum_{i=1}^{n} \frac{z_{2i}^2}{n} \\
&+ \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right)^2 \sum_{i=1}^{n} \frac{z_{1i} z_{2i}}{n} \\
&+ \left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right)^2 \sum_{i=1}^{n} \frac{z_{1i} z_{2i}}{n} \\
&\xrightarrow[n\to\infty]{P} 2\left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right)\left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right) = \rho,
\end{aligned}$$

which confirms the conventional large sample asymptotic property of the statistic $\hat{\rho}_1^{(d)}$,

$$\hat{\rho}_1^{(d)} \xrightarrow[d,n\to\infty]{P} \rho.$$

3.4 PROOF

We state here the theorem on the HDLSS asymptotic behavior of the sample eigenvalues and vectors (or sample PC variances and directions) included in [24] as frequently referred in this chapter.

Theorem 2 (HDLSS asymptotic result of engenvalues and engenvectors.)**.** *Under the Gaussian assumption and the one spike case of (3.5) and (3.6),*
(i) the limit of the first sample eigenvalue $\hat{\lambda}_{Xi}^{(d)}$ (similarly for $\hat{\lambda}_{Yi}^{(d)}$) described in (3.12) depends on α,

$$\frac{\hat{\lambda}_{X1}^{(d)}}{\max(d^\alpha, d)} \Longrightarrow \begin{cases} \sigma_X^2 \frac{\chi_n^2}{n}, & \alpha > 1, \\ \sigma_X^2 \frac{\chi_n^2}{n} + \frac{\tau_X^2}{n}, & \alpha = 1, \\ \frac{\tau_X^2}{n}, & \alpha < 1, \end{cases}$$

as $d \to \infty$, where $\Longrightarrow$ denotes the convergence in distribution, and χ_n^2 denotes the chi-square distribution with degree of freedom n. The rest of eigenvalues converge to the same quantities when scaled, that is, for any $\alpha \in [0, \infty), i = 2, 3, \ldots, n$,

$$\frac{\hat{\lambda}_{Xi}^{(d)}}{n} \to \frac{\tau_X^2}{n}, \text{ as } d \to \infty,$$

in probability.
(ii) The limit of the first eigenvectors $\hat{\xi}_{Xi}^{(d)}$ (similarly for $\hat{\xi}_{Yi}^{(d)}$) described in (3.12) depends on α,

$$\langle \hat{\xi}_{Xi}^{(d)}, \xi_{Xi}^{(d)} \rangle \Longrightarrow \begin{cases} 1, & \alpha > 1, \\ \left(1 + \frac{\tau_X^2}{\sigma_X^2 \chi_n^2}\right)^{-\frac{1}{2}}, & \alpha = 1, \\ 0, & \alpha < 1, \end{cases}$$

as $d \to \infty$. The rest of the eigenvectors are strongly inconsistent with their population counterpart, for any $\alpha \in [0, \infty), i = 2, 3, \ldots, n$,

$$\langle \hat{\xi}_{Xi}^{(d)}, \xi_{Xi}^{(d)} \rangle \to 0, \text{ as } \infty,$$

in probability.

3.4.1 Settings

Let $\mathbf{X}^{(d)}$ and $\mathbf{Y}^{(d)}$ be $d \times n$ matrices that collect d-dimensional vector observations measured on the common set of n samples. Following the representation (3.11) of the two random vectors $X^{(d)}$ and $Y^{(d)}$, we, using covariance and cross-covariance structures given in (3.8), (3.9) and (3.10), write $\mathbf{X}^{(d)}$ and $\mathbf{Y}^{(d)}$ as,

$$\begin{bmatrix} \mathbf{X}^{(d)} \\ \mathbf{Y}^{(d)} \end{bmatrix} = \left(\mathbf{\Sigma}_T^{(2d)}\right)^{\frac{1}{2}} \mathbf{Z}^{(2d)} = \begin{bmatrix} \mathbf{\Sigma}_X^{(d)} & \mathbf{\Sigma}_{XY}^{(d)} \\ \left(\mathbf{\Sigma}_{XY}^{(d)}\right)^T & \mathbf{\Sigma}_Y^{(d)} \end{bmatrix}^{\frac{1}{2}} \mathbf{Z}^{(2d)}, \tag{3.16}$$

where $\mathbf{Z}^{(2d)}$ is a $2d \times n$ matrix with columns of $\mathbf{Z}^{(2d)}$ being i.i.d. observations from an $2d$-dimensional standard normal distribution. We introduce notations for elements of the matrix $\mathbf{Z}^{(2d)}$ for a later use in the proof,

$$\mathbf{Z}^{(2d)} = \begin{bmatrix} \mathbf{Z}_X^{(d)} \\ \mathbf{Z}_Y^{(d)} \end{bmatrix} = \begin{bmatrix} z_{X11} & z_{X12} & \cdots & z_{X1n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{Xd1} & z_{Xd2} & \cdots & z_{Xdn} \\ z_{Y11} & z_{Y12} & \cdots & z_{Y1n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{Yd1} & z_{Yd2} & \cdots & z_{Ydn} \end{bmatrix} = \begin{bmatrix} z_{X1\bullet} \\ \vdots \\ z_{Xd\bullet} \\ z_{Y1\bullet} \\ \vdots \\ z_{Yd\bullet} \end{bmatrix}, \tag{3.17}$$

where $\mathbf{Z}_X^{(d)}$ denotes the upper half of $\mathbf{Z}^{(2d)}$ corresponding to $\mathbf{X}^{(d)}$ (similarly for $\mathbf{Z}_Y^{(d)}$), z_{Xij}, for $i = 1, 2, \ldots, d$ and $j = 1, 2, \ldots, n$, represents the ith element of the j observation in $\mathbf{Z}_X^{(d)}$ (similarly for z_{Yij}), and $z_{Xk\bullet}$ denotes the kth row of $\mathbf{Z}^{(2d)}$ (similarly for $z_{Yk\bullet}$. To investigate the asymptotic behavior of the sample covariance and cross-covariance of $\mathbf{X}^{(d)}$ and $\mathbf{Y}^{(d)}$, we want to expand the matrix in (3.16) to get an explicit expression elementwise. This can be down either by manual or using symbolic operations in Matlab. The results, however, is too long to be included in this page, so we are going to work with a big O, small o representation for the elements that have lengthy expressions. The covariance matrix $\mathbf{\Sigma}_T^{(2d)}$ takes the following eigendecompsition,

$$\mathbf{\Sigma}_T^{(2d)} = \sum_{i=1}^{2d} \lambda_{Ti}^{(2d)} \xi_{Ti}^{(2d)} \left(\xi_{Ti}^{(2d)}\right)^T, \tag{3.18}$$

where

$$\lambda_{T1}^{(2d)} = O_P(d^\alpha),\ \lambda_{T2}^{(2d)} = O_P(d^\alpha),\ \lambda_{T4}^{(2d)} = O_P(1),\ \lambda_{T4}^{(2d)} = O_P(1),$$
$$\lambda_{Ti}^{(2d)} = \tau_X^2,\ i = 5, 6, \ldots, d+2$$

$$
\begin{aligned}
\lambda_{Tj}^{(2d)} &= \tau_Y^2,\ j = d+3, d+4, \ldots, 2d,\\
\xi_{T1}^{(2d)} &= \begin{bmatrix} O_P(1) & O_P(\frac{1}{\sqrt{d^\alpha}}) & O_P(1) & \ldots & O_P(1) & O_P(1) & O_P(\frac{1}{\sqrt{d^\alpha}}) & O_P(1) & \ldots & O_P(1) \end{bmatrix}^T,\\
\xi_{T2}^{(2d)} &= \begin{bmatrix} O_P(1) & O_P(\frac{1}{\sqrt{d^\alpha}}) & O_P(1) & \ldots & O_P(1) & O_P(1) & O_P(\frac{1}{\sqrt{d^\alpha}}) & O_P(1) & \ldots & O_P(1) \end{bmatrix}^T,\\
\xi_{T3}^{(2d)} &= \begin{bmatrix} O_P(1) & O_P(1) & \ldots & O_P(1) & O_P(1) \end{bmatrix}^T,\\
\xi_{T4}^{(2d)} &= \begin{bmatrix} O_P(1) & O_P(1) & \ldots & O_P(1) & O_P(1) \end{bmatrix}^T,\\
\xi_{Ti}^{(2d)} &= e_i^{(2d)},\ i = 5, 6, \ldots, d+2,\\
\xi_{Tj}^{(2d)} &= e_j^{(2d)},\ j = d+3, d+4, \ldots, 2d.
\end{aligned}
$$

Then, using (3.16), (3.17) and (3.18), the data matrix $\mathbf{X}^{(d)}$ and $\mathbf{Y}^{(d)}$ can be expressed as,

$$
\begin{bmatrix} \mathbf{X}^{(d)} \\ \mathbf{Y}^{(d)} \end{bmatrix} = \begin{bmatrix}
a_1 z_{X1\bullet} + a_2 z_{X2\bullet} + a_3 z_{Y1\bullet} + a_4 z_{Y2\bullet} \\
a_5 z_{X1\bullet} + a_6 z_{X2\bullet} + a_7 z_{Y1\bullet} + a_8(1) z_{Y2\bullet} \\
\tau_X z_{X3\bullet} \\
\vdots \\
\tau_X z_{Xd\bullet} \\
b_1 z_{X1\bullet} + b_2 z_{X2\bullet} + b_3 z_{Y1\bullet} + b_4 z_{Y2\bullet} \\
b_5 z_{X1\bullet} + b_6 z_{X2\bullet} + b_7 z_{Y1\bullet} + b_8 z_{Y2\bullet} \\
\tau_Y z_{Y3\bullet} \\
\vdots \\
\tau_Y z_{Yd\bullet}
\end{bmatrix}, \tag{3.19}
$$

where a_1, a_3, b_1 and b_3 are random variable of magnitude of $O_P(\sqrt{d^\alpha})$, and the rest of a_i and b_j are of magnitude of $O_P(1)$. Let $c_1, c_2, d1$ and $d2$ denote the first, second, $(d+1)$th and $(d+2)$th row of the matrix (3.19), respectively.

$$
\begin{aligned}
c_1 &= a_1 z_{X1\bullet} + a_2 z_{X2\bullet} + a_3 z_{Y1\bullet} + a_4 z_{Y2\bullet},\\
c_2 &= a_5 z_{X1\bullet} + a_6 z_{X2\bullet} + a_7 z_{Y1\bullet} + a_8 z_{Y2\bullet},\\
d_1 &= b_1 z_{X1\bullet} + b_2 z_{X2\bullet} + b_3 z_{Y1\bullet} + b_4 z_{Y2\bullet},\\
d_2 &= b_5 z_{X1\bullet} + b_6 z_{X2\bullet} + b_7 z_{Y1\bullet} + b_8 z_{Y2\bullet}.
\end{aligned} \tag{3.20}
$$

The sample covariance and cross-covariance matrices $\hat{\mathbf{\Sigma}}_X^{(d)}$, $\hat{\mathbf{\Sigma}}_Y^{(d)}$ and $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$ are found as blocks of the following matrix,

$$
\begin{bmatrix} n\hat{\mathbf{\Sigma}}_X^{(d)} & n\hat{\mathbf{\Sigma}}_{XY}^{(d)} \\ \left(n\hat{\mathbf{\Sigma}}_{XY}^{(d)}\right)^T & n\hat{\mathbf{\Sigma}}_Y^{(d)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(d)} \\ \mathbf{Y}^{(d)} \end{bmatrix} \begin{bmatrix} \mathbf{X}^{(d)} \\ \mathbf{Y}^{(d)} \end{bmatrix}^T, \tag{3.21}
$$

where elementwise-explicit matrices are given as,

$$\hat{\mathbf{\Sigma}}_X^{(d)} =$$
$$\frac{1}{n}\begin{bmatrix} \langle c_1, c_1\rangle & \langle c_1, c_2\rangle & \tau_X\langle c_1, z_{X3\bullet}\rangle & \tau_X\langle c_1, z_{X4\bullet}\rangle & \dots & \tau_X\langle c_1, z_{Xd\bullet}\rangle \\ \langle c_2, c_1\rangle & \langle c_2, c_2\rangle & \tau_X\langle c_2, z_{X3\bullet}\rangle & \tau_X\langle c_2, z_{X4\bullet}\rangle & \dots & \tau_X\langle c_2, z_{Xd\bullet}\rangle \\ \tau_X\langle z_{X3\bullet}, c_1\rangle & \tau_X\langle z_{X3\bullet}, c_2\rangle & \tau_X^2\langle z_{X3\bullet}, z_{X3\bullet}\rangle & \tau_X^2\langle z_{X3\bullet}, z_{4,\bullet}\rangle & \dots & \tau_X^2\langle z_{X3\bullet}, z_{Xd\bullet}\rangle \\ \tau_X\langle z_{X4\bullet}, c_1\rangle & \tau_X\langle z_{X4\bullet}, c_2\rangle & \tau_X^2\langle z_{X4\bullet}, z_{X3\bullet}\rangle & \tau_X^2\langle z_{X4\bullet}, z_{X4\bullet}\rangle & \dots & \tau_X^2\langle z_{X4\bullet}, z_{Xd\bullet}\rangle \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \tau_X\langle z_{Xd\bullet}, c_1\rangle & \tau_X\langle z_{Xd\bullet}, c_2\rangle & \tau_X^2\langle z_{Xd\bullet}, z_{3,\bullet}\rangle & \tau_X^2\langle z_{Xd\bullet}, z_{4,\bullet}\rangle & \dots & \tau_X^2\langle z_{Xd\bullet}, z_{Xd\bullet}\rangle \end{bmatrix},$$

$$\hat{\mathbf{\Sigma}}_Y^{(d)} =$$
$$\frac{1}{n}\begin{bmatrix} \langle d_1, d_1\rangle & \langle d_1, d_2\rangle & \tau_Y\langle d_1, z_{Y3\bullet}\rangle & \tau_Y\langle d_1, z_{Y4\bullet}\rangle & \dots & \tau_Y\langle d_1, z_{Yd\bullet}\rangle \\ \langle d_2, d_1\rangle & \langle d_2, d_2\rangle & \tau_Y\langle d_2, z_{Y3\bullet}\rangle & \tau_Y\langle d_2, z_{Y4\bullet}\rangle & \dots & \tau_Y\langle d_2, z_{Yd\bullet}\rangle \\ \tau_Y\langle z_{Y3\bullet}, d_1\rangle & \tau_Y\langle z_{Y3\bullet}, d_2\rangle & \tau_Y^2\langle z_{Y3\bullet}, z_{Y3\bullet}\rangle & \tau_Y^2\langle z_{Y3\bullet}, z_{Y4\bullet}\rangle & \dots & \tau_Y^2\langle z_{Y3\bullet}, z_{Yd\bullet}\rangle \\ \tau_Y\langle z_{Y4\bullet}, d_1\rangle & \tau_Y\langle z_{Y4\bullet}, d_2\rangle & \tau_Y^2\langle z_{Y4\bullet}, z_{Y3\bullet}\rangle & \tau_Y^2\langle z_{Y4\bullet}, z_{Y4\bullet}\rangle & \dots & \tau_Y^2\langle z_{Y4\bullet}, z_{Yd\bullet}\rangle \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \tau_Y\langle z_{Yd\bullet}, d_1\rangle & \tau_Y\langle z_{Yd\bullet}, d_2\rangle & \tau_Y^2\langle z_{Yd\bullet}, z_{Y3\bullet}\rangle & \tau_Y^2\langle z_{Yd\bullet}, z_{Y4\bullet}\rangle & \dots & \tau_Y^2\langle z_{Yd\bullet}, z_{Yd\bullet}\rangle \end{bmatrix}, \tag{3.22}$$

$$\hat{\mathbf{\Sigma}}_{XY}^{(d)} =$$
$$\frac{1}{n}\begin{bmatrix} \langle c_1, d_1\rangle & \langle c_1, d_2\rangle & \tau_Y\langle c_1, z_{Y3\bullet}\rangle & \tau_Y\langle c_1, z_{Y4\bullet}\rangle & \dots & \tau_Y\langle c_1, z_{Yd\bullet}\rangle \\ \langle c_2, d_1\rangle & \langle c_2, d_2\rangle & \tau_Y\langle c_2, z_{Y3\bullet}\rangle & \tau_Y\langle c_2, z_{Y4\bullet}\rangle & \dots & \tau_Y\langle c_2, z_{Yd\bullet}\rangle \\ \tau_X\langle z_{X3\bullet}, d_1\rangle & \tau_X\langle z_{X3\bullet}, d_2\rangle & \tau_X\tau_Y\langle z_{X3\bullet}, z_{Y3\bullet}\rangle & \tau_X\tau_Y\langle z_{X3\bullet}, z_{Y4\bullet}\rangle & \dots & \tau_X\tau_Y\langle z_{X3\bullet}, z_{Yd\bullet}\rangle \\ \tau_X\langle z_{X4\bullet}, d_1\rangle & \tau_X\langle z_{X4\bullet}, d_2\rangle & \tau_X\tau_Y\langle z_{X4\bullet}, z_{Y3\bullet}\rangle & \tau_X\tau_Y\langle z_{X4\bullet}, z_{Y4\bullet}\rangle & \dots & \tau_X\tau_Y\langle z_{X4\bullet}, z_{Yd\bullet}\rangle \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \tau_X\langle z_{Xd\bullet}, d_1\rangle & \tau_X\langle z_{Xd\bullet}, d_2\rangle & \tau_X\tau_Y\langle z_{Xd\bullet}, z_{Y3\bullet}\rangle & \tau_X\tau_Y\langle z_{Xd\bullet}, z_{Y4\bullet}\rangle & \dots & \tau_X\tau_Y\langle z_{Xd\bullet}, z_{Yd\bullet}\rangle \end{bmatrix}.$$

3.4.2 Notation

Here, we introduce notations to be used in the proof for HDLSS asymptotic behaviors of a random variable.

- For a d-dimensional random variable $X^{(d)}$, we say that $X^{(d)} = o_P(d^\alpha)$, for $\alpha \in R$, if, $\forall \epsilon > 0$,
$$\lim_{d\to\infty} P\left(\left|\frac{X^{(d)}}{d^\alpha} > \epsilon\right|\right) = 0.$$

- For a d-dimensional random variable $X^{(d)}$, we say that $X^{(d)} = O_P(d^\alpha)$, for $\alpha \in R$, if, $\forall \epsilon > 0$, there exists a finite $C > 0$ such that,

$$P\left(\left|\frac{X^{(d)}}{d^\alpha} > M\right|\right) < \epsilon, \ \forall d.$$

- For a d-dimensional random variable $X^{(d)}$, we say that $X^{(d)} \asymp d^\alpha$, for $\alpha \in R$, if,

$$X^{(d)} = O_P(d^\alpha) \text{ and } X^{(d)} \neq o_P(d^\alpha).$$

3.4.3 Case of $\alpha > 1$

First, we prove Theorem 1 under the condition of $\alpha > 1$ with the spiked model of the population covariance structure of $X^{(d)}$ and $Y^{(d)}$ described in (3.9).

3.4.3.1 Behavior of the sample cross-covariance matrix We now investigate the HDLSS asymptotic behavior of the sample cross-covariance matrix $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$ given in (3.22), in specific, its sample singular values and singular vectors. The singular value decomposition (SVD) of $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$ gives,

$$\hat{\mathbf{\Sigma}}_{XY}^{(d)} = \sum_{i=1}^{n} \hat{\lambda}_{XYi}^{(d)} \hat{\eta}_{Xi}^{(d)} \left(\hat{\eta}_{Yi}^{(d)}\right)^T, \tag{3.23}$$

where $\hat{\lambda}_{XY1}^{(d)} \geq \hat{\lambda}_{XY2}^{(d)} \geq \cdots \geq \hat{\lambda}_{XYn}^{(d)} \geq 0$, $\|\hat{\eta}_{Xi}^{(d)}\|_2 = \|\hat{\eta}_{Yi}^{(d)}\|_2 = 1$ for $i = 1, 2, \ldots, n$, and $\langle \hat{\eta}_{Xi}^{(d)}, \hat{\eta}_{Xj}^{(d)} \rangle = \langle \hat{\eta}_{Yi}^{(d)}, \hat{\eta}_{Yj}^{(d)} \rangle = 0$, for $i \neq j$.

Lemma 2 (CCA HDLSS Asymptotic Lemma 1.)**.** *Let C_{XY} and $\mathbf{M}_{XY}$ be,*

$$C_{XY} = \lim_{d\to\infty} \frac{\langle c_1, d_1 \rangle}{d^\alpha}, \ \mathbf{M}_{XY} = \begin{bmatrix} C_{XY} & \underset{1\times(d-1)}{\mathbf{0}} \\ \underset{(d-1)\times 1}{\mathbf{0}} & \underset{(d-1)\times(d-1)}{\mathbf{0}} \end{bmatrix},$$

where c_1 and d_1 are defined in (3.20) and so C_{XY} is a non-degenerate random variable. Then, for $\alpha > 1$,

$$\left\| \frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \mathbf{M}_{XY} \right\|_F^2 \xrightarrow[d\to\infty]{p} 0,$$

where $\| \bullet \|_F$ is the Frobenious norm of a matrix.

Proof. Let $\mathbf{M}_{XY}^{(d)}$ be,

$$\mathbf{M}_{XY}^{(d)} = \begin{bmatrix} \frac{\langle c_1, d_1 \rangle}{d^\alpha} & \underset{1\times(d-1)}{\mathbf{0}} \\ \underset{(d-1)\times 1}{\mathbf{0}} & \underset{(d-1)\times(d-1)}{\mathbf{0}} \end{bmatrix}.$$

It is obvious to see that,

$$\left\| \mathbf{M}_{XY}^{(d)} - \mathbf{M}_{XY} \right\|_F^2 \xrightarrow[d\to\infty]{p} 0.$$

Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left\| \frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \mathbf{M}_{XY}^{(d)} \right\|_F^2 =& \frac{\langle c_1, d_2 \rangle^2}{d^{2\alpha}} + \frac{\langle c_2, d_1 \rangle^2}{d^{2\alpha}} + \frac{\langle c_2, d_2 \rangle^2}{d^{2\alpha}} + \tau_Y^2 \sum_{i=3}^{d} \frac{\langle c_1, z_{Yi\bullet} \rangle^2}{d^{2\alpha}} \\
&+ \tau_Y^2 \sum_{i=3}^{d} \frac{\langle c_2, z_{Yi\bullet} \rangle^2}{d^{2\alpha}} + \tau_X^2 \sum_{i=3}^{d} \frac{\langle z_{Xi\bullet}, d_1 \rangle^2}{d^{2\alpha}} \\
&+ \tau_X^2 \sum_{i=3}^{d} \frac{\langle z_{Xi\bullet}, d_2 \rangle^2}{d^{2\alpha}} + \tau_X^2 \tau_Y^2 \sum_{i=3}^{d} \sum_{j=3}^{d} \frac{\langle z_{Xi\bullet}, z_{Yj\bullet} \rangle^2}{d^{2\alpha}} \\
\leq& \frac{\langle c_1, d_2 \rangle^2}{d^{2\alpha}} + \frac{\langle c_2, d_1 \rangle^2}{d^{2\alpha}} + \frac{\langle c_2, d_2 \rangle^2}{d^{2\alpha}} + \tau_Y^2 \sum_{i=3}^{d} \frac{\|c_1\|_2^2 \|z_{Yi\bullet}\|_2^2}{d^{2\alpha}} \\
&+ \tau_Y^2 \sum_{i=3}^{d} \frac{\|c_2\|_2^2 \|z_{Yi\bullet}\|_2^2}{d^{2\alpha}} + \tau_X^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2 \|d_1\|_2^2}{d^{2\alpha}} \\
&+ \tau_X^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2 \|d_2\|_2^2}{d^{2\alpha}} + \tau_X^2 \tau_Y^2 \sum_{i=3}^{d} \sum_{j=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2 \|z_{Yj\bullet}\|_2^2}{d^{2\alpha}}.
\end{aligned}$$

Since $\langle c_1, d_2 \rangle^2$, $\langle c_2, d_1 \rangle^2$ are $O_P(\sqrt{d^\alpha})$, and $\langle c_2, d_2 \rangle^2$ is $O_P(1)$,

$$\frac{\langle c_1, d_2 \rangle^2}{d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0, \ \frac{\langle c_2, d_1 \rangle^2}{d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0, \ \frac{\langle c_2, d_2 \rangle^2}{d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0.$$

It is not hard to see that $\|z_{Xi\bullet}\|_2^2 \sim \chi_n^2$ and $\|z_{Yi\bullet}\|_2^2 \sim \chi_n^2$, since the elements of $z_{Xi\bullet}$ (respectively $z_{Yi\bullet}$) are the ith components of the i.i.d. multivariate standard normal samples of size n. Note that the magnitudes of $\|c_1\|_2^2, \|c_2\|_2^2, \|d_1\|_2^2$ and $\|d_2\|_2^2$ are of $O_P(d^\alpha)$. Applying the law of large numbers, we have,

$$\begin{aligned}
\tau_Y^2 \sum_{i=3}^{d} \frac{\|c_1\|_2^2 \|z_{Yi\bullet}\|_2^2}{d^{2\alpha}} &= \frac{\tau_Y^2}{d^{\alpha-1}} \left\| \frac{c_1}{\sqrt{d^\alpha}} \right\|_2^2 \sum_{i=3}^{d} \frac{\|z_{Yi\bullet}\|_2^2}{d} \\
&\xrightarrow[d\to\infty]{p} 0 \times O_P(1) \times E(\chi_n^2) = 0, \\
\tau_Y^2 \sum_{i=3}^{d} \frac{\|c_2\|_2^2 \|z_{Yi\bullet}\|_2^2}{d^{2\alpha}} &= \frac{\tau_Y^2}{d^{\alpha-1}} \left\| \frac{c_2}{\sqrt{d^\alpha}} \right\|_2^2 \sum_{i=3}^{d} \frac{\|z_{Yi\bullet}\|_2^2}{d}
\end{aligned}$$

$$\xrightarrow[d\to\infty]{p} 0 \times 0 \times E(\chi_n^2) = 0,$$

$$\tau_X^2 \sum_{i=3}^{d} \frac{\|d_1\|_2^2 \|z_{Xi\bullet}\|_2^2}{d^{2\alpha}} = \frac{\tau_X^2}{d^{\alpha-1}} \left\| \frac{d_1}{\sqrt{d^\alpha}} \right\|_2^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2}{d}$$

$$\xrightarrow[d\to\infty]{p} 0 \times O_P(1) \times E(\chi_n^2) = 0,$$

$$\tau_X^2 \sum_{i=3}^{d} \frac{\|d_2\|_2^2 \|z_{Xi\bullet}\|_2^2}{d^{2\alpha}} = \frac{\tau_X^2}{d^{\alpha-1}} \left\| \frac{d_2}{\sqrt{d^\alpha}} \right\|_2^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2}{d}$$

$$\xrightarrow[d\to\infty]{p} 0 \times 0 \times E(\chi_n^2) = 0.$$

Using $\alpha > 1$ and the law of large numbers,

$$\tau_X^2 \tau_Y^2 \sum_{i=3}^{d} \sum_{j=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2 \|z_{Yj\bullet}\|_2^2}{d^{2\alpha}} = \frac{\tau_X^2 \tau_Y^2}{d^{2\alpha-2}} \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2}{d} \sum_{j=3}^{d} \frac{\|z_{Yj\bullet}\|_2^2}{d}$$

$$\xrightarrow[d\to\infty]{p} 0 \times E(\chi_n^2) \times E(\chi_n^2) = 0.$$

Therefore,

$$\left\| \frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \mathbf{M}_{XY}^{(d)} \right\|_F^2 \xrightarrow[d\to\infty]{p} 0.$$

□

Lemma 3 (CCA HDLSS Asymptotic Lemma 2.)**.** *Let $\mathbf{M}_{XY}$ be the matrix defined in Lemma 2. Let $\hat{\lambda}_{XY1}^{(d)}$, $\hat{\eta}_{X1}^{(d)}$ and $\hat{\eta}_{Y1}^{(d)}$ be the first sample singular value and singular vectors from (3.23). Then, for $\alpha > 1$,*

$$\left\| \frac{n\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \hat{\eta}_{X1}^{(d)} \left(\hat{\eta}_{Y1}^{(d)}\right)^T - \mathbf{M}_{XY} \right\|_F^2 \xrightarrow[d\to\infty]{p} 0.$$

Proof. Using $\mathbf{M}_{XY}^{(d)}$ defined in Lemma 2 and the triangle inequality,

$$\begin{aligned}
&\left\| \frac{n\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \hat{\eta}_{X1}^{(d)} \left(\hat{\eta}_{Y1}^{(d)}\right)^T - \mathbf{M}_{XY}^{(d)} \right\|_F \\
&\quad = \left\| \left(\frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \mathbf{M}_{XY}^{(d)} \right) - \left(\frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \hat{\eta}_{X1}^{(d)} \left(\hat{\eta}_{Y1}^{(d)}\right)^T \right) \right\|_F \\
&\quad \leq \left\| \frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \mathbf{M}_{XY}^{(d)} \right\|_F + \left\| \frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \hat{\eta}_{X1}^{(d)} \left(\hat{\eta}_{Y1}^{(d)}\right)^T \right\|_F.
\end{aligned}$$

By Lemma 2,

$$\left\| \frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \mathbf{M}_{XY}^{(d)} \right\|_F^2 \xrightarrow[d\to\infty]{p} 0.$$

Write $\mathbf{M}_{XY}^{(d)}$ as,

$$\mathbf{M}_{XY}^{(d)} = \frac{\langle c_1, d_1 \rangle}{d^\alpha} e_1^{(d)} \left(e_1^{(d)}\right)^T.$$

Since the first sample singular value $\hat{\lambda}_{XY1}^{(d)}$ and singular vectors $\hat{\eta}_{X1}^{(d)}$ and $\hat{\eta}_{Y1}^{(d)}$ provide the best rank-1 approximation to $\hat{\lambda}_{XY1}^{(d)}/d^\alpha$,

$$\begin{aligned}
\left\| \frac{\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \hat{\eta}_{X1}^{(d)} \left(\hat{\eta}_{Y1}^{(d)}\right)^T \right\|_F^2 &\leq \left\| \frac{\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \frac{\langle c_1, d_1 \rangle}{nd^\alpha} e_1^{(d)} \left(e_1^{(d)}\right)^T \right\|_F^2 \\
&= \left\| \frac{\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d^\alpha} - \frac{\mathbf{M}_{XY}^{(d)}}{n} \right\|_F^2 \\
&\xrightarrow[d\to\infty]{p} 0.
\end{aligned}$$

□

Lemma 4 (CCA HDLSS Asymptotic Lemma 3.)**.** *For $\alpha > 1$, the first singular value $\hat{\lambda}_{XY1}^{(d)}$ and singular vectors $\hat{\eta}_{X1}^{(d)}$ and $\hat{\eta}_{Y1}^{(d)}$ converge in probability to the following quantities as $d \to \infty$,*

$$\frac{n\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \xrightarrow[d\to\infty]{p} C_{XY}, \ \left\| \hat{\eta}_{X1}^{(d)} - e_1^{(d)} \right\|_2^2 \xrightarrow[d\to\infty]{p} 0, \ \left\| \hat{\eta}_{Y1}^{(d)} - e_1^{(d)} \right\|_2^2 \xrightarrow[d\to\infty]{p} 0,$$

where C_{XY} is defined in Lemma 2.

Proof. Let $\tilde{\eta}_{X1}^{(d)}$ and $\tilde{\eta}_{Y1}^{(d)}$ be $\hat{\eta}_{X1}^{(d)}$ and $\hat{\eta}_{Y1}^{(d)}$ where their first entries are set to 0. Let $\hat{\eta}_{X1}^{(d)}(i)$ and $\hat{\eta}_{Y1}^{(d)}(j)$ be the ith and jth entries of $\hat{\eta}_{X1}^{(d)}$ and $\hat{\eta}_{Y1}^{(d)}$. By Lemma 3,

$$\begin{aligned}
\left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \right)^2 \left\| \tilde{\eta}_{X1}^{(d)} \right\|_2^2 \left\| \tilde{\eta}_{Y1}^{(d)} \right\|_2^2 &= \left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \right)^2 \sum_{i=2}^{d} \left(\hat{\eta}_{X1}^{(d)}(i) \right)^2 \sum_{j=2}^{d} \left(\hat{\eta}_{Y1}^{(d)}(j) \right)^2 \\
&= \left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \right)^2 \sum_{i=2}^{d} \sum_{j=2}^{d} \left(\hat{\eta}_{X1}^{(d)}(i) \right)^2 \left(\hat{\eta}_{Y1}^{(d)}(j) \right)^2 \\
&\xrightarrow[d\to\infty]{p} 0.
\end{aligned}$$

First, we show that $(\hat{\lambda}_{XY1}^{(d)}/d^\alpha)^2 > 0$ in probability. Suppose that $(\hat{\lambda}_{XY1}^{(d)}/d^\alpha)^2$ converges in probability to 0. Then, noting that $\|\hat{\eta}_{X1}^{(d)}\|_2^2 = 1$ and $\|\hat{\eta}_{Y1}^{(d)}\|_2^2 = 1$,

$$\begin{aligned}
\left\| \frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \hat{\eta}_{X1}^{(d)} \left(\hat{\eta}_{Y1}^{(d)}\right)^T \right\|_F^2 &= \left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \right)^2 \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\hat{\eta}_{X1}^{(d)}(i) \right)^2 \left(\hat{\eta}_{Y1}^{(d)}(j) \right)^2 \\
&= \left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha} \right)^2 \sum_{i=1}^{d} \left(\hat{\eta}_{X1}^{(d)}(i) \right)^2 \sum_{j=1}^{d} \left(\hat{\eta}_{Y1}^{(d)}(j) \right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}}\right)^2 \left\|\hat{\eta}_{X1}^{(d)}\right\|_2^2 \left\|\hat{\eta}_{Y1}^{(d)}\right\|_2^2 \\
&= \left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}}\right)^2 \\
&\xrightarrow[d\to\infty]{p} 0,
\end{aligned}
$$

which contradicts to Lemma 3 (note that the limiting matrix M is not a degenerate matrix). Therefore,

$$
\left\|\tilde{\eta}_{X1}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{p} 0, \text{ or } \left\|\hat{\eta}_{Y1}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{p} 0.
$$

We want to show that both $\|\tilde{\eta}_{X1}^{(d)}\|_2^2$ and $\|\tilde{\eta}_{Y1}^{(d)}\|_2^2$ converge to 0 in probability. Suppose that $\|\tilde{\eta}_{X1}^{(d)}\|_2^2$ converges to 0 and $\|\tilde{\eta}_{Y1}^{(d)}\|_2^2 > 0$ in probability. Since the norm of $\hat{\eta}_{X1}^{(d)}$ is 1, its first entry $\hat{\eta}_{X1}^{(d)}(1)$ converges in probability to,

$$
\left(\hat{\eta}_{X1}^{(d)}(1)\right)^2 = 1 - \left\|\tilde{\eta}_{X1}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{p} 1.
$$

Then, the squared sum of the entries in the first row of $(\hat{\lambda}_{XY1}^{(d)}/d^{\alpha})\hat{\eta}_{X1}^{(d)}\left(\hat{\eta}_{Y1}^{(d)}\right)^T$ with the first entry excluded becomes,

$$
\left\|\frac{\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}}\hat{\eta}_{X1}^{(d)}(1)\tilde{\eta}_{Y1}^{(d)}\right\|_2^2 = \left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}}\right)^2 \left(\hat{\eta}_{X1}^{(d)}(1)\right)^2 \sum_{j=2}^{d} \left(\hat{\eta}_{Y1}^{(d)}(j)\right)^2 \xrightarrow[d\to\infty]{p} > 0,
$$

which contradicts to Lemma 3, according to which the above quantity should converge to 0. Argument is similar for the case where $\|\tilde{\eta}_{Y1}^{(d)}\|_2^2$ converges to 0 and $\|\tilde{\eta}_{X1}^{(d)}\|_2^2 > 0$ in probability. Hence we have

$$
\left\|\tilde{\eta}_{X1}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{p} 0, \text{ and } \left\|\hat{\eta}_{Y1}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{p} 0.
$$

Note that, since the norm of $\hat{\eta}_{Y1}^{(d)}$ is 1,

$$
\left(\hat{\eta}_{Y1}^{(d)}(1)\right)^2 = 1 - \left\|\tilde{\eta}_{Y1}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{p} 1.
$$

Therefore

$$
\begin{aligned}
\left\|\hat{\eta}_{X1}^{(d)} - e_1^{(d)}\right\|_2^2 &= \left(\hat{\eta}_{X1}^{(d)}(1) - 1\right)^2 + \left\|\tilde{\eta}_{X1}^{(d)}(i)\right\|_2^2 \xrightarrow[d\to\infty]{p} 0, \\
\left\|\hat{\eta}_{Y1}^{(d)} - e_1^{(d)}\right\|_2^2 &= \left(\hat{\eta}_{Y1}^{(d)}(1) - 1\right)^2 + \left\|\tilde{\eta}_{Y1}^{(d)}(i)\right\|_2^2 \xrightarrow[d\to\infty]{p} 0.
\end{aligned}
$$

To calculate the limiting value of $\hat{\lambda}_{XY1}^{(d)}/d^{\alpha}$ as $d \to \infty$, using the unitary invariance property of the Frobenius norm and the previous result about the limiting vectors of $\hat{\eta}_{X1}^{(d)}$ and $\hat{\eta}_{Y1}^{(d)}$ as $d \to \infty$,

$$\begin{aligned}
&\left\| \frac{n\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}} \hat{\eta}_{X1}^{(d)} \left(\hat{\eta}_{Y1}^{(d)}\right)^T - \mathbf{M}_{XY}^{(d)} \right\|_F^2 \\
&= \left\| \frac{n\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}} \hat{\eta}_{X1}^{(d)} \left(\hat{\eta}_{Y1}^{(d)}\right)^T - \frac{\langle c_1, d_1 \rangle}{d^{\alpha}} e_1^{(d)} \left(e_1^{(d)}\right)^T \right\|_F^2 \\
&= \left\| \frac{n\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}} \left(\hat{\eta}_{Y1}^{(d)}\right)^T \hat{\eta}_{X1}^{(d)} \left(\hat{\eta}_{Y1}^{(d)}\right)^T \hat{\eta}_{X1}^{(d)} - \frac{\langle c_1, d_1 \rangle}{d^{\alpha}} \left(\hat{\eta}_{Y1}^{(d)}\right)^T e_1^{(d)} \left(e_1^{(d)}\right)^T \hat{\eta}_{X1}^{(d)} \right\|_F^2 \\
&= \left\| \frac{n\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}} - \frac{\langle c_1, d_1 \rangle}{d^{\alpha}} \left(\hat{\eta}_{Y1}^{(d)}\right)^T e_1^{(d)} \left(e_1^{(d)}\right)^T \hat{\eta}_{X1}^{(d)} \right\|_F^2 \\
&= \left(\frac{\langle c_1, d_1 \rangle}{d^{\alpha}} \right)^2 \left\| \frac{n\hat{\lambda}_{XY1}^{(d)}/d^{\alpha}}{\langle c_1, d_1 \rangle/d^{\alpha}} - \left(\hat{\eta}_{Y1}^{(d)}\right)^T e_1^{(d)} \left(e_1^{(d)}\right)^T \hat{\eta}_{X1}^{(d)} \right\|_F^2 \\
&\xrightarrow[d\to\infty]{p} \mathrm{A}^2 \left(\frac{\mathrm{B}}{\mathrm{C}} - 1 \right)^2 .
\end{aligned}$$

where,

$$\mathrm{A} = \lim_{d\to\infty} \frac{\langle c_1, d_1 \rangle}{d^{\alpha}}, \ \mathrm{B} = \lim_{d\to\infty} \frac{n\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}}, \ \mathrm{C} = \lim_{d\to\infty} \frac{\langle c_1, d_1 \rangle}{d^{\alpha}}.$$

By Lemma 2,

$$\left\| \frac{n\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}} \hat{\eta}_{X1}^{(d)} \hat{\eta}_{Y1}^{(d)} - \mathbf{M}_{XY}^{(d)} \right\|_F^2 \xrightarrow[d\to\infty]{p} 0.$$

Since $\mathrm{A} \asymp O_p(1)$,

$$\left(\frac{B}{C} - 1 \right)^2 \xrightarrow[d\to\infty]{p} 0.$$

Therefore,

$$\frac{B}{C} \xrightarrow[d\to\infty]{p} 1.$$

□

Lemma 5 (CCA HDLSS Asymptotic Lemma 4.)**.** *The magnitudes of the entries $\hat{\eta}_{X1}^{(d)}(i)$ and $\hat{\eta}_{Y1}^{(d)}(i)$, $i = 2, 3, .., d$, of the first sample singular vectors $\hat{\eta}_{X1}^{(d)}$ and $\hat{\eta}_{Y1}^{(d)}$ are of $O_p(1/\sqrt{d^{\alpha}})$ as $d \to \infty$.*

Proof. Let $\hat{\mathbf{\Sigma}}_{XY}^{(d)}(i,j)$ be the entry of the ith row and jth column of $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$. Consider $\hat{\mathbf{\Sigma}}_{XY}^{(d)}(1,3)$, which is $O_p(\sqrt{d^{\alpha}})$. The contribution of $\hat{\lambda}_{XY1}^{(d)}$, $\hat{\eta}_{X1}^{(d)}$ and $\hat{\eta}_{Y1}^{(d)}$ to $\hat{\mathbf{\Sigma}}_{XY}^{(d)}(1,3)$ is,

$$\hat{\lambda}_{XY1}^{(d)} \hat{\eta}_{X1}^{(d)}(1) \hat{\eta}_{Y1}^{(d)}(3).$$

From Lemma 4,

$$\hat{\lambda}_{XY1}^{(d)} \asymp O_p(d^{\alpha}),\ \hat{\eta}_{X1}^{(d)}(1) \xrightarrow[d\to\infty]{p} 1.$$

Therefore,

$$\hat{\eta}_{Y1}^{(d)}(3) = O_p\left(\frac{1}{\sqrt{d^{\alpha}}}\right).$$

Argument is similar for the rest of entries. □

Lemma 6 (CCA HDLSS Asymptotic Lemma 5.)**.** *The magnitudes of the sample singular values* $\hat{\lambda}_{XYi}^{(d)}$ *for* $i = 2, 3, .., n$ *are of* $O_p(d)$.

Proof. Let $\tilde{\mathbf{\Sigma}}_{XY}^{(d)}$ be the $(d-1)\times(d-1)$ sample cross-covariance matrix $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$ from which the first row and column are set to 0. Let $\tilde{\eta}_{Xi}^{(d)}$ and $\tilde{\eta}_{Yi}^{(d)}$ be $\hat{\eta}_{Xi}^{(d)}$ and $\hat{\eta}_{Yi}^{(d)}$ where their first entries are set to 0. Then,

$$\begin{aligned}
&\frac{n\hat{\lambda}_{XY1}^{(d)}}{d}\tilde{\eta}_{X1}^{(d)}\left(\tilde{\eta}_{Y1}^{(d)}\right)^T + \sum_{i=2}^{n}\frac{n\hat{\lambda}_{XYi}^{(d)}}{d}\tilde{\eta}_{Xi}^{(d)}\left(\tilde{\eta}_{Yi}^{(d)}\right)^T = \frac{n\tilde{\mathbf{\Sigma}}_{XY}^{(d)}}{d},\\
&\sum_{i=2}^{n}\frac{n\hat{\lambda}_{XYi}^{(d)}}{d}\tilde{\eta}_{Xi}^{(d)}\left(\tilde{\eta}_{Yi}^{(d)}\right)^T = \frac{n\tilde{\mathbf{\Sigma}}_{XY}^{(d)}}{d} - \frac{n\hat{\lambda}_{XY1}^{(d)}}{d}\tilde{\eta}_{X1}^{(d)}\left(\tilde{\eta}_{Y1}^{(d)}\right)^T,\\
&\left\|\sum_{i=2}^{n}\frac{n\hat{\lambda}_{XYi}^{(d)}}{d}\tilde{\eta}_{Xi}^{(d)}\left(\tilde{\eta}_{Yi}^{(d)}\right)^T\right\|_F \le \left\|\frac{n\tilde{\mathbf{\Sigma}}_{XY}^{(d)}}{d}\right\|_F + \left\|\frac{n\hat{\lambda}_{XY1}^{(d)}}{d}\tilde{\eta}_{X1}^{(d)}\left(\tilde{\eta}_{Y1}^{(d)}\right)^T\right\|_F.
\end{aligned}$$

Note that $\|c_2\|_2^2$ and $\|d_2\|_2^2$ are $O_P(1)$ and that $\|z_{Yi\bullet}\|_2^2 \sim \chi_n^2$ and $\|z_{Xi\bullet}\|_2^2 \sim \chi_n^2$. Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left\|\frac{n\tilde{\mathbf{\Sigma}}_{XY}^{(d)}}{d}\right\|_F^2 &= \frac{\langle c_2, d_2\rangle^2}{d^2} + \tau_Y^2\sum_{i=3}^{d}\frac{\langle c_2, z_{Yi\bullet}\rangle^2}{d^2}\\
&\quad + \tau_X^2\sum_{i=3}^{d}\frac{\langle z_{Xi\bullet}, d_2\rangle^2}{d^2} + \tau_X^2\tau_Y^2\sum_{i=3}^{d}\sum_{j=3}^{d}\frac{\langle z_{Xi\bullet}, z_{Yj\bullet}\rangle^2}{d^2}\\
&\le \frac{\langle c_2, d_2\rangle^2}{d^2} + \tau_Y^2\sum_{i=3}^{d}\frac{\|c_2\|_2^2\|z_{Yi\bullet}\|_2^2}{d^2}\\
&\quad + \tau_X^2\sum_{i=3}^{d}\frac{\|z_{Xi\bullet}\|_2^2\|d_2\|_2^2}{d^2} + \tau_X^2\tau_Y^2\sum_{i=3}^{d}\sum_{j=3}^{d}\frac{\|z_{Xi\bullet}\|_2^2\|z_{Yj\bullet}\|_2^2}{d^2}\\
&= \frac{\langle c_2, d_2\rangle^2}{d^2} + \tau_Y^2\left\|\frac{c_2}{\sqrt{d}}\right\|_2^2\sum_{i=3}^{d}\frac{\|z_{Yi\bullet}\|_2^2}{d}\\
&\quad + \tau_Y^2\left\|\frac{d_2}{\sqrt{d}}\right\|_2^2\sum_{i=3}^{d}\frac{\|z_{Yi\bullet}\|_2^2}{d} + \tau_X^2\tau_Y^2\sum_{i=3}^{d}\frac{\|z_{Xi\bullet}\|_2^2}{d}\sum_{j=3}^{d}\frac{\|z_{Yj\bullet}\|_2^2}{d}
\end{aligned}$$

$$\xrightarrow[d\to\infty]{p} 0+0+0+\tau_X^2\tau_Y^2E^2(\chi_n^2)$$
$$=\tau_X^2\tau_Y^2n^2.$$

Using Lemma 4 and 5 on the magnitudes of $\hat{\lambda}_{XY1}^{(d)}$ and the entries of $\hat{\eta}_{X1}^{(d)}(i)$ and $\hat{\eta}_{Y1}^{(d)}(i)$,

$$\hat{\lambda}_{XY1}^{(d)} \asymp O_p(d^\alpha),\ \hat{\eta}_{X1}^{(d)}(i) = O_p(\frac{1}{\sqrt{d^\alpha}}), i=2,3,..,d.$$

Then we have,

$$\begin{aligned}
\left\|\frac{\hat{\lambda}_{XY1}^{(d)}}{d}\tilde{\eta}_{X1}^{(d)}\left(\tilde{\eta}_{Y1}^{(d)}\right)^T\right\|_F^2 &= \left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d}\right)^2\sum_{i=2}^d\sum_{j=2}^d\left(\hat{\eta}_{X1}^{(d)}(i)\right)^2\left(\hat{\eta}_{X1}^{(d)}(j)\right)^2\\
&=\left(\hat{\lambda}_{XY1}^{(d)}\right)^2\sum_{i=2}^d\frac{\left(\hat{\eta}_{X1}^{(d)}(i)\right)^2}{d}\sum_{j=2}^d\frac{\left(\hat{\eta}_{X1}^{(d)}(j)\right)^2}{d}\\
&=\left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d^\alpha}\right)^2\sum_{i=2}^d\frac{\left(\sqrt{d^\alpha}\hat{\eta}_{X1}^{(d)}(i)\right)^2}{d}\sum_{j=2}^d\frac{\left(\sqrt{d^\alpha}\hat{\eta}_{X1}^{(d)}(j)\right)^2}{d}\\
&=O_p(1).
\end{aligned}$$

Hence

$$\left\|\sum_{i=2}^n\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\tilde{\eta}_{Xi}^{(d)}\left(\tilde{\eta}_{Y1}^{(d)}\right)^T\right\|_F^2 = O_p(1). \tag{3.24}$$

Expanding the squared Frobenius norm of the above matrix,

$$\begin{aligned}
\left\|\sum_{i=2}^n\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\tilde{\eta}_{Xi}^{(d)}\left(\tilde{\eta}_{Y1}^{(d)}\right)^T\right\|_F^2 &= \sum_{j=2}^d\sum_{k=2}^d\left(\sum_{i=2}^n\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\hat{\eta}_{Xi}^{(d)}(j)\hat{\eta}_{Yi}^{(d)}(k)\right)^2\\
&=\sum_{i=2}^n\left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\right)^2\sum_{j=2}^d\sum_{k=2}^d\left(\hat{\eta}_{Xi}^{(d)}(j)\right)^2\left(\hat{\eta}_{Yi}^{(d)}(k)\right)^2\\
&\quad+2\sum_{i\neq i'}\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\frac{\hat{\lambda}_{XYi'}^{(d)}}{d}\sum_{j=2}^d\sum_{k=2}^d\left(\hat{\eta}_{Xi}^{(d)}(j)\hat{\eta}_{Xi'}^{(d)}(j)\right)\left(\hat{\eta}_{Yi}^{(d)}(k)\hat{\eta}_{Yi'}^{(d)}(k)\right)\\
&=\sum_{i=2}^n\left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\right)^2\sum_{j=2}^d\left(\hat{\eta}_{Xi}^{(d)}(j)\right)^2\sum_{k=2}^d\left(\hat{\eta}_{Yi}^{(d)}(k)\right)^2\\
&\quad+2\sum_{i\neq i'}\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\frac{\hat{\lambda}_{XYi'}^{(d)}}{d}\sum_{j=2}^d\left(\hat{\eta}_{Xi}^{(d)}(j)\hat{\eta}_{Xi'}^{(d)}(j)\right)\sum_{k=2}^d\left(\hat{\eta}_{Yi}^{(d)}(k)\hat{\eta}_{Yi'}^{(d)}(k)\right)\\
&=\sum_{i=2}^n\left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\right)^2\left\|\tilde{\eta}_{Xi}^{(d)}\right\|_2^2\left\|\tilde{\eta}_{Yi}^{(d)}\right\|_2^2
\end{aligned}$$

$$+2\sum_{2\le i,i'\le n,\, i\neq i'}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\frac{\hat{\lambda}^{(d)}_{XYi'}}{d}\left\langle\tilde{\eta}^{(d)}_{Xi},\tilde{\eta}^{(d)}_{Xi'}\right\rangle\left\langle\tilde{\eta}^{(d)}_{Yi},\hat{\eta}^{(d)}_{Yi'}\right\rangle.$$

We, to simplify the last part in the above equalities, show that $\|\tilde{\eta}^{(d)}_{Xi}\|_2^2$ and $\|\tilde{\eta}^{(d)}_{Yi}\|_2^2$, for $i=2,3,\ldots,n$, approach to 1 and that $\langle\tilde{\eta}^{(d)}_{Xi},\tilde{\eta}^{(d)}_{Xi'}\rangle$ and $\langle\tilde{\eta}^{(d)}_{Yi},\tilde{\eta}^{(d)}_{Yi'}\rangle$, for $i,i'=2,3,\ldots,n$ and $i\neq i'$, approach to 0 as $d\to\infty$. Consider $\|\tilde{\eta}^{(d)}_{X2}\|_2^2$. Since $\|\hat{\eta}^{(d)}_{X1}-e^{(d)}_1\|_2$ converges in probability to 0 by Lemma 4 and $\|\hat{\eta}^{(d)}_{X1}\|_2^2=1$,

$$\left\|\tilde{\eta}^{(d)}_{X1}\right\|_2^2=1-\left(\hat{\eta}^{(d)}_{X1}(1)\right)^2\xrightarrow[d\to\infty]{p}0,$$

which leads to,

$$\left\langle\tilde{\eta}^{(d)}_{X1},\tilde{\eta}^{(d)}_{X2}\right\rangle^2\le\left\|\tilde{\eta}^{(d)}_{X1}\right\|_2^2\left\|\tilde{\eta}^{(d)}_{X2}\right\|_2^2\le\left\|\tilde{\eta}^{(d)}_{X1}\right\|_2^2\times 1\xrightarrow[d\to\infty]{p}0.$$

Using the orthogonality of $\hat{\eta}^{(d)}_{X1}$ and $\hat{\eta}^{(d)}_{X2}$,

$$\left\langle\hat{\eta}^{(d)}_{X1},\hat{\eta}^{(d)}_{X2}\right\rangle=\hat{\eta}^{(d)}_{X1}(1)\hat{\eta}^{(d)}_{X2}(1)+\left\langle\tilde{\eta}^{(d)}_{X1},\tilde{\eta}^{(d)}_{X2}\right\rangle=0.$$

Since $\hat{\eta}^{(d)}_{X1}(1)$ converges in probability to 1 by Lemma 4, we have,

$$\hat{\eta}^{(d)}_{X2}(1)\xrightarrow[d\to\infty]{p}0.$$

Since $\|\hat{\eta}^{(d)}_{X2}\|_2^2=1$,

$$\left\|\tilde{\eta}^{(d)}_{X2}\right\|_2^2\xrightarrow[d\to\infty]{p}1.$$

Simiarly,

$$\left\|\tilde{\eta}^{(d)}_{Xi}\right\|_2^2\xrightarrow[d\to\infty]{p}1,\ \left\|\tilde{\eta}^{(d)}_{Yi}\right\|_2^2\xrightarrow[d\to\infty]{p}1,\ i=2,3,\ldots,n.$$

Now consider $\langle\tilde{\eta}^{(d)}_{X2},\tilde{\eta}^{(d)}_{X3}\rangle$. Since $\hat{\eta}^{(d)}_{X2}$ and $\hat{\eta}^{(d)}_{X3}$ are orthogonal,

$$\left\langle\hat{\eta}^{(d)}_{X2},\hat{\eta}^{(d)}_{X3}\right\rangle=\hat{\eta}^{(d)}_{X2}(1)\hat{\eta}^{(d)}_{X3}(1)+\left\langle\tilde{\eta}^{(d)}_{X2},\tilde{\eta}^{(d)}_{X3}\right\rangle=0.$$

We showed that $\|\tilde{\eta}^{(d)}_{X2}\|_2^2$ and $\|\tilde{\eta}^{(d)}_{X3}\|_2^2$ converge in probability to 0, which implies,

$$\left\langle\tilde{\eta}^{(d)}_{X2},\tilde{\eta}^{(d)}_{X3}\right\rangle\xrightarrow[d\to\infty]{p}0.$$

Simiarly,

$$\left\langle\tilde{\eta}^{(d)}_{Xi},\tilde{\eta}^{(d)}_{Xi'}\right\rangle\xrightarrow[d\to\infty]{p}0,\left\langle\tilde{\eta}^{(d)}_{Yi},\ \tilde{\eta}^{(d)}_{Yi'}\right\rangle\xrightarrow[d\to\infty]{p}0,\ i,i'=2,3,\ldots,n,\ i\neq i'.$$

Then, by (3.24),

$$\begin{aligned}
&\left\| \sum_{i=2}^{n} \frac{\hat{\lambda}_{XYi}^{(d)}}{d} \tilde{\eta}_{Xi}^{(d)} \left(\tilde{\eta}_{Y1}^{(d)}\right)^T \right\|_F^2 \\
&\quad = \sum_{i=2}^{n} \left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d} \right)^2 \left\| \tilde{\eta}_{Xi}^{(d)} \right\|_2^2 \left\| \tilde{\eta}_{Yi}^{(d)} \right\|_2^2 + 2 \sum_{2 \le i,i' \le n,\, i \ne i'} \frac{\hat{\lambda}_{XYi}^{(d)}}{d} \frac{\hat{\lambda}_{XYi'}^{(d)}}{d} \left\langle \tilde{\eta}_{Xi}^{(d)}, \tilde{\eta}_{Xi'}^{(d)} \right\rangle \left\langle \tilde{\eta}_{Yi}^{(d)}, \hat{\eta}_{Yi'}^{(d)} \right\rangle \\
&\quad \xrightarrow[d\to\infty]{p} \sum_{i=2}^{n} \left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d} \right)^2 = O_p(1).
\end{aligned}$$

Hence, the magnitudes of $\hat{\lambda}_{XYi}^{(d)}$, $i = 2, 3, ..n$ are of $O_p(d)$ as $d \to \infty$. □

Lemma 7 (CCA HDLSS Asymptotic Lemma 6.)**.** *This lemma improves Lemma 6 by providing a precise limiting value. The sample singular values $\hat{\lambda}_{XYi}^{(d)}$ scaled by $1/d$, for $i = 2, 3, .., n$, converge in probability to the following quantity as $d \to \infty$,*

$$\frac{\hat{\lambda}_{XYi}^{(d)}}{d} \xrightarrow[d\to\infty]{p} 0, \ i = 2, 3, \ldots, n.$$

Proof. Let $\tilde{\mathbf{\Sigma}}_{XY}^{(d)}$ be the $(d-1) \times (d-1)$ sample cross-covariance matrix $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$ from which the first row and column are set to 0. We showed that,

$$\left\| \frac{\tilde{\mathbf{\Sigma}}_{XY}^{(d)}}{d} \right\|_F^2 \xrightarrow[d\to\infty]{p} \le \tau_X^2 \tau_Y^2. \tag{3.25}$$

Let $\tilde{\eta}_{Xi}^{(d)}$ and $\tilde{\eta}_{Yi}^{(d)}$ be $\hat{\eta}_{Xi}^{(d)}$ and $\hat{\eta}_{Yi}^{(d)}$ where their first entries are set to 0. Then,

$$\begin{aligned}
\left\| \frac{\tilde{\mathbf{\Sigma}}_{XY}^{(d)}}{d} \right\|_F^2 &= \sum_{j=2}^{d} \sum_{k=2}^{d} \left(\sum_{i=1}^{n} \frac{\hat{\lambda}_{XYi}^{(d)}}{d} \hat{\eta}_{Xi}^{(d)}(j) \hat{\eta}_{Yi}^{(d)}(k) \right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d} \right)^2 \sum_{j=2}^{d} \sum_{k=2}^{d} \left(\hat{\eta}_{Xi}^{(d)}(j) \right)^2 \left(\hat{\eta}_{Yi}^{(d)}(k) \right)^2 \\
&\quad + 2 \sum_{1 \le i,i' \le n,\, i \ne i'} \frac{\hat{\lambda}_{XYi}^{(d)}}{d} \frac{\hat{\lambda}_{XYi'}^{(d)}}{d} \sum_{j=2}^{d} \sum_{k=2}^{d} \left(\hat{\eta}_{Xi}^{(d)}(j) \hat{\eta}_{Xi'}^{(d)}(j) \right) \left(\hat{\eta}_{Yi}^{(d)}(k) \hat{\eta}_{Yi'}^{(d)}(k) \right) \\
&= \sum_{i=1}^{n} \left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d} \right)^2 \sum_{j=2}^{d} \left(\hat{\eta}_{Xi}^{(d)}(j) \right)^2 \sum_{k=2}^{d} \left(\hat{\eta}_{Yi}^{(d)}(k) \right)^2 \\
&\quad + 2 \sum_{1 \le i,i' \le n,\, i \ne i'} \frac{\hat{\lambda}_{XYi}^{(d)}}{d} \frac{\hat{\lambda}_{XYi'}^{(d)}}{d} \sum_{j=2}^{d} \left(\hat{\eta}_{Xi}^{(d)}(j) \hat{\eta}_{Xi'}^{(d)}(j) \right) \sum_{k=2}^{d} \left(\hat{\eta}_{Yi}^{(d)}(k) \hat{\eta}_{Yi'}^{(d)}(k) \right) \\
&= \underbrace{\left(\frac{\hat{\lambda}_{XY1}^{(d)}}{d} \right)^2 \sum_{j=2}^{d} \left(\tilde{\eta}_{X1}^{(d)}(j) \right)^2 \sum_{k=2}^{d} \left(\tilde{\eta}_{Y1}^{(d)}(k) \right)^2}_{(1)}
\end{aligned}$$

$$+\underbrace{2\sum_{i=2}^{n}\frac{\hat{\lambda}_{XY1}^{(d)}}{d}\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\left\langle\tilde{\eta}_{X1}^{(d)},\tilde{\eta}_{Xi}^{(d)}\right\rangle\left\langle\tilde{\eta}_{Y1}^{(d)},\tilde{\eta}_{Yi}^{(d)}\right\rangle}_{(2)}+\underbrace{\sum_{i=2}^{n}\left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\right)^{2}\left\|\tilde{\eta}_{Xi}^{(d)}\right\|_2^2\left\|\tilde{\eta}_{Yi}^{(d)}\right\|_2^2}_{(3)}$$

$$+\underbrace{2\sum_{2\le i,i'\le n,\,i\neq i'}\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\frac{\hat{\lambda}_{XYi'}^{(d)}}{d}\left\langle\tilde{\eta}_{Xi}^{(d)},\tilde{\eta}_{Xi'}^{(d)}\right\rangle\left\langle\tilde{\eta}_{Yi}^{(d)},\tilde{\eta}_{Yi'}^{(d)}\right\rangle}_{(4)}.$$

In the proof of Lemma 6, we showed that $\|\tilde{\eta}_{Xi}^{(d)}\|_2^2$ and $\|\tilde{\eta}_{Yi}^{(d)}\|_2^2$, for $i=2,3,\ldots,n$, converge in probability to 1 and that $\langle\tilde{\eta}_{Xi}^{(d)},\tilde{\eta}_{Xi'}^{(d)}\rangle$ and $\langle\tilde{\eta}_{Yi}^{(d)},\tilde{\eta}_{Xi'}^{(d)}\rangle$, for $i,i'=2,3,\ldots,n$ and $i\neq i'$, converge in probability to 1 as $d\to\infty$. Since $\hat{\lambda}_{XYi}$, for $i=2,3,\ldots,n$, is of magnitude of $O_p(d)$ by Lemma 6, we have for (3) and (4),

$$\left|\sum_{i=2}^{n}\left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\right)^{2}\left\|\tilde{\eta}_{Xi}^{(d)}\right\|_2^2\left\|\tilde{\eta}_{Yi}^{(d)}\right\|_2^2-\sum_{i=2}^{n}\left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\right)^{2}\right|\xrightarrow[d\to\infty]{p}0,$$

$$2\sum_{2\le i,i'\le n,\,i\neq i'}\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\frac{\hat{\lambda}_{XYi'}^{(d)}}{d}\left\langle\tilde{\eta}_{Xi}^{(d)},\tilde{\eta}_{Xi'}^{(d)}\right\rangle\left\langle\tilde{\eta}_{Yi}^{(d)},\tilde{\eta}_{Yi'}^{(d)}\right\rangle\xrightarrow[d\to\infty]{p}0.$$

Since $\|\tilde{\eta}_{Xi}^{(d)}\|_2^2$ and $\|\tilde{\eta}_{Yi}^{(d)}\|_2^2$, for $i=2,3,\ldots,n$, converge in probability to 1, the probability that an infinite number of entries of $\tilde{\eta}_{Xi}^{(d)}$ and $\tilde{\eta}_{Yi}^{(d)}$ are of $(\asymp 1/d^a)$, for $0\le a<1/2$, is 0. Suppose that an infinite number of entries of $\tilde{\eta}_{Xi}^{(d)}$ are of $(\asymp 1/d^a)$, for $0\le a<1/2$. Then, the squared sum of its entries blows up,

$$\sum_{j=2}^{d}\left(\tilde{\eta}_{Xi}^{(d)}(j)\right)^2=\frac{1}{d^{2a-1}}\sum_{j=2}^{d}\frac{d^{2a}\tilde{\eta}_{Xi}^{(d)}(j)}{d}\xrightarrow[d\to\infty]{p}\infty,\ 0\le a<\frac{1}{2}.$$

Hence, only a finite number of entries of $\tilde{\eta}_{Xi}^{(d)}$ and $\tilde{\eta}_{Yi}^{(d)}$ are of magnitude of $(\asymp 1/d^a)$, for $0\le a<1/2$ and the rest of entries are of $O_P(1/\sqrt{d})$ as $d\to\infty$. Using this fact and Lemma 5,

$$\begin{aligned}\left\langle\tilde{\eta}_{X1}^{(d)},\tilde{\eta}_{Xi}^{(d)}\right\rangle&=\sum_{j=2}^{d}\tilde{\eta}_{X1}^{(d)}(j)\tilde{\eta}_{Xi}^{(d)}(j)\\&=\sum_{j\in I}\tilde{\eta}_{X1}^{(d)}(j)\tilde{\eta}_{Xi}^{(d)}(j)+\sum_{j\notin I}\tilde{\eta}_{X1}^{(d)}(j)\tilde{\eta}_{Xi}^{(d)}(j)\\&=\sum_{j\in I}\tilde{\eta}_{X1}^{(d)}(j)\tilde{\eta}_{Xi}^{(d)}(j)+\frac{1}{\sqrt{d^{\alpha-1}}}\sum_{j\notin I}\frac{\left(\sqrt{d^{\alpha}}\tilde{\eta}_{X1}^{(d)}(j)\right)\left(\sqrt{d}\tilde{\eta}_{Xi}^{(d)}(j)\right)}{d}\\&=O_P(1/\sqrt{d^{\alpha-1}}),\ i=2,3,\ldots,n,\end{aligned}$$

where I is an index set denoting the entries of $\tilde{\eta}_{Xi}^{(d)}$ of magnitude of $(\asymp 1/d^a)$, for $0\le a<1/2$. Similarly, the magnitude of $\left\langle\tilde{\eta}_{Y1}^{(d)},\tilde{\eta}_{Yi}^{(d)}\right\rangle$ is $O_P(1/\sqrt{d^{\alpha-1}})$, for $i=2,3,\ldots,n$. Hence, by Lemma 4

and 6,

$$
\begin{aligned}
(2) &= 2\sum_{i=2}^{n} \frac{\hat{\lambda}_{XY1}^{(d)}}{d} \frac{\hat{\lambda}_{XYi}^{(d)}}{d} \left\langle \tilde{\eta}_{X1}^{(d)}, \tilde{\eta}_{Xi}^{(d)} \right\rangle \left\langle \tilde{\eta}_{Y1}^{(d)}, \tilde{\eta}_{Yi}^{(d)} \right\rangle \\
&= 2\sum_{i=2}^{n} \frac{\hat{\lambda}_{XY1}^{(d)}}{d^{\alpha}} \frac{\hat{\lambda}_{XYi}^{(d)}}{d} \left(\sqrt{d^{\alpha-1}} \left\langle \tilde{\eta}_{X1}^{(d)}, \tilde{\eta}_{Xi}^{(d)} \right\rangle \right) \left(\sqrt{d^{\alpha-1}} \left\langle \tilde{\eta}_{Y1}^{(d)}, \tilde{\eta}_{Yi}^{(d)} \right\rangle \right) \\
&= O_P(1).
\end{aligned}
$$

We prove that the term (2) above indeed converges to 0 as $d \to \infty$. With Lemma 5 and the fact that $\langle \hat{\eta}_{X1}^{(d)}, \hat{\eta}_{Xi}^{(d)} \rangle = 0$ and $\|\tilde{\eta}_{Xi}^{(d)}\|$ converges in probability to 0, for $i = 2, 3, \ldots, n$, the Cauchy-Schwarz inequality implies,

$$
\begin{aligned}
\left(\hat{\eta}_{X1}^{(d)}(1) \hat{\eta}_{Xi}^{(d)}(1) \right)^2 &= \left(-\sum_{j=2}^{d} \hat{\eta}_{X1}^{(d)}(j) \hat{\eta}_{Xi}^{(d)}(j) \right)^2 \\
&= \left\langle \tilde{\eta}_{X1}^{(d)}, \tilde{\eta}_{Xi}^{(d)} \right\rangle^2 \\
&\leq \left\| \tilde{\eta}_{X1}^{(d)} \right\|_2^2 \left\| \tilde{\eta}_{Xi}^{(d)} \right\|_2^2 \\
&= \sum_{j=2}^{d} \left(\hat{\eta}_{X1}^{(d)}(j) \right)^2 \left\| \tilde{\eta}_{Xi}^{(d)} \right\|_2^2 \\
&= \frac{1}{d^{\alpha-1}} \sum_{j=2}^{d} \frac{\left(\sqrt{d^{\alpha}} \hat{\eta}_{X1}^{(d)}(j) \right)^2}{d} \left\| \tilde{\eta}_{Xi}^{(d)} \right\|_2^2 \\
&= O_P(\frac{1}{d^{\alpha-1}}).
\end{aligned}
$$

Since $\hat{\eta}_{X1}^{(d)}(1)$ converges in probability to 1 by Lemma 4, we have,

$$
\hat{\eta}_{Xi}^{(d)}(1) = O_P(\frac{1}{\sqrt{d^{\alpha-1}}}), \ i = 2, 3, \ldots, n. \tag{3.26}
$$

Similarly,

$$
\hat{\eta}_{Yi}^{(d)}(1) = O_P(\frac{1}{\sqrt{d^{\alpha-1}}}), \ i = 2, 3, \ldots, n.
$$

Consider $\hat{\mathbf{\Sigma}}_{XY}^{(d)}(3,1)$, an entry in the third row and first column of the sample cross-covariance matrix $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$ given in (3.22). Then, we have,

$$
\begin{aligned}
\frac{\hat{\mathbf{\Sigma}}_{XY}^{(d)}(3,1)}{\sqrt{d^{\alpha}}} &= \frac{\tau_X \langle z_{X3\bullet}, d_1 \rangle}{n\sqrt{d^{\alpha}}} \\
&= \frac{1}{\sqrt{d^{\alpha}}} \sum_{i=1}^{n} \hat{\lambda}_{XYi}^{(d)} \hat{\eta}_{Xi}^{(d)}(3) \hat{\eta}_{Yi}^{(d)}(1)
\end{aligned}
$$

$$= \frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}} \left(\sqrt{d^{\alpha}} \hat{\eta}^{(d)}_{X1}(3) \right) \hat{\eta}^{(d)}_{Y1}(1) + \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{\sqrt{d^{2\alpha-1}}} \hat{\eta}^{(d)}_{Xi}(3) \left(\sqrt{d^{\alpha-1}} \hat{\eta}^{(d)}_{Yi}(1) \right),$$

which implies that,

$$\sqrt{d^{\alpha}} \hat{\eta}^{(d)}_{X1}(3) = \frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1} \hat{\eta}^{(d)}_{Y1}(1)} \left(\frac{\tau_X \langle z_{X3\bullet}, d_1 \rangle}{n\sqrt{d^{\alpha}}} - \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{\sqrt{d^{2\alpha-1}}} \hat{\eta}^{(d)}_{Xi}(3) \left(\sqrt{d^{\alpha-1}} \hat{\eta}^{(d)}_{Yi}(1) \right) \right).$$

In a similar argument as above, it can be shown that,

$$\sqrt{d^{\alpha}} \hat{\eta}^{(d)}_{X1}(j) = \frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1} \hat{\eta}^{(d)}_{Y1}(1)} \left(\frac{\tau_X \langle z_{Xj\bullet}, d_1 \rangle}{n\sqrt{d^{\alpha}}} - \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{\sqrt{d^{2\alpha-1}}} \hat{\eta}^{(d)}_{Xi}(j) \left(\sqrt{d^{\alpha-1}} \hat{\eta}^{(d)}_{Yi}(1) \right) \right), \ j = 4, 5, \ldots, d,$$

$$\sqrt{d^{\alpha}} \hat{\eta}^{(d)}_{Y1}(k) = \frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1} \hat{\eta}^{(d)}_{X1}(1)} \left(\frac{\tau_Y \langle c_1, z_{Yk\bullet} \rangle}{n\sqrt{d^{\alpha}}} - \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{\sqrt{d^{2\alpha-1}}} \left(\sqrt{d^{\alpha-1}} \hat{\eta}^{(d)}_{Xi}(1) \right) \hat{\eta}^{(d)}_{Yi}(k) \right), \ k = 3, 4, \ldots, d.$$

Then, the term (2) goes as follows,

$$\begin{aligned}
(2) &= 2 \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XY1}}{d} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \left\langle \tilde{\eta}^{(d)}_{X1}, \tilde{\eta}^{(d)}_{Xi} \right\rangle \left\langle \tilde{\eta}^{(d)}_{Y1}, \tilde{\eta}^{(d)}_{Yi} \right\rangle \\
&= 2 \frac{\hat{\lambda}^{(d)}_{XY1}}{d} \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \left(\sum_{j=2}^{d} \hat{\eta}^{(d)}_{X1}(j) \hat{\eta}^{(d)}_{Xi}(j) \right) \left(\sum_{k=2}^{d} \hat{\eta}^{(d)}_{Y1}(k) \hat{\eta}^{(d)}_{Yi}(k) \right) \\
&= 2 \frac{\hat{\lambda}^{(d)}_{XY1}}{d} \hat{\eta}^{(d)}_{X1}(2) \hat{\eta}^{(d)}_{Y1}(2) \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \hat{\eta}^{(d)}_{Xi}(2) \hat{\eta}^{(d)}_{Yi}(2) \\
&\quad + 2 \frac{\hat{\lambda}^{(d)}_{XY1}}{d} \hat{\eta}^{(d)}_{X1}(2) \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \hat{\eta}^{(d)}_{Xi}(2) \left(\sum_{k=3}^{d} \hat{\eta}^{(d)}_{Y1}(k) \hat{\eta}^{(d)}_{Yi}(k) \right) \\
&\quad + 2 \frac{\hat{\lambda}^{(d)}_{XY1}}{d} \hat{\eta}^{(d)}_{Y1}(2) \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \hat{\eta}^{(d)}_{Yi}(2) \left(\sum_{k=3}^{d} \hat{\eta}^{(d)}_{X1}(k) \hat{\eta}^{(d)}_{Xi}(k) \right) \\
&\quad + 2 \frac{\hat{\lambda}^{(d)}_{XY1}}{d} \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \left(\sum_{j=3}^{d} \hat{\eta}^{(d)}_{X1}(j) \hat{\eta}^{(d)}_{Xi}(j) \right) \left(\sum_{k=3}^{d} \hat{\eta}^{(d)}_{Y1}(k) \hat{\eta}^{(d)}_{Yi}(k) \right).
\end{aligned}$$

Using Lemma 4, 5 and the fact that only a finite number of entries of $\hat{\eta}^{(d)}_{Xi}$ and $\hat{\eta}^{(d)}_{Yi}$, for $i = 2, 3, \ldots, n$, can be of magnitude ($\asymp 1/d^{a}$), for $0 \leq a < 1/2$, the first three terms in the last part of the above equalities becomes,

$$\begin{aligned}
&2 \frac{\hat{\lambda}^{(d)}_{XY1}}{d} \hat{\eta}^{(d)}_{X1}(2) \hat{\eta}^{(d)}_{Y1}(2) \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \hat{\eta}^{(d)}_{Xi}(2) \hat{\eta}^{(d)}_{Yi}(2) \\
&\quad = \frac{2}{d} \frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}} \left(\sqrt{d^{\alpha}} \hat{\eta}^{(d)}_{X1}(2) \right) \left(\sqrt{d^{\alpha}} \hat{\eta}^{(d)}_{Y1}(2) \right) \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \hat{\eta}^{(d)}_{Xi}(2) \hat{\eta}^{(d)}_{Yi}(2) \xrightarrow[d \to \infty]{p} 0,
\end{aligned}$$

$$2 \frac{\hat{\lambda}^{(d)}_{XY1}}{d} \hat{\eta}^{(d)}_{X1}(2) \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \hat{\eta}^{(d)}_{Xi}(2) \left(\sum_{k=3}^{d} \hat{\eta}^{(d)}_{Y1}(k) \hat{\eta}^{(d)}_{Yi}(k) \right)$$

$$= \frac{2}{d}\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}}\left(\sqrt{d^{\alpha}}\hat{\eta}^{(d)}_{X1}(2)\right)\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\left(\sqrt{d}\hat{\eta}^{(d)}_{Xi}(2)\right)\left(\sum_{k=3}^{d}\frac{\left(\sqrt{d^{\alpha}}\hat{\eta}^{(d)}_{Y1}(k)\right)\left(\sqrt{d}\hat{\eta}^{(d)}_{Yi}(k)\right)}{d}\right)\xrightarrow[d\to\infty]{p}0,$$

$$2\frac{\hat{\lambda}^{(d)}_{XY1}}{d}\hat{\eta}^{(d)}_{Y1}(2)\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\hat{\eta}^{(d)}_{Yi}(2)\left(\sum_{k=3}^{d}\hat{\eta}^{(d)}_{X1}(k)\hat{\eta}^{(d)}_{Xi}(k)\right)$$
$$= \frac{2}{d}\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}}\left(\sqrt{d^{\alpha}}\hat{\eta}^{(d)}_{Y1}(2)\right)\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\left(\sqrt{d}\hat{\eta}^{(d)}_{Yi}(2)\right)\left(\sum_{k=3}^{d}\frac{\left(\sqrt{d^{\alpha}}\hat{\eta}^{(d)}_{Y1}(k)\right)\left(\sqrt{d}\hat{\eta}^{(d)}_{Yi}(k)\right)}{d}\right)\xrightarrow[d\to\infty]{p}0.$$

The fourth term expands as,

$$2\frac{\hat{\lambda}^{(d)}_{XY1}}{d}\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\left(\sum_{j=3}^{d}\hat{\eta}^{(d)}_{X1}(j)\hat{\eta}^{(d)}_{Xi}(j)\right)\left(\sum_{k=3}^{d}\hat{\eta}^{(d)}_{Y1}(k)\hat{\eta}^{(d)}_{Yi}(k)\right)$$
$$= 2\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}}\frac{1}{d}\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\left(\sum_{j=3}^{d}\sqrt{d^{\alpha}}\hat{\eta}^{(d)}_{X1}(j)\hat{\eta}^{(d)}_{Xi}(j)\right)\left(\sum_{k=3}^{d}\sqrt{d^{\alpha}}\hat{\eta}^{(d)}_{Y1}(k)\hat{\eta}^{(d)}_{Yi}(k)\right)$$
$$= 2\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}}\frac{1}{d}\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}$$
$$\times\left(\sum_{j=3}^{d}\frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{Y1}(1)}\left(\frac{\tau_X\langle z_{Xj\bullet},d_1\rangle}{n\sqrt{d^{\alpha}}}-\sum_{l=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYl}}{\sqrt{d^{2\alpha-1}}}\hat{\eta}^{(d)}_{Xl}(j)\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Yl}(1)\right)\right)\hat{\eta}^{(d)}_{Xi}(j)\right)$$
$$\times\left(\sum_{k=3}^{d}\frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{X1}(1)}\left(\frac{\tau_Y\langle c_1,z_{Yk\bullet}\rangle}{n\sqrt{d^{\alpha}}}-\sum_{l=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYl}}{\sqrt{d^{2\alpha-1}}}\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Xl}(1)\right)\hat{\eta}^{(d)}_{Yl}(k)\right)\hat{\eta}^{(d)}_{Yi}(k)\right).$$

Using the the Cauchy-Schwarz inequality on $\langle\hat{\eta}^{(d)}_{Xi},\hat{\eta}^{(d)}_{Yj}\rangle$, for $i,j=2,3,\ldots,n$, the law of large number and the fact that $\alpha>1$ and only a finite number of entries of $\hat{\eta}^{(d)}_{Xi}$ and $\hat{\eta}^{(d)}_{Yi}$, for $i=2,3,\ldots,n$, can be of magnitude ($\asymp 1/d^{a}$), for $0\le a<1/2$,, it is not hard to see that,

$$2\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}}\frac{1}{d}\sum_{i=2}^{n}\left[\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\left(\sum_{j=3}^{d}\frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{Y1}(1)}\left(-\sum_{l=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYl}}{\sqrt{d^{2\alpha-1}}}\hat{\eta}^{(d)}_{Xl}(j)\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Yl}(1)\right)\right)\hat{\eta}^{(d)}_{Xi}(j)\right)\right.$$
$$\left.\times\left(\sum_{k=3}^{d}\frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{X1}(1)}\left(-\sum_{l=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYl}}{\sqrt{d^{2\alpha-1}}}\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Xl}(1)\right)\hat{\eta}^{(d)}_{Yl}(k)\right)\hat{\eta}^{(d)}_{Yi}(k)\right)\right]$$
$$= 2\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}}\frac{1}{d^{\alpha-1}}\left(\frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{Y1}(1)}\right)^{2}\sum_{i=2}^{n}\left[\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\left(\sum_{j=3}^{d}\left(-\sum_{l=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYl}}{d}\hat{\eta}^{(d)}_{Xl}(j)\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Yl}(1)\right)\right)\hat{\eta}^{(d)}_{Xi}(j)\right)\right.$$
$$\left.\times\left(\sum_{k=3}^{d}\left(-\sum_{l=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYl}}{d}\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Xl}(1)\right)\hat{\eta}^{(d)}_{Yl}(k)\right)\hat{\eta}^{(d)}_{Yi}(k)\right)\right]\xrightarrow[d\to\infty]{p}0,$$

$$2\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}}\frac{1}{d}\sum_{i=2}^{n}\left[\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\sum_{j=3}^{d}\left(\frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{Y1}(1)}\frac{\tau_X\langle z_{Xj\bullet},d_1\rangle}{n\sqrt{d^{\alpha}}}\hat{\eta}^{(d)}_{Xi}(j)\right)\right.$$

$$\times \left(\sum_{k=3}^{d} \frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{X1}(1)} \left(- \sum_{l=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYl}}{\sqrt{d^{2\alpha-1}}} \left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Xl}(1) \right) \hat{\eta}^{(d)}_{Yi}(k) \right) \hat{\eta}^{(d)}_{Yl}(k) \right) \Bigg]$$

$$= 2\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}} \frac{1}{\sqrt{d^{\alpha-1}}} \left(\frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{Y1}(1)} \right)^2 \sum_{i=2}^{n} \Bigg[\frac{\hat{\lambda}^{(d)}_{XYi}}{d} \sum_{j=3}^{d} \left(\frac{\tau_X \langle z_{Xj\bullet}, d_1 \rangle}{n\sqrt{d^{\alpha}}} \frac{\sqrt{d}\hat{\eta}^{(d)}_{Xi}(j)}{d} \right)$$

$$\times \left(\sum_{k=3}^{d} \left(- \sum_{l=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYl}}{d} \left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Xl}(1) \right) \hat{\eta}^{(d)}_{Yl}(k) \right) \hat{\eta}^{(d)}_{Yi}(k) \right) \Bigg] \xrightarrow[d\to\infty]{p} 0,$$

$$2\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}} \frac{1}{d} \sum_{i=2}^{n} \Bigg[\frac{\hat{\lambda}^{(d)}_{XYi}}{d} \sum_{j=3}^{d} \left(\frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{Y1}(1)} \frac{\tau_Y \langle c_1, z_{Yk\bullet} \rangle}{n\sqrt{d^{\alpha}}} \hat{\eta}^{(d)}_{Yi}(j) \right)$$

$$\times \left(\sum_{k=3}^{d} \frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{X1}(1)} \left(- \sum_{l=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYl}}{\sqrt{d^{2\alpha-1}}} \left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Xl}(1) \right) \hat{\eta}^{(d)}_{Yl}(k) \right) \hat{\eta}^{(d)}_{Yi}(k) \right) \Bigg] \xrightarrow[d\to\infty]{p} 0,$$

$$2\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}} \frac{1}{d} \sum_{i=2}^{n} \frac{\hat{\lambda}^{(d)}_{XYi}}{d} \left(\sum_{j=3}^{d} \frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{Y1}(1)} \left(\frac{\tau_X \langle z_{Xj\bullet}, d_1 \rangle}{n\sqrt{d^{\alpha}}} \right) \hat{\eta}^{(d)}_{Xi}(j) \right)$$

$$\times \left(\sum_{k=3}^{d} \frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{X1}(1)} \left(\frac{\tau_Y \langle c_1, z_{Yk\bullet} \rangle}{n\sqrt{d^{\alpha}}} \right) \hat{\eta}^{(d)}_{Yi}(k) \right)$$

$$= 2\frac{\hat{\lambda}^{(d)}_{XY1}}{d^{\alpha}} \left(\frac{d^{\alpha}}{\hat{\lambda}^{(d)}_{XY1}\hat{\eta}^{(d)}_{Y1}(1)} \right)^2 \sum_{i=2}^{n} \Bigg[\frac{\hat{\lambda}^{(d)}_{XYi}}{d} \left(\sum_{j=3}^{d} \left(\frac{\tau_X \langle z_{Xj\bullet}, d_1 \rangle}{n\sqrt{d^{\alpha}}} \right) \frac{\sqrt{d}\hat{\eta}^{(d)}_{Xi}(j)}{d} \right)$$

$$\times \left(\sum_{k=3}^{d} \left(\frac{\tau_Y \langle c_1, z_{Yk\bullet} \rangle}{n\sqrt{d^{\alpha}}} \right) \frac{\sqrt{d}\hat{\eta}^{(d)}_{Yi}(k)}{d} \right) \Bigg] \xrightarrow[d\to\infty]{p} 0.$$

We now prove that the term (1) converges in probability to $\tau_X^2\tau_Y^2$ as $d \to \infty$,

$$(1) = \left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d} \right)^2 \sum_{j=2}^{d} \left(\hat{\eta}^{(d)}_{X1}(j) \right)^2 \sum_{k=2}^{d} \left(\hat{\eta}^{(d)}_{Y1}(k) \right)^2 \xrightarrow[d\to\infty]{p} \tau_X^2\tau_Y^2.$$

which implies that $\hat{\lambda}^{(d)}_{XYi}/d$, for $i = 2, 3, \ldots, n$, are squeezed, by the condition of (3.25), to converge in probability to 0 as $d \to \infty$. The term (1) can be written as,

$$\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d} \right)^2 \sum_{j=2}^{d} \left(\hat{\eta}^{(d)}_{X1}(j) \right)^2 \sum_{k=2}^{d} \left(\hat{\eta}^{(d)}_{Y1}(k) \right)^2$$

$$= \left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d} \right)^2 \left(\hat{\eta}^{(d)}_{X1}(2) \right)^2 \left(\hat{\eta}^{(d)}_{Y1}(2) \right)^2 + \left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d} \right)^2 \left(\hat{\eta}^{(d)}_{X1}(2) \right)^2 \sum_{k=3}^{d} \left(\hat{\eta}^{(d)}_{Y1}(k) \right)^2$$

$$+ \left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d} \right)^2 \left(\hat{\eta}^{(d)}_{Y1}(2) \right)^2 \sum_{j=3}^{d} \left(\hat{\eta}^{(d)}_{X1}(j) \right)^2 + \left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d} \right)^2 \sum_{j=3}^{d} \left(\hat{\eta}^{(d)}_{X1}(j) \right)^2 \sum_{k=3}^{d} \left(\hat{\eta}^{(d)}_{Y1}(k) \right)^2.$$

By the use of Lemma 4 and 5,

$$\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d}\right)^2\left(\hat{\eta}^{(d)}_{X1}(2)\right)^2\left(\hat{\eta}^{(d)}_{Y1}(2)\right)^2=\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d^\alpha}\right)^2\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{X1}(2)\right)^2\frac{1}{d^2}\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{Y1}(2)\right)^2\xrightarrow[d\to\infty]{p}0,$$

$$\begin{aligned}\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d}\right)^2\left(\hat{\eta}^{(d)}_{X1}(2)\right)^2\sum_{k=3}^{d}\left(\hat{\eta}^{(d)}_{Y1}(k)\right)^2&=\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d^\alpha}\right)^2\frac{\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{X1}(2)\right)^2}{d}\sum_{k=3}^{n}\frac{\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{Y1}(k)\right)^2}{d}\\&=\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d^\alpha}\right)^2\frac{1}{d}\frac{\sum_{k=3}^{n}\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{X1}(2)\right)^2\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{Y1}(k)\right)^2}{d}\xrightarrow[d\to\infty]{p}0,\end{aligned}$$

$$\begin{aligned}\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d}\right)^2\left(\hat{\eta}^{(d)}_{Y1}(2)\right)^2\sum_{k=3}^{d}\left(\hat{\eta}^{(d)}_{X1}(k)\right)^2&=\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d^\alpha}\right)^2\frac{\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{Y1}(2)\right)^2}{d}\sum_{k=3}^{n}\frac{\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{X1}(k)\right)^2}{d}\\&=\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d^\alpha}\right)^2\frac{1}{d}\frac{\sum_{k=3}^{n}\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{Y1}(2)\right)^2\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{X1}(k)\right)^2}{d}\xrightarrow[d\to\infty]{p}0.\end{aligned}$$

Note that the magnitudes of $\tau_X\langle z_{X3\bullet},d_1\rangle/n\sqrt{d^\alpha}$ and $\tau_Y\langle c_1,z_{Y3\bullet}\rangle/n\sqrt{d^\alpha}$ are of $(\asymp 1)$ from (3.20). Using Lemma 4, 5 and (3.26),

$$\begin{aligned}&\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d}\right)^2\sum_{j=3}^{d}\left(\hat{\eta}^{(d)}_{X1}(j)\right)^2\sum_{k=3}^{d}\left(\hat{\eta}^{(d)}_{Y1}(k)\right)^2\\&=\left(\frac{\hat{\lambda}^{(d)}_{XY1}}{d^\alpha}\right)^2\frac{1}{d^2}\sum_{j=3}^{d}\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{X1}(j)\right)^2\sum_{k=3}^{d}\left(\sqrt{d^\alpha}\hat{\eta}^{(d)}_{Y1}(k)\right)^2\\&=\left(\frac{d^\alpha}{\hat{\lambda}^{(d)}_{XY1}}\right)^2\frac{1}{d^2}\sum_{j=3}^{d}\left(\frac{\tau_X\langle z_{Xj\bullet},d_1\rangle}{n\sqrt{d^\alpha}}-\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{\sqrt{d^{2\alpha-1}}}\hat{\eta}^{(d)}_{Xi}(j)\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Yi}(1)\right)\right)^2\\&\quad\times\sum_{k=3}^{d}\left(\frac{\tau_Y\langle c_1,z_{Yk\bullet}\rangle}{n\sqrt{d^\alpha}}-\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{\sqrt{d^{2\alpha-1}}}\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Xi}(1)\right)\hat{\eta}^{(d)}_{Yi}(k)\right)^2\\&=\left(\frac{d^\alpha}{\hat{\lambda}^{(d)}_{XY1}}\right)^2\frac{1}{d^2}\sum_{j=3}^{d}\left(\frac{\tau_X\langle z_{Xj\bullet},d_1\rangle}{n\sqrt{d^\alpha}}\right)^2\sum_{k=3}^{d}\left(\frac{\tau_Y\langle c_1,z_{Yk\bullet}\rangle}{n\sqrt{d^\alpha}}\right)^2\\&\quad-4\left(\frac{d^\alpha}{\hat{\lambda}^{(d)}_{XY1}}\right)^2\frac{1}{d^{2\alpha-1}}\left(\sum_{j=3}^{d}\frac{\tau_X\langle z_{Xj\bullet},d_1\rangle}{n\sqrt{d^\alpha}}\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\hat{\eta}^{(d)}_{Xi}(j)\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Yi}(1)\right)\right)\\&\quad\times\left(\sum_{k=3}^{d}\frac{\tau_Y\langle c_1,z_{Yk\bullet}\rangle}{n\sqrt{d^\alpha}}\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Xi}(1)\right)\hat{\eta}^{(d)}_{Yi}(k)\right)\\&\quad+\left(\frac{d^\alpha}{\hat{\lambda}^{(d)}_{XY1}}\right)^2\frac{1}{d^{4(\alpha-1)}}\sum_{j=3}^{d}\left(\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\hat{\eta}^{(d)}_{Xi}(j)\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Yi}(1)\right)\right)^2\\&\quad\times\sum_{k=3}^{d}\left(\sum_{i=2}^{n}\frac{\hat{\lambda}^{(d)}_{XYi}}{d}\left(\sqrt{d^{\alpha-1}}\hat{\eta}^{(d)}_{Xi}(1)\right)\hat{\eta}^{(d)}_{Yi}(k)\right)^2.\end{aligned}$$

The second term in the last equality above converges in probability to 0. Using the fact that $\alpha > 1$ and only a finite number of entries of $\hat{\eta}_{Xi}^{(d)}$ and $\hat{\eta}_{Yi}^{(d)}$, for $i = 2, 3, \ldots, n$, can be of magnitude ($\asymp 1/d^a$), for $0 \leq a < 1/2$,

$$\frac{1}{d^{2\alpha-2}} \left(\sum_{i=2}^{n} \frac{\hat{\lambda}_{XYi}^{(d)}}{d} \left(\sqrt{d^{\alpha-1}} \hat{\eta}_{Yi}^{(d)}(1) \right) \sum_{j=3}^{d} \frac{\tau_X \langle z_{Xj\bullet}, d_1 \rangle}{n\sqrt{d^\alpha}} \frac{\sqrt{d}\hat{\eta}_{Xi}^{(d)}(j)}{d} \right)$$
$$\times \left(\sum_{i=2}^{n} \frac{\hat{\lambda}_{XYi}^{(d)}}{d} \left(\sqrt{d^{\alpha-1}} \hat{\eta}_{Xi}^{(d)}(1) \right) \sum_{k=3}^{d} \frac{\tau_Y \langle c_1, z_{Yk\bullet} \rangle}{n\sqrt{d^\alpha}} \frac{\sqrt{d}\hat{\eta}_{Yi}^{(d)}(k)}{d} \right) \xrightarrow[d\to\infty]{p} 0.$$

The third term also converges in probability to 0. It easy to see the result with the following equivalent form, noting that each inner product is bounded by 1 by the Cauchy-Schwarz inequality and that $\alpha > 1$,

$$\frac{1}{d^{4(\alpha-1)}} \left(\sum_{i,j=2}^{n} \left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d} \right)^2 \left(\frac{\hat{\lambda}_{XYj}^{(d)}}{d} \right)^2 \left(\sqrt{d^{\alpha-1}} \hat{\eta}_{Yi}^{(d)}(1) \right)^2 \left(\sqrt{d^{\alpha-1}} \hat{\eta}_{Xj}^{(d)}(1) \right)^2 \sum_{k=3}^{d} \left(\hat{\eta}_{Xi}^{(d)}(k) \right)^2 \left(\hat{\eta}_{Yj}^{(d)}(k) \right)^2 \right)$$
$$+ \frac{4}{d^{4(\alpha-1)}} \left(\sum_{2\leq a,a'\leq n, a\neq a'} \sum_{2\leq b,b'\leq n, b\neq b'} \frac{\hat{\lambda}_{XYa}^{(d)}}{d} \frac{\hat{\lambda}_{XYa'}^{(d)}}{d} \frac{\hat{\lambda}_{XYb}^{(d)}}{d} \frac{\hat{\lambda}_{XYb'}^{(d)}}{d} \sqrt{d^{\alpha-1}} \hat{\eta}_{Ya}^{(d)}(1) \sqrt{d^{\alpha-1}} \hat{\eta}_{Ya'}^{(d)}(1) \right.$$
$$\left. \times \sqrt{d^{\alpha-1}} \hat{\eta}_{Xb}^{(d)}(1) \sqrt{d^{\alpha-1}} \hat{\eta}_{Xb'}^{(d)}(1) \sum_{k=3}^{d} \left\langle \hat{\eta}_{Xa}^{(d)}(k), \hat{\eta}_{Xa'}^{(d)}(k) \right\rangle \left\langle \hat{\eta}_{Yb}^{(d)}(k), \hat{\eta}_{Yb'}^{(d)}(k) \right\rangle \right) \xrightarrow[d\to\infty]{p} 0.$$

Now, look at the first term. Denote by $d_1(i)$ and $c_1(i)$ the ith entries of the vectors d_1 and c_1 given in (3.20). Then, recalling that $z_{Xij}^2 \sim \chi_1^2$ (similarly z_{Yij}^2) and that z_{Xki} and z_{Xkj} are independent (similarly for z_{Yki} and z_{Ykj}) and using the limiting quantity of the first sample singular value in Lemma 4,

$$\tau_X^2 \tau_Y^2 \left(\frac{d^\alpha}{\hat{\lambda}_{XY1}^{(d)}} \right)^2 \frac{1}{d^2} \sum_{j=3}^{d} \left(\frac{\langle z_{Xj\bullet}, d_1 \rangle}{n\sqrt{d^\alpha}} \right)^2 \sum_{k=3}^{d} \left(\frac{\langle c_1, z_{Yk\bullet} \rangle}{n\sqrt{d^\alpha}} \right)^2$$
$$= \tau_X^2 \tau_Y^2 \frac{1}{n^2} \left(\frac{d^\alpha}{\hat{\lambda}_{XY1}^{(d)}} \right)^2 \left(\sum_{i=1}^{n} \frac{d_1^2(i)}{d^\alpha} \sum_{j=3}^{d} \frac{z_{Xji}^2}{d} + 2 \sum_{1\leq i,j\leq n, i\neq j} \sum_{k=1}^{d} \frac{d_1(i)d_1(j)}{d^\alpha} \frac{z_{Xki} z_{Xkj}}{d} \right)$$
$$\times \left(\sum_{i=1}^{n} \frac{c_1^2(i)}{d^\alpha} \sum_{j=3}^{d} \frac{z_{Yji}^2}{d} + 2 \sum_{1\leq i,j\leq n, i\neq j} \sum_{k=1}^{d} \frac{c_1(i)c_1(j)}{d^\alpha} \frac{z_{Yki} z_{Ykj}}{d} \right)$$
$$\xrightarrow[d\to\infty]{p} \tau_X^2 \tau_Y^2 \frac{1}{n^2} \frac{n^2 \|c_1\|_2^2 \|d_1\|_2^2}{\langle c_1, d_1 \rangle^2}$$
$$\geq \tau_X^2 \tau_Y^2,$$

where the last inequality results from the Cauchy-Schwarz inequality of $\langle c_1, d_1 \rangle^2 \leq \|c_1\|_2^2 \|d_1\|_2^2$. Then, the condition of (3.25) completes the proof. □

3.4.3.2 Behavior of the sample covariance matrices We investigate the HDLSS asymptotic behavior of the sample covariance matrices $\hat{\mathbf{\Sigma}}_X^{(d)}$ and $\hat{\mathbf{\Sigma}}_Y^{(d)}$ given in (3.22), in specific, its sample eigenvalues and eigenvectors. Here, we only include the result of the analysis of $\hat{\mathbf{\Sigma}}_X^{(d)}$ as that of $\hat{\mathbf{\Sigma}}_Y^{(d)}$ is similar. The eigendecomposition of $\hat{\mathbf{\Sigma}}_X^{(d)}$ gives,

$$\hat{\mathbf{\Sigma}}_X^{(d)} = \sum_{i=1}^{n} \hat{\lambda}_{Xi}^{(d)} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T, \tag{3.27}$$

where $\hat{\lambda}_{X1}^{(d)} \geq \hat{\lambda}_{X2}^{(d)} \geq \cdots \geq \hat{\lambda}_{Xn}^{(d)} \geq 0$, $\|\hat{\xi}_{Xi}^{(d)}\|_2 = 1$ and $\langle \hat{\xi}_{Xi}^{(d)}, \hat{\xi}_{Xj}^{(d)} \rangle = 0$ for $i \neq j$.

Lemma 8 (CCA HDLSS Asymptotic Lemma 7.)**.** *Let C_X and $\mathbf{M}_X$ be,*

$$C_X = \lim_{d\to\infty} \frac{\|c_1\|_2^2}{d^\alpha}, \ \mathbf{M}_X = \begin{bmatrix} C_X & \mathbf{0}_{1\times(d-1)} \\ \mathbf{0}_{(d-1)\times 1} & \mathbf{0}_{(d-1)\times(d-1)} \end{bmatrix}.$$

where c_1 is defined in (3.20) and so C_X is a non-degenerate random variable. Then, for $\alpha > 1$,

$$\left\| \frac{n\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \mathbf{M}_X \right\|_F^2 \xrightarrow[d\to\infty]{p} 0,$$

Proof. Let $\mathbf{M}_X^{(d)}$ be,

$$\mathbf{M}_X^{(d)} = \begin{bmatrix} \frac{\|c_1\|_2^2}{d^\alpha} & \mathbf{0}_{1\times(d-1)} \\ \mathbf{0}_{(d-1)\times 1} & \mathbf{0}_{(d-1)\times(d-1)} \end{bmatrix}.$$

It is obvious to see that,

$$\left\| \mathbf{M}_X^{(d)} - \mathbf{M}_X \right\|_F^2 \xrightarrow[d\to\infty]{p} 0.$$

Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left\| \frac{n\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \mathbf{M}_X^{(d)} \right\|_F^2 =& \frac{\langle c_1, c_2\rangle^2}{d^{2\alpha}} + \frac{\langle c_2, c_1\rangle^2}{d^{2\alpha}} + \frac{\langle c_2, c_2\rangle^2}{d^{2\alpha}} + 2\tau_X^2 \sum_{i=3}^{d} \frac{\langle c_1, z_{Xi\bullet}\rangle^2}{d^{2\alpha}} \\
&+ 2\tau_X^2 \sum_{i=3}^{d} \frac{\langle c_2, z_{Xi\bullet}\rangle^2}{d^{2\alpha}} + \tau_X^4 \sum_{i=3}^{d}\sum_{j=3}^{d} \frac{\langle z_{Xi\bullet}, z_{Xj\bullet}\rangle^2}{d^{2\alpha}} \\
\leq& \frac{\langle c_1, c_2\rangle^2}{d^{2\alpha}} + \frac{\langle c_2, c_1\rangle^2}{d^{2\alpha}} + \frac{\langle c_2, c_2\rangle^2}{d^{2\alpha}} + 2\tau_X^2 \sum_{i=3}^{d} \frac{\|c_1\|_2^2 \|z_{Xi\bullet}\|_2^2}{d^{2\alpha}} \\
&+ 2\tau_X^2 \sum_{i=3}^{d} \frac{\|c_2\|_2^2 \|z_{Xi\bullet}\|_2^2}{d^{2\alpha}} + \tau_X^4 \sum_{i=3}^{d}\sum_{j=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2 \|z_{Xj\bullet}\|_2^2}{d^{2\alpha}}.
\end{aligned}$$

Since $\langle c_1, c_2\rangle^2$, $\langle c_2, c_1\rangle^2$ are $O_P(\sqrt{d^\alpha})$, and $\langle c_2, c_2\rangle^2$ is $O_P(1)$,

$$\frac{\langle c_1, c_2\rangle^2}{d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0, \quad \frac{\langle c_2, c_1\rangle^2}{d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0, \quad \frac{\langle c_2, c_2\rangle^2}{d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0.$$

Note that $\|z_{Xi\bullet}\|_2^2 \sim \chi_n^2$ and that $\|c_1\|_2^2$ and $\|c_2\|_2^2$ are of $O_P(d^\alpha)$. By the law of large numbers,

$$\begin{aligned} 2\tau_X^2 \sum_{i=3}^{d} \frac{\|c_1\|_2^2 \|z_{Xi\bullet}\|_2^2}{d^{2\alpha}} &= \frac{\tau_X^2}{d^{\alpha-1}} \left\| \frac{d_1}{\sqrt{d^\alpha}} \right\|_2^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2}{d} \\ &\xrightarrow[d\to\infty]{p} 0 \times O_P(1) \times E(\chi_n^2) = 0, \\ 2\tau_X^2 \sum_{i=3}^{d} \frac{\|c_2\|_2^2 \|z_{Xi\bullet}\|_2^2}{d^{2\alpha}} &= \frac{\tau_X^2}{d^{\alpha-1}} \left\| \frac{d_2}{\sqrt{d^\alpha}} \right\|_2^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2}{d} \\ &\xrightarrow[d\to\infty]{p} 0 \times 0 \times E(\chi_n^2) = 0. \end{aligned}$$

Using the fact that $\alpha > 1$, $\|z_{Xi\bullet}\|_2^2 \sim \chi_n^2$ and $\|z_{Yi\bullet}\|_2^2 \sim \chi_n^2$ and applying the law of large numbers,

$$\begin{aligned} \tau_X^4 \sum_{i=3}^{d} \sum_{j=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2 \|z_{Xj\bullet}\|_2^2}{d^{2\alpha}} &= \frac{\tau_X^4}{d^{2\alpha-2}} \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2}{d} \sum_{j=3}^{d} \frac{\|z_{Xj\bullet}\|_2^2}{d} \\ &\xrightarrow[d\to\infty]{p} 0 \times n \times n = 0. \end{aligned}$$

Therefore,

$$\left\| \frac{n\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \mathbf{M}_X^{(d)} \right\|_F^2 \xrightarrow[d\to\infty]{p} 0.$$

□

Lemma 9 (CCA HDLSS Asymptotic Lemma 8.)**.** *Let $\mathbf{M}_X$ be the matrix defined in Lemma 8. Let $\hat{\lambda}_{X1}^{(d)}$ and $\hat{\xi}_{X1}^{(d)}$ be the first sample eigenvalue and eigenvectors from (3.27). Then, for $\alpha > 1$,*

$$\left\| \frac{n\hat{\lambda}_{X1}^{(d)}}{d^\alpha} \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T - \mathbf{M}_X \right\|_F^2 \xrightarrow[d\to\infty]{p} 0.$$

Proof. Using $\mathbf{M}_X^{(d)}$ defined in Lemma 8 and the triangle inequality,

$$\begin{aligned} &\left\| \frac{n\hat{\lambda}_{X1}^{(d)}}{d^\alpha} \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T - \mathbf{M}_X^{(d)} \right\|_F \\ &\quad = \left\| \left(\frac{n\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \mathbf{M}_X^{(d)} \right) - \left(\frac{n\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \frac{n\hat{\lambda}_{X1}^{(d)}}{d^\alpha} \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T \right) \right\|_F \\ &\quad \le \left\| \frac{n\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \mathbf{M}_X^{(d)} \right\|_F + \left\| \frac{n\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \frac{n\hat{\lambda}_{X1}^{(d)}}{d^\alpha} \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T \right\|_F. \end{aligned}$$

By Lemma 8,

$$\left\| \frac{n\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \mathbf{M}_X^{(d)} \right\|_F^2 \xrightarrow[d\to\infty]{p} 0.$$

Write $\mathbf{M}_X^{(d)}$ as,

$$\mathbf{M}_X^{(d)} = \frac{\|c_1\|_2^2}{d^\alpha} e_1^{(d)} \left(e_1^{(d)}\right)^T.$$

Since the first eigenvalue $\hat{\lambda}_{X1}^{(d)}$ and eigenvector $\hat{\xi}_{X1}^{(d)}$ provide the best rank-1 approximation to the matrix $\hat{\lambda}_{X1}^{(d)}/d^\alpha$,

$$\begin{aligned}
\left\| \frac{\hat{\lambda}_{X1}^{(d)}}{d^\alpha} - \frac{\hat{\lambda}_{X1}^{(d)}}{d^\alpha} \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T \right\|_F^2 &\leq \left\| \frac{\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \frac{\|c_1\|_2^2}{nd^\alpha} e_1^{(d)} \left(e_1^{(d)}\right)^T \right\|_F^2 \\
&\leq \left\| \frac{\hat{\mathbf{\Sigma}}_X^{(d)}}{d^\alpha} - \frac{\mathbf{M}_X^{(d)}}{n} \right\|_F^2 \\
&\xrightarrow[d\to\infty]{p} 0.
\end{aligned}$$

□

Lemma 10 (CCA HDLSS Asymptotic Lemma 9.)**.** *For $\alpha > 1$, the first eigenvalue $\hat{\lambda}_{X1}^{(d)}$ and eigenvectors $\hat{\xi}_{X1}^{(d)}$ converge in probability to the following quantities as $d \to \infty$,*

$$\frac{n\hat{\lambda}_{X1}^{(d)}}{d^\alpha} \xrightarrow[d\to\infty]{p} C_X, \ \left\| \hat{\xi}_{X1}^{(d)} - e_1^{(d)} \right\|_2 \xrightarrow[d\to\infty]{p} 0,$$

where C_X is defined in Lemma 8.

Proof. Let $\tilde{\xi}_{X1}^{(d)}$ be $\hat{\xi}_{X1}^{(d)}$ where their first entries are set to 0. Let $\hat{\xi}_{X1}^{(d)}(i)$ be the ith and jth entries of $\hat{\xi}_{X1}^{(d)}$. By Lemma 9,

$$\begin{aligned}
\left(\frac{\hat{\lambda}_{X1}^{(d)}}{d^\alpha}\right)^2 \left\|\tilde{\xi}_{X1}^{(d)}\right\|_2^2 \left\|\tilde{\xi}_{X1}^{(d)}\right\|_2^2 &= \left(\frac{\hat{\lambda}_{X1}^{(d)}}{d^\alpha}\right)^2 \sum_{i=2}^d \left(\hat{\xi}_{X1}^{(d)}(i)\right)^2 \sum_{j=2}^d \left(\hat{\xi}_{X1}^{(d)}(j)\right)^2 \\
&= \left(\frac{\hat{\lambda}_{X1}^{(d)}}{d^\alpha}\right)^2 \sum_{i=2}^d \sum_{j=2}^d \left(\hat{\xi}_{X1}^{(d)}(i)\right)^2 \left(\hat{\xi}_{X1}^{(d)}(j)\right)^2 \\
&= \left\| \frac{\hat{\lambda}_{X1}^{(d)}}{d^\alpha} \tilde{\xi}_{X1}^{(d)} \left(\tilde{\xi}_{X1}^{(d)}\right)^T \right\|_F^2 \\
&\xrightarrow[d\to\infty]{p} 0.
\end{aligned}$$

First, we show that $(\hat{\lambda}_{X1}^{(d)}/d^{\alpha})^2 > 0$ in probability. Suppose that $(\hat{\lambda}_{X1}^{(d)}/d^{\alpha})^2$ converges in probability to 0. Then, noting that $\|\hat{\xi}_{X1}^{(d)}\|_2^2 = 1$ and $\|\hat{\xi}_{X1}^{(d)}\|_2^2 = 1$,

$$\begin{aligned}
\left\| \frac{\hat{\lambda}_{X1}^{(d)}}{d^{\alpha}} \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T \right\|_F^2 &= \left(\frac{\hat{\lambda}_{X1}^{(d)}}{d^{\alpha}}\right)^2 \sum_{i=1}^{d}\sum_{j=1}^{d} \left(\hat{\xi}_{X1}^{(d)}(i)\right)^2 \left(\hat{\xi}_{X1}^{(d)}(j)\right)^2 \\
&= \left(\frac{\hat{\lambda}_{X1}^{(d)}}{d^{\alpha}}\right)^2 \sum_{i=1}^{d} \left(\hat{\xi}_{X1}^{(d)}\right)^2 \sum_{j=1}^{d} \left(\hat{\xi}_{X1}^{(d)}(j)\right)^2 \\
&= \left(\frac{\hat{\lambda}_{X1}^{(d)}}{d^{\alpha}}\right)^2 \left\|\hat{\xi}_{X1}^{(d)}\right\|_2^2 \left\|\hat{\xi}_{X1}^{(d)}\right\|_2^2 \\
&= \left(\frac{\hat{\lambda}_{X1}^{(d)}}{d^{\alpha}}\right)^2 \\
&\xrightarrow[d\to\infty]{p} 0,
\end{aligned}$$

which contradicts to Lemma 8 (note that the limiting matrix M is not a degenerate matrix). Therefore,

$$\left\|\tilde{\xi}_{X1}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{p} 0.$$

Note that, since the norm of $\hat{\xi}_{X1}^{(d)}$ is 1,

$$\left(\hat{\xi}_{X1}^{(d)}(1)\right)^2 = 1 - \left\|\tilde{\xi}_{X1}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{p} 1.$$

Therefore

$$\left\|\hat{\xi}_{X1}^{(d)} - e_1^{(d)}\right\|_2^2 = \left(\hat{\xi}_{X1}^{(d)}(1) - 1\right)^2 + \left\|\tilde{\xi}_{X1}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{p} 0.$$

To find the limiting value of $\hat{\lambda}_{X1}^{(d)}/d^{\alpha}$ as $d \to \infty$, using the unitary invariance property of the Frobenius norm and the previous result about the limiting vectors of $\hat{\xi}_{X1}^{(d)}$ as $d \to \infty$,

$$\begin{aligned}
&\left\| \frac{n\hat{\lambda}_{X1}^{(d)}}{d^{\alpha}} \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T - \mathbf{M}_X^{(d)} \right\|_F^2 \\
&= \left\| \frac{n\hat{\lambda}_{X1}^{(d)}}{d^{\alpha}} \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T - \frac{\|c_1\|_2^2}{d^{\alpha}} e_1^{(d)} \left(e_1^{(d)}\right)^T \right\|_F^2 \\
&= \left\| \frac{n\hat{\lambda}_{X1}^{(d)}}{d^{\alpha}} \left(\hat{\xi}_{X1}^{(d)}\right)^T \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T \hat{\xi}_{X1}^{(d)} - \frac{\|c_1\|_2^2}{d^{\alpha}} \left(\hat{\xi}_{X1}^{(d)}\right)^T e_1^{(d)} \left(e_1^{(d)}\right)^T \hat{\xi}_{X1}^{(d)} \right\|_F^2 \\
&= \left\| \frac{n\hat{\lambda}_{X1}^{(d)}}{d^{\alpha}} - \frac{\|c_1\|_2^2}{d^{\alpha}} \left(\hat{\xi}_{X1}^{(d)}\right)^T e_1^{(d)} \left(e_1^{(d)}\right)^T \hat{\xi}_{X1}^{(d)} \right\|_F^2
\end{aligned}$$

$$= \left(\frac{\|c_1\|_2^2}{d^\alpha}\right)^2 \left\| \frac{n\hat{\lambda}_{X1}^{(d)}/d^\alpha}{\|c_1\|_2^2/d^\alpha} - \left(\hat{\xi}_{X1}^{(d)}\right)^T e_1^{(d)} \left(e_1^{(d)}\right)^T \hat{\xi}_{X1}^{(d)} \right\|_F^2$$
$$\xrightarrow[d\to\infty]{p} \mathrm{A}^2 \left(\frac{\mathrm{B}}{\mathrm{C}} - 1\right)^2,$$

where,

$$\mathrm{A} = \lim_{d\to\infty} \frac{\|c_1\|_2^2}{d^\alpha},\ \mathrm{B} = \lim_{d\to\infty} \frac{n\hat{\lambda}_{X1}^{(d)}}{d^\alpha},\ \mathrm{C} = \lim_{d\to\infty} \frac{\|c_1\|_2^2}{d^\alpha}.$$

By Lemma 8,

$$\left\| \frac{n\hat{\lambda}_{X1}^{(d)}}{d^\alpha} \hat{\xi}_{X1}^{(d)} \left(\hat{\xi}_{X1}^{(d)}\right)^T - \mathbf{M}_X^{(d)} \right\|_F^2 \xrightarrow[d\to\infty]{p} 0,$$

and the fact that $\mathrm{A} \asymp O_p(1)$,

$$\left(\frac{\mathrm{B}}{\mathrm{C}} - 1\right)^2 \xrightarrow[d\to\infty]{p} 0.$$

Therefore,

$$\frac{\mathrm{B}}{\mathrm{C}} \xrightarrow[d\to\infty]{p} 1.$$

□

Lemma 11 (CCA HDLSS Asymptotic Lemma 10.)**.** *The sample eigenvalue* $\hat{\lambda}_{Xi}^{(d)}$ *and eigenvector* $\hat{\xi}_{Xi}^{(d)}$, $i = 2, 3, .., d$, *converge in probability to the following quantities as* $d \to \infty$,

$$\frac{n\hat{\lambda}_{Xi}^{(d)}}{d} \xrightarrow[d\to\infty]{P} \tau_X^2,\ \langle \hat{\xi}_{Xi}^{(d)}, \xi^{(d)} \rangle \xrightarrow[d\to\infty]{P} 0,\ i = 2, 3, \ldots, n,$$

where $\xi^{(d)}$ *is an any given vector in* R^d.

Proof. Recall that the underlying random variable $X^{(d)}$ of the sample covariance matrix $\hat{\mathbf{\Sigma}}_X^{(d)}$ has a simple spiked covariance structure of (3.9). Then, the asymptotic behavior of a sample eigenvalue is a direct consequence of Theorem 1 for $\alpha > 1$. The result on the behavior of $\langle \hat{\xi}_{Xi}^{(d)}, \xi^{(d)} \rangle$ is indicated in the proof of Theorem 1 given in [24], even though $\xi^{(d)}$ is fixed at the population counterpart of $\hat{\xi}_{Xi}^{(d)}$ in Theorem 1 for comparison purpose. Note that an inner product of two unit vectors measures a cosine of the angle between the two. Lemma 11 implies that the sample eigenvector $\hat{\xi}_{Xi}^{(d)}$ becomes completely random such that the probability of it being consistent with any given vector is 0 as $d \to \infty$. □

3.4.3.3 Behavior of the matrix $\hat{\mathbf{R}}^{(d)}$ The definition of the matrix $\hat{\mathbf{R}}^{(d)}$ is given in (3.3) with the inverse matrix we are going to use explained in (3.13). The sample canonical correlation coefficients are found as the singular values of $\hat{\mathbf{R}}^{(d)}$ and sample canonical weight vectors are obtained via unscaling and normalizing the singular vectors of $\hat{\mathbf{R}}^{(d)}$ shown in (3.4).

$$\begin{aligned}\hat{\mathbf{R}}^{(d)} &= \left(\hat{\mathbf{\Sigma}}_X^{(d)}\right)^{-\frac{1}{2}} \hat{\mathbf{\Sigma}}_{XY}^{(d)} \left(\hat{\mathbf{\Sigma}}_Y^{(d)}\right)^{-\frac{1}{2}} \\ &= \sum_{i=1}^{n} \sqrt{\frac{1}{\hat{\lambda}_{Xi}^{(d)}}} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T \sum_{i=1}^{n} \hat{\lambda}_{XYi}^{(d)} \hat{\eta}_{Xi}^{(d)} \left(\hat{\eta}_{Yi}^{(d)}\right)^T \sum_{i=1}^{n} \sqrt{\frac{1}{\hat{\lambda}_{Yi}^{(d)}}} \hat{\xi}_{Yi}^{(d)} \left(\hat{\xi}_{Yi}^{(d)}\right)^T\end{aligned} \tag{3.28}$$

Consider the two parts of the matrix $\hat{\mathbf{R}}^{(d)}$ in the following,

$$\begin{bmatrix} \frac{1}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\hat{\lambda}_{X2}^{(d)}}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\hat{\lambda}_{Xn}^{(d)}}} \end{bmatrix} \begin{bmatrix} \left(\hat{\xi}_{X1}^{(d)}\right)^T \\ \left(\hat{\xi}_{X2}^{(d)}\right)^T \\ \vdots \\ \left(\hat{\xi}_{Xn}^{(d)}\right)^T \end{bmatrix} \begin{bmatrix} \left(\hat{\eta}_{X1}^{(d)}\right)^T \\ \left(\hat{\eta}_{X2}^{(d)}\right)^T \\ \vdots \\ \left(\hat{\eta}_{Xn}^{(d)}\right)^T \end{bmatrix}^T \begin{bmatrix} \sqrt{\hat{\lambda}_{XY1}^{(d)}} & 0 & \dots & 0 \\ 0 & \sqrt{\hat{\lambda}_{XY2}^{(d)}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\hat{\lambda}_{XYn}^{(d)}} \end{bmatrix},$$

$$\begin{bmatrix} \sqrt{\hat{\lambda}_{XY1}^{(d)}} & 0 & \dots & 0 \\ 0 & \sqrt{\hat{\lambda}_{XY2}^{(d)}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\hat{\lambda}_{XYn}^{(d)}} \end{bmatrix} \begin{bmatrix} \left(\hat{\eta}_{X1}^{(d)}\right)^T \\ \left(\hat{\eta}_{X2}^{(d)}\right)^T \\ \vdots \\ \left(\hat{\eta}_{Xn}^{(d)}\right)^T \end{bmatrix}^T \begin{bmatrix} \left(\hat{\xi}_{Y1}^{(d)}\right)^T \\ \left(\hat{\xi}_{Y2}^{(d)}\right)^T \\ \vdots \\ \left(\hat{\xi}_{Yn}^{(d)}\right)^T \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{\hat{\lambda}_{Y1}^{(d)}}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\hat{\lambda}_{Y2}^{(d)}}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\hat{\lambda}_{Yn}^{(d)}}} \end{bmatrix}.$$

The first part reduces to the following matrix (call it $\hat{\mathbf{R}}_1^{(d)}$),

$$\hat{\mathbf{R}}_1^{(d)} = \begin{bmatrix} \frac{\sqrt{\hat{\lambda}_{XY1}^{(d)}}}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \langle \hat{\xi}_{X1}^{(d)}, \hat{\eta}_{X1}^{(d)} \rangle & \frac{\sqrt{\hat{\lambda}_{XY2}^{(d)}}}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \langle \hat{\xi}_{X1}^{(d)}, \hat{\eta}_{X2}^{(d)} \rangle & \dots & \frac{\sqrt{\hat{\lambda}_{XYn}^{(d)}}}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \langle \hat{\xi}_{X1}^{(d)}, \hat{\eta}_{Xn}^{(d)} \rangle \\ \frac{\sqrt{\hat{\lambda}_{XY1}^{(d)}}}{\sqrt{\hat{\lambda}_{X2}^{(d)}}} \langle \hat{\xi}_{X2}^{(d)}, \hat{\eta}_{X1}^{(d)} \rangle & \frac{\sqrt{\hat{\lambda}_{XY2}^{(d)}}}{\sqrt{\hat{\lambda}_{X2}^{(d)}}} \langle \hat{\xi}_{X2}^{(d)}, \hat{\eta}_{X2}^{(d)} \rangle & \dots & \frac{\sqrt{\hat{\lambda}_{XYn}^{(d)}}}{\sqrt{\hat{\lambda}_{X2}^{(d)}}} \langle \hat{\xi}_{X1}^{(d)}, \hat{\eta}_{Xn}^{(d)} \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sqrt{\hat{\lambda}_{XY1}^{(d)}}}{\sqrt{\hat{\lambda}_{Xn}^{(d)}}} \langle \hat{\xi}_{Xn}^{(d)}, \hat{\eta}_{X1}^{(d)} \rangle & \frac{\sqrt{\hat{\lambda}_{XY2}^{(d)}}}{\sqrt{\hat{\lambda}_{Xn}^{(d)}}} \langle \hat{\xi}_{Xn}^{(d)}, \hat{\eta}_{X2}^{(d)} \rangle & \dots & \frac{\sqrt{\hat{\lambda}_{XYn}^{(d)}}}{\sqrt{\hat{\lambda}_{Xn}^{(d)}}} \langle \hat{\xi}_{X1}^{(d)}, \hat{\eta}_{Xn}^{(d)} \rangle \end{bmatrix}. \tag{3.29}$$

We investigate which value each entry of $\hat{\mathbf{R}}_1^{(d)}(i,j)$ converges in probability to as $d \to \infty$. Referring to Lemma 4, 7, 10 and 11 on the magnitudes of the sample eigenvalues and singular values, it can be easily noticed that,

$$\frac{\sqrt{\hat{\lambda}_{XY1}^{(d)}}}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} = \frac{\sqrt{\langle c_1, d_1 \rangle / d^\alpha}}{\sqrt{\langle c_1, c_1 \rangle / d^\alpha}} \asymp O_P(1), \ \frac{\sqrt{\hat{\lambda}_{XYi}^{(d)}}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}} = o_P(1), \ i,j = 2,3,\dots,n,$$

$$\frac{\sqrt{\hat{\lambda}_{XYi}^{(d)}}}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} = O_P(1/\sqrt{d^{\alpha-1}}),\ \frac{\sqrt{\hat{\lambda}_{XY1}^{(d)}}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}} = O_P(\sqrt{d^{\alpha-1}}),\ i = 2, 3, \ldots, n.$$

Then, $\hat{\mathbf{R}}_1^{(d)}$ can be written as,

$$\left\| \hat{\mathbf{R}}_1^{(d)} - \begin{bmatrix} \frac{\sqrt{\langle c_1, d_1 \rangle}}{\|c_1\|_2} \langle \hat{\xi}_{X1}^{(d)}, \hat{\eta}_{X1}^{(d)} \rangle & \underset{1\times(d-1)}{\mathbf{0}} \\ \frac{\sqrt{\hat{\lambda}_{XY1}^{(d)}}}{\sqrt{\hat{\lambda}_{X2}^{(d)}}} \langle \hat{\xi}_{X2}^{(d)}, \hat{\eta}_{X1}^{(d)} \rangle & \underset{1\times(d-1)}{\mathbf{0}} \\ \vdots & \vdots \\ \frac{\sqrt{\hat{\lambda}_{XY1}^{(d)}}}{\sqrt{\hat{\lambda}_{Xn}^{(d)}}} \langle \hat{\xi}_{Xn}^{(d)}, \hat{\eta}_{X1}^{(d)} \rangle & \underset{1\times(d-1)}{\mathbf{0}} \end{bmatrix} \right\|_2 \xrightarrow[d\to\infty]{P} 0.$$

Since $(\sqrt{\hat{\lambda}_{XY1}^{(d)}}/\sqrt{\hat{\lambda}_{Xj}^{(d)}})\langle \hat{\xi}_{Xi}^{(d)}, \hat{\eta}_{X1}^{(d)} \rangle$, for $i = 2, 3, \ldots, n$, appears to blow up, we further investigate their magnitudes.

Lemma 12 (CCA HDLSS Asymptotic Lemma 11.)**.** *The inner product of the first sample singular vector $\hat{\eta}_{X1}^{(d)}$ and the ith eigenvector $\hat{\xi}_{Xi}^{(d)}$, $i = 2, 3, ..n$, converges to 0 with the speed of $o_P(1/\sqrt{d^{\alpha-1}})$ as $d \to \infty$.*

Proof. Consider $\mathbf{X}^{(d)}$ given in (3.16), which contains the observations of the random variable $X^{(d)}$. The singular value decomposition of $\mathbf{X}^{(d)}$ gives,

$$\underset{d\times n}{\mathbf{X}^{(d)}} = \begin{bmatrix} \hat{\xi}_{X1}^{(d)} & \hat{\xi}_{X2}^{(d)} & \cdots & \hat{\xi}_{Xn}^{(d)} \end{bmatrix} \cdot \begin{bmatrix} \sqrt{\hat{\lambda}_{X1}^{(d)}} & 0 & \ldots & 0 \\ 0 & \sqrt{\hat{\lambda}_{X2}^{(d)}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{\hat{\lambda}_{Xn}^{(d)}} \end{bmatrix} \begin{bmatrix} \left(Z_{\hat{\xi}_{X1}}^{(d)}\right)^T \\ \left(Z_{\hat{\xi}_{X2}}^{(d)}\right)^T \\ \vdots \\ \left(Z_{\hat{\xi}_{Xn}}^{(d)}\right)^T \end{bmatrix},$$

where $\hat{\xi}_{Xi}^{(d)}$ and $\hat{\lambda}_{Xi}^{(d)}$) are eigenvectors and eigenvalues of the sample covariance matrix $\hat{\mathbf{\Sigma}}_X^{(d)}$, $\|Z_{\hat{\xi}_{Xi}}^{(d)}\| = 1$ and $\langle Z_{\hat{\xi}_{Xi}}^{(d)}, Z_{\hat{\xi}_{Xj}}^{(d)} \rangle = 0$ for $i \neq j$. Note that $Z_{\hat{\xi}_{Xi}}^{(d)}$ is the standardized scores of the projections of the n observations in $\hat{\xi}_{Xi}^{(d)}$ onto the eigenvectors $\hat{\xi}_{Xi}^{(d)}$,

$$Z_{\hat{\xi}_{Xi}}^{(d)} = \frac{\left(\mathbf{X}^{(d)}\right)^T \hat{\xi}_{Xi}^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}},\ i = 1, 2, .., n.$$

Define $Z_{\hat{\eta}_{X1}}^{(d)}$ to be the standardized scores of the projections of the n observations in $\hat{\xi}_{Xi}^{(d)}$ onto the first singular vector $\hat{\eta}_{X1}^{(d)}$ obtained from the sample cross-covariance matrix $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$,

$$Z_{\hat{\eta}_{X1}}^{(d)} = \frac{\left(\mathbf{X}^{(d)}\right)^T \hat{\eta}_{X1}^{(d)}}{\sqrt{\hat{\lambda}_{X1}^{(d)}}}.$$

First we show that $Z^{(d)}_{\hat{\xi}_{X1}}$ converges in probability to $Z^{(d)}_{\hat{\eta}_{X1}}$ as $d \to \infty$. By the triangle inequality,

$$\begin{aligned}\left\| Z^{(d)}_{\hat{\xi}_{X1}} - Z^{(d)}_{\hat{\eta}_{X1}} \right\|_2 &= \left\| \frac{\left(\mathbf{X}^{(d)}\right)^T \hat{\xi}^{(d)}_{X1}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} - \frac{\left(\mathbf{X}^{(d)}\right)^T \hat{\eta}^{(d)}_{X1}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right\|_2 \\ &\leq \left\| \frac{\left(\mathbf{X}^{(d)}\right)^T}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} (\hat{\xi}^{(d)}_{X1} - e^{(d)}_1) \right\|_2 + \left\| \frac{\left(\mathbf{X}^{(d)}\right)^T}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} (\hat{\eta}^{(d)}_{X1} - e^{(d)}_1) \right\|_2 .\end{aligned}$$

The first term can be written as,

$$\left\| \frac{\left(\mathbf{X}^{(d)}\right)^T}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} (\hat{\xi}^{(d)}_{X1} - e^{(d)}_1) \right\|^2_2 = \sum_{i=1}^{n} \left\langle \frac{\mathbf{X}^{(d)}(\bullet, i)}{\sqrt{\hat{\lambda}^{(d)}_{X1}}}, \hat{\xi}^{(d)}_{X1} - e^{(d)}_1 \right\rangle^2 ,$$

where $\mathbf{X}^{(d)}(\bullet, i)$ stands for the ith column of $\mathbf{X}^{(d)}$. Note that the first component in each column of $\mathbf{X}^{(d)}$ is of $O_P(\sqrt{d^\alpha})$, which is that of $\sqrt{\hat{\lambda}^{(d)}_{X1}}$ by Lemma 10, and the rest of the elements are of $O_P(1)$ by Lemma 11. Since $\|\hat{\xi}^{(d)}_{X1} - e^{(D)}_1\|_2$ converge in probability to 0 by Lemma 10, using the Cauchy-Schwarz inequality and the fact that $\alpha > 1$,

$$\begin{aligned}\left\langle \frac{\mathbf{X}^{(d)}(\bullet, i)}{\sqrt{\hat{\lambda}^{(d)}_{X1}}}, \hat{\xi}^{(d)}_{X1} - e^{(d)}_1 \right\rangle^2 &\leq \left\| \frac{\mathbf{X}^{(d)}(\bullet, i)}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right\|^2_2 \left\| \hat{\xi}^{(d)}_{X1} - e^{(d)}_1 \right\|^2_2 \\ &= \left(\left(\frac{\mathbf{X}^{(d)}(1, i)}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right)^2 + \sum_{j=2}^{d} \left(\frac{\mathbf{X}^{(d)}(j, i)}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right)^2 \right) \left\| \hat{\xi}^{(d)}_{X1} - e^{(d)}_1 \right\|^2_2 \\ &= \left(\left(\frac{\mathbf{X}^{(d)}(1, i)}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right)^2 + \frac{d^\alpha}{\hat{\lambda}^{(d)}_{X1}} \frac{1}{d^{\alpha-1}} \sum_{j=2}^{d} \frac{\left(\mathbf{X}^{(d)}(j, i)\right)^2}{d} \right) \left\| \hat{\xi}^{(d)}_{X1} - e^{(d)}_1 \right\|^2_2 \\ &\xrightarrow[d\to\infty]{p} 0, \ i = 1, 2, .., n.\end{aligned}$$

Therefore,

$$\left\| \frac{\left(\mathbf{X}^{(d)}\right)^T}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} (\hat{\xi}^{(d)}_{X1} - e^{(d)}_1) \right\|^2_2 \xrightarrow[d\to\infty]{p} 0.$$

Similarly, using the results given in Lemma 4, 10 and 11, we have,

$$\left\| \frac{\left(\mathbf{X}^{(d)}\right)^T}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} (\hat{\eta}^{(d)}_{X1} - e^{(d)}_1) \right\|^2_2 \xrightarrow[d\to\infty]{p} 0,$$

which leads to,

$$\left\| Z^{(d)}_{\hat{\xi}_{X1}} - Z^{(d)}_{\hat{\eta}_{X1}} \right\|_2 \xrightarrow[d\to\infty]{p} 0.$$

Now we show that the inner product $\langle Z^{(d)}_{\hat{\xi}_{Xi}}, Z^{(d)}_{\hat{\eta}_{X1}} \rangle$ for $i = 2, 3, .., n$, converges to 0 as $d \to \infty$. Using $\langle Z^{(d)}_{\hat{\xi}_{Xi}}, Z^{(d)}_{\hat{\xi}_{X1}} \rangle = 0$ for $i = 2, 3, .., n$,

$$\begin{aligned}
\left\langle Z^{(d)}_{\hat{\xi}_{Xi}}, Z^{(d)}_{\hat{\eta}_{X1}} \right\rangle &= \left\langle Z^{(d)}_{\hat{\xi}_{Xi}}, Z^{(d)}_{\hat{\eta}_{X1}} - Z^{(d)}_{\hat{\xi}_{X1}} + Z^{(d)}_{\hat{\xi}_{X1}} \right\rangle \\
&= \left\langle Z^{(d)}_{\hat{\xi}_{Xi}}, Z^{(d)}_{\hat{\eta}_{X1}} - Z^{(d)}_{\hat{\xi}_{X1}} \right\rangle + \left\langle Z^{(d)}_{\hat{\xi}_{Xi}}, Z^{(d)}_{\hat{\xi}_{X1}} \right\rangle \\
&= \left\langle Z^{(d)}_{\hat{\xi}_{Xi}}, Z^{(d)}_{\hat{\eta}_{X1}} - Z^{(d)}_{\hat{\xi}_{X1}} \right\rangle.
\end{aligned}$$

Using the Cauchy-Schwarz inequality and the fact that $Z^{(d)}_{\hat{\xi}_{Xi}}$ for $i = 1, 2, .., n$, are standerdized scores (of unit variance),

$$\begin{aligned}
\left\langle Z^{(d)}_{\hat{\xi}_{Xi}}, Z^{(d)}_{\hat{\eta}_{X1}} - Z^{(d)}_{\hat{\xi}_{X1}} \right\rangle^2 &\leq \left\| Z^{(d)}_{\hat{\xi}_{Xi}} \right\|_2^2 \left\| Z^{(d)}_{\hat{\eta}_{X1}} - Z^{(d)}_{\hat{\xi}_{X1}} \right\|_2^2 \\
&= 1 \times \left\| Z^{(d)}_{\hat{\eta}_{X1}} - Z^{(d)}_{\hat{\xi}_{X1}} \right\|_2^2 \\
&\xrightarrow[d \to \infty]{p} 0, \ i = 1, 2, .., n.
\end{aligned}$$

We now show that $\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}} / \sqrt{\hat{\lambda}^{(d)}_{X1}}$ converges in probability to $\hat{\eta}^{(d)}_{X1}$ as $d \to \infty$. By the triangle inequality,

$$\begin{aligned}
\left\| \hat{\eta}^{(d)}_{X1} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right\|_2 &= \left\| \left(\hat{\eta}^{(d)}_{X1} - \hat{\xi}^{(d)}_{X1} \right) - \left(\hat{\xi}^{(d)}_{X1} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right) \right\|_2 \\
&\leq \left\| \hat{\eta}^{(d)}_{X1} - \hat{\xi}^{(d)}_{X1} \right\|_2 + \left\| \hat{\xi}^{(d)}_{X1} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right\|_2.
\end{aligned}$$

As $\|\hat{\xi}^{(d)}_{X1} - e^{(d)}_1\|_2$ and $\|\hat{\eta}^{(d)}_{X1} - e^{(d)}_1\|_2$ converge in probability to 0 by Lemma 4 and 11, so does $\|\hat{\eta}^{(d)}_{X1} - \hat{\xi}^{(d)}_{X1}\|_2$. By the triangle inequality, the second term becomes,

$$\begin{aligned}
\left\| \hat{\xi}^{(d)}_{X1} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right\|_2 &= \left\| \left(\hat{\xi}^{(d)}_{X1} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\xi}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right) - \left(\frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\xi}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right) \right\|_2 \\
&\leq \left\| \hat{\xi}^{(d)}_{X1} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\xi}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right\|_2 + \left\| \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\xi}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right\|_2.
\end{aligned}$$

The first term is 0 since $\mathbf{X}^{(d)} Z^{(d)}_{\hat{\xi}_{X1}} / \sqrt{\hat{\lambda}^{(d)}_{X1}} = \hat{\xi}^{(d)}_{X1}$ in the singular decomposition of $\mathbf{X}^{(d)}$. The second term goes as follows,

$$\left\| \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\xi}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} \right\|_2^2 = \left\| \frac{\mathbf{X}^{(d)}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} (Z^{(d)}_{\hat{\xi}_{X1}} - Z^{(d)}_{\hat{\eta}_{X1}}) \right\|_2^2$$

$$= \sum_{i=1}^{d} \left\langle \left(\frac{\mathbf{X}^{(d)}(i,\bullet)}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \right)^T, Z_{\hat{\xi}_{X1}}^{(d)} - Z_{\hat{\eta}_{X1}}^{(d)} \right\rangle^2,$$

where $\mathbf{X}^{(d)}(i,\bullet)$ stands for the ith row of the $d \times n$ data matrix $\mathbf{X}^{(d)}$. Note that only the first row of the data matrix $\mathbf{X}^{(d)}$ has its components of magnitude of $O_P(\sqrt{d^\alpha})$, and the components of the rest of the rows are all of $O_P(1)$. Since we showed that $\|Z_{\hat{\xi}_{X1}}^{(d)} - Z_{\hat{\eta}_{X1}}^{(d)}\|_2$ converges in probability to 0 as $d \to \infty$,

$$\left\langle \left(\frac{\mathbf{X}^{(d)}(1,\bullet)}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \right)^T, Z_{\hat{\xi}_{X1}}^{(d)} - Z_{\hat{\eta}_{X1}}^{(d)} \right\rangle^2 \leq \left\| \left(\frac{\mathbf{X}^{(d)}(1,\bullet)}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \right)^T \right\|_2^2 \left\| Z_{\hat{\xi}_{X1}}^{(d)} - Z_{\hat{\eta}_{X1}}^{(d)} \right\|_2^2$$
$$\xrightarrow[d\to\infty]{p} 0.$$

Similarly,

$$\left\langle \left(\frac{\mathbf{X}^{(d)}(i,\bullet)}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \right)^T, Z_{\hat{\xi}_{X1}}^{(d)} - Z_{\hat{\eta}_{X1}}^{(d)} \right\rangle^2 \leq \left\| \left(\frac{\mathbf{X}^{(d)}(i,\bullet)}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \right)^T \right\|_2^2 \left\| Z_{\hat{\xi}_{X1}}^{(d)} - Z_{\hat{\eta}_{X1}}^{(d)} \right\|_2^2$$
$$\xrightarrow[d\to\infty]{p} 0, \ i = 2, 3, .., d.$$

Therefore, we proved that,

$$\left\| \hat{\eta}_{X1}^{(d)} - \frac{\mathbf{X}^{(d)} Z_{\hat{\eta}_{X1}}^{(d)}}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \right\|_2 \xrightarrow[d\to\infty]{p} 0.$$

Finally to determine the speed of $\langle \hat{\eta}_{X1}^{(d)}, \hat{\xi}_{Xi}^{(d)} \rangle$ converging to 0 as $d \to \infty$,

$$\begin{aligned}
\left(Z_{\hat{\eta}_{X1}}^{(d)} \right)^T Z_{\hat{\xi}_{Xi}}^{(d)} &= \frac{\left(Z_{\hat{\eta}_{X1}}^{(d)} \right)^T \left(\mathbf{X}^{(d)} \right)^T \hat{\xi}_{Xi}^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}} \\
&= \frac{\sqrt{\hat{\lambda}_{X1}^{(d)}} \frac{\left(\mathbf{X}^{(d)} Z_{\hat{\eta}_{X1}}^{(d)} \right)^T}{\sqrt{\hat{\lambda}_{X1}^{(d)}}} \hat{\xi}_{Xi}^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}} \\
&= \frac{\sqrt{d^\alpha}}{\sqrt{d}} \frac{\frac{\sqrt{\hat{\lambda}_{X1}^{(d)}}}{\sqrt{d^\alpha}} \frac{\left\langle \mathbf{X}^{(d)} Z_{\hat{\eta}_{X1}}^{(d)}, \hat{\xi}_{Xi}^{(d)} \right\rangle}{\sqrt{\hat{\lambda}_{X1}^{(d)}}}}{\frac{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}{\sqrt{d}}}, \ i = 2, 3, ..n.
\end{aligned}$$

By Lemma 10 and 11, the magnitudes of $\sqrt{\hat{\lambda}_{X1}^{(d)}}/\sqrt{d^\alpha}$ and $\sqrt{\hat{\lambda}_{Xi}^{(d)}}/\sqrt{d}$ are of $O_P(1)$. Recall that

$\langle Z^{(d)}_{\hat{\eta}_{X1}}, Z^{(d)}_{\hat{\xi}_{Xi}} \rangle$ converges in probability to 0 as $d \to \infty$. Hence,

$$\left\langle \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}}, \hat{\xi}^{(d)}_{Xi} \right\rangle \xrightarrow[d\to\infty]{p} o_P\left(\frac{1}{\sqrt{d^{\alpha-1}}}\right), \ i = 2, 3, .., n.$$

However, we know that,

$$\left\| \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} - \hat{\eta}^{(d)}_{X1} \right\|_2 = \left\| \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{X1}}} - \frac{\mathbf{X}^{(d)} Z^{(d)}_{\hat{\eta}_{X1}}}{\sqrt{\hat{\lambda}^{(d)}_{XY1}}} \right\|_2 \xrightarrow[d\to\infty]{p} 0,$$

where the magnitude of the sample singular value $\hat{\lambda}^{(d)}_{XY1}$ is the same as that of $\hat{\lambda}^{(d)}_{X1}$ as shown in Lemma 4. Therefore,

$$\langle \hat{\eta}^{(d)}_{X1}, \hat{\xi}^{(d)}_{Xi} \rangle = o_P\left(\frac{1}{\sqrt{d^{\alpha-1}}}\right), \ i = 2, 3, .., n.$$

□

By Lemma 12, the matrix $\hat{\mathbf{R}}^{(d)}_1$ in (3.29) has the following form as $d \to \infty$,

$$\left\| \hat{\mathbf{R}}^{(d)}_1 - \begin{bmatrix} \frac{\sqrt{\langle c_1, d_1 \rangle}}{\|c_1\|_2} & \underset{1\times(d-1)}{\mathbf{0}} \\ \underset{(d-1)\times 1}{\mathbf{0}} & \underset{(d-1)\times(d-1)}{\mathbf{0}} \end{bmatrix} \right\|_2 \xrightarrow[d\to\infty]{P} 0. \tag{3.30}$$

Similarly, the corresponding part of $\hat{\mathbf{R}}^{(d)}_1$ on the right side of the matrix $\hat{\mathbf{R}}^{(d)}$ (call it $\hat{\mathbf{R}}^{(d)}_2$) reduces to a similar form of (3.30),

$$\left\| \hat{\mathbf{R}}^{(d)}_2 - \begin{bmatrix} \frac{\sqrt{\langle c_1, d_1 \rangle}}{\|c_2\|_2} & \underset{1\times(d-1)}{\mathbf{0}} \\ \underset{(d-1)\times 1}{\mathbf{0}} & \underset{(d-1)\times(d-1)}{\mathbf{0}} \end{bmatrix} \right\|_2 \xrightarrow[d\to\infty]{P} 0.$$

Then, the matrix $\hat{\mathbf{R}}^{(d)}$ is written as,

$$\hat{\mathbf{R}}^{(d)} = \begin{bmatrix} \hat{\xi}^{(d)}_{X1} & \hat{\xi}^{(d)}_{X2} & \dots & \hat{\xi}^{(d)}_{Xn} \end{bmatrix} \hat{\mathbf{R}}^{(d)}_1 \hat{\mathbf{R}}^{(d)}_2 \begin{bmatrix} \left(\hat{\xi}^{(d)}_{Y1}\right)^T \\ \left(\hat{\xi}^{(d)}_{Y2}\right)^T \\ \vdots \\ \left(\hat{\xi}^{(d)}_{Yn}\right)^T \end{bmatrix},$$

and its limiting matrix is,

$$\left\| \hat{\mathbf{R}}^{(d)} - \begin{bmatrix} \hat{\xi}^{(d)}_{X1} & \hat{\xi}^{(d)}_{X2} & \dots & \hat{\xi}^{(d)}_{Xn} \end{bmatrix} \begin{bmatrix} \frac{\langle c_1, d_1 \rangle}{\|c_1\|_2 \|c_2\|_2} & \underset{1\times(d-1)}{\mathbf{0}} \\ \underset{(d-1)\times 1}{\mathbf{0}} & \underset{(d-1)\times(d-1)}{\mathbf{0}} \end{bmatrix} \begin{bmatrix} \left(\hat{\xi}^{(d)}_{Y1}\right)^T \\ \left(\hat{\xi}^{(d)}_{Y2}\right)^T \\ \vdots \\ \left(\hat{\xi}^{(d)}_{Yn}\right)^T \end{bmatrix} \right\|_2 \xrightarrow[d\to\infty]{P} 0, \tag{3.31}$$

where $\hat{\xi}_{Yi}^{(d)}$ is a sample eigenvector obtained from the eigendecomposition of the sample covariance matrix shown in (3.22) from the data matrix $\mathbf{Y}^{(d)}$ of (3.16).

Perform the SVD of $\hat{\mathbf{R}}^{(d)}$ and denote its sample left and right singular vectors by $\hat{\eta}_{RXi}^{(d)}$ and $\hat{\eta}_{RYi}^{(d)}$ for $i = 1, 2, .., n$. From (3.31), the sample singular vectors of $\hat{\mathbf{R}}^{(d)}$ are linear combinations of $\hat{\xi}_{Xi}^{(d)}$'s or $\hat{\xi}_{Yi}^{(d)}$'s,

$$\hat{\eta}_{RXi}^{(d)} = \sum_{j=1}^{n} a_j^{(d)} \hat{\xi}_{Xj}^{(d)}, \ \hat{\eta}_{RYi}^{(d)} = \sum_{k=1}^{n} b_k^{(d)} \hat{\xi}_{Yk}^{(d)}, \ i = 1, 2, \ldots, n, \tag{3.32}$$

where $(a_1^{(d)})^2 + (a_2^{(d)})^2 + \cdots + (a_n^{(d)})^2 = (b_1^{(d)})^2 + (b_2^{(d)})^2 + \cdots + (b_n^{(d)})^2 = 1$ to ensure that norms of $\hat{\eta}_{RXi}^{(d)}$ and $\hat{\eta}_{RYi}^{(d)}$ are of unit length.

3.4.3.4 Behavior of the first sample canonical weight vector The sample canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and $\hat{\psi}_{Yi}^{(d)}$ are found by unscaling and normalizing $\hat{\eta}_{RX1}^{(d)}$ and $\hat{\eta}_{RY1}^{(d)}$ as in (3.4) using the pseudoinverse of (3.13),

$$\hat{\psi}_{Xi}^{(d)} = \frac{\left(\hat{\mathbf{\Sigma}}_X^{(d)}\right)^{-\frac{1}{2}} \hat{\eta}_{RXi}^{(d)}}{\left\| \left(\hat{\mathbf{\Sigma}}_X^{(d)}\right)^{-\frac{1}{2}} \hat{\eta}_{RXi}^{(d)} \right\|_2}, \ \hat{\psi}_{Yi}^{(d)} = \frac{\left(\hat{\mathbf{\Sigma}}_Y^{(d)}\right)^{-\frac{1}{2}} \hat{\eta}_{RYi}^{(d)}}{\left\| \left(\hat{\mathbf{\Sigma}}_Y^{(d)}\right)^{-\frac{1}{2}} \hat{\eta}_{RYi}^{(d)} \right\|_2}, \ i = 1, 2, .., n. \tag{3.33}$$

To see if $\hat{\psi}_{X1}^{(d)}$ and $\hat{\psi}_{Y1}^{(d)}$ is consistent to their population counterparts $\psi_{X1}^{(d)}$ and $\psi_{Y1}^{(d)}$ given in (3.7), we investigate the limiting value of their inner products $\langle \hat{\psi}_{X1}^{(d)}, \psi_{X1}^{(d)} \rangle$ and $\langle \hat{\psi}_{Y1}^{(d)}, \psi_{Y1}^{(d)} \rangle$, which measure the cosine of the angle formed by the two. Expanding (3.33),

$$\begin{aligned}
\langle \hat{\psi}_{X1}^{(d)}, \psi_{X1}^{(d)} \rangle &= \frac{\left(\psi_{X1}^{(d)}\right)^T \left(\hat{\mathbf{\Sigma}}_X^{(d)}\right)^{-\frac{1}{2}} \hat{\eta}_{RX1}^{(d)}}{\left\|\psi_{X1}^{(d)}\right\|_2 \left\| \left(\hat{\mathbf{\Sigma}}_X^{(d)}\right)^{-\frac{1}{2}} \hat{\eta}_{RX1}^{(d)} \right\|_2} \\
&= \frac{\left(\psi_{X1}^{(d)}\right)^T \left(\sum_{i=1}^{n} \hat{\lambda}_{Xi}^{(d)} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T \right)^{-\frac{1}{2}} \hat{\eta}_{RX1}^{(d)}}{\left\|\psi_{X1}^{(d)}\right\|_2 \left\| \left(\sum_{i=1}^{n} \hat{\lambda}_{Xi}^{(d)} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T \right)^{-\frac{1}{2}} \hat{\eta}_{RX1}^{(d)} \right\|_2} \\
&= \frac{\left(\cos\theta_X e1^{(d)} + \sin\theta_X e_2^{(d)}\right)^T \left(\sum_{i=1}^{n} \frac{1}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T \right) \hat{\eta}_{RX1}^{(d)}}{\left\|\cos\theta_X e_1^{(d)} + \sin\theta_X e_2^{(d)}\right\|_2 \left\| \left(\sum_{i=1}^{n} \frac{1}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T \right) \hat{\eta}_{RX1}^{(d)} \right\|_2}.
\end{aligned}$$

Let's take a close look at the denominator term. It is easy to see that $\left\|\cos\theta_X e_1^{(d)} + \sin\theta_X e_2^{(d)}\right\|_2 = 1$. Using (3.32),

$$\begin{aligned}\left\|\left(\sum_{i=1}^{n}\frac{1}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\hat{\xi}_{Xi}^{(d)}\left(\hat{\xi}_{Xi}^{(d)}\right)^T\right)\hat{\eta}_{RX1}^{(d)}\right\|_2 &= \left\|\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{a_j}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle\hat{\xi}_{Xi}^{(d)},\hat{\xi}_{Xj}^{(d)}\rangle\hat{\xi}_{Xi}^{(d)}\right\|_2 \\ &= \left\|\sum_{i=1}^{n}\frac{a_i}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\hat{\xi}_{Xi}^{(d)}\right\|_2.\end{aligned}$$

Now take a look at the numerator term,

$$\begin{aligned}&\left(\cos\theta_X e_1^{(d)} + \sin\theta_X e_2^{(d)}\right)^T\left(\sum_{i=1}^{n}\frac{1}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\hat{\xi}_{Xi}^{(d)}\left(\hat{\xi}_{Xi}^{(d)}\right)^T\right)\hat{\eta}_{RX1}^{(d)} \\ &= \left(\cos\theta_X e_1^{(d)} + \sin\theta_X e_2^{(d)}\right)^T\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{a_j^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle\hat{\xi}_{Xi}^{(d)},\hat{\xi}_{Xj}^{(d)}\rangle\hat{\xi}_{Xi}^{(d)}\right) \\ &= \left(\cos\theta_X e_1^{(d)}\right)^T\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{a_j^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle\hat{\xi}_{Xi}^{(d)},\hat{\xi}_{Xj}^{(d)}\rangle\hat{\xi}_{Xi}^{(d)}\right) + \left(\sin\theta_X e_2^{(d)}\right)^T\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{a_j^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle\hat{\xi}_{Xi}^{(d)},\hat{\xi}_{Xj}^{(d)}\rangle\hat{\xi}_{Xi}^{(d)}\right) \\ &= \left(\cos\theta_X e_1^{(d)}\right)^T\left(\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\hat{\xi}_{Xi}^{(d)}\right) + \left(\sin\theta_X e_2^{(d)}\right)^T\left(\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\hat{\xi}_{Xi}^{(d)}\right) \\ &= \cos\theta_X\left(\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle e_1^{(d)},\hat{\xi}_{Xi}^{(d)}\rangle\right) + \sin\theta_X\left(\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle e_2^{(d)},\hat{\xi}_{Xi}^{(d)}\rangle\right).\end{aligned}$$

Hence, we have,

$$\langle\hat{\psi}_{X1}^{(d)},\psi_{X1}^{(d)}\rangle = \frac{\cos\theta_X\left(\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle e_1^{(d)},\hat{\xi}_{Xi}^{(d)}\rangle\right) + \sin\theta_X\left(\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle e_2^{(d)},\hat{\xi}_{Xi}^{(d)}\rangle\right)}{\left\|\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\hat{\xi}_{Xi}^{(d)}\right\|_2}. \tag{3.34}$$

Simiarly,

$$\langle\hat{\psi}_{Y1}^{(d)},\psi_{Y1}^{(d)}\rangle = \frac{\cos\theta_Y\left(\sum_{i=1}^{n}\frac{b_i^{(d)}}{\sqrt{\hat{\lambda}_{Yi}^{(d)}}}\langle e_1^{(d)},\hat{\xi}_{Yi}^{(d)}\rangle\right) + \sin\theta_Y\left(\sum_{i=1}^{n}\frac{b_i^{(d)}}{\sqrt{\hat{\lambda}_{Yi}^{(d)}}}\langle e_2^{(d)},\hat{\xi}_{Yi}^{(d)}\rangle\right)}{\left\|\sum_{i=1}^{n}\frac{b_i^{(d)}}{\sqrt{\hat{\lambda}_{Yi}^{(d)}}}\hat{\xi}_{Yi}^{(d)}\right\|_2}.$$

Observe from (3.31) that, for $\hat{\eta}_{RX1}^{(d)}$, $a_1^{(d)}$ converges to 1 and $a_i^{(d)}$ for $i = 2, 3, \ldots, n$, does to 0 as $d \to \infty$, which implies,

$$\left(\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle e_1^{(d)},\hat{\xi}_{Xi}^{(d)}\rangle - \frac{1}{\sqrt{\hat{\lambda}_{X1}^{(d)}}}\langle e_1^{(d)},\hat{\xi}_{X1}^{(d)}\rangle\right)^2 \xrightarrow[d\to\infty]{P} 0,$$

$$\left(\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle e_2^{(d)},\hat{\xi}_{Xi}^{(d)}\rangle-\frac{1}{\sqrt{\hat{\lambda}_{X1}^{(d)}}}\langle e_2^{(d)},\hat{\xi}_{X1}^{(d)}\rangle\right)^2\xrightarrow[d\to\infty]{P}0,$$

$$\left\|\sum_{i=1}^{n}\frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\hat{\xi}_{Xi}^{(d)}-\frac{1}{\sqrt{\hat{\lambda}_{X1}^{(d)}}}\hat{\xi}_{X1}^{(d)}\right\|_2^2\xrightarrow[d\to\infty]{P}0.$$

Then, (3.34) becomes,

$$\left(\langle\hat{\psi}_{X1}^{(d)},\psi_{X1}^{(d)}\rangle-\frac{\cos\theta_X\left(\frac{d^\alpha}{\sqrt{\hat{\lambda}_{X1}^{(d)}}}\langle e_1^{(d)},\hat{\xi}_{X1}^{(d)}\rangle\right)+\sin\theta_X\left(\frac{d^\alpha}{\sqrt{\hat{\lambda}_{X1}^{(d)}}}\langle e_2^{(d)},\hat{\xi}_{X1}^{(d)}\rangle\right)}{\left\|\frac{d^\alpha}{\sqrt{\hat{\lambda}_{X1}^{(d)}}}\hat{\xi}_{X1}^{(d)}\right\|_2}\right)^2\xrightarrow[d\to\infty]{P}0.$$

Recall from Lemma 10 and 11 that $d^\alpha/\sqrt{\hat{\lambda}_{X1}^{(d)}}$ is of ($\asymp O_P(1)$), $\langle e_1^{(d)},\hat{\xi}_{X1}^{(d)}\rangle$ and $\langle e_2^{(d)},\hat{\xi}_{X1}^{(d)}\rangle$ converge in probability to 1 and 0, respectively. Therefore, we finally have,

$$\left(\langle\hat{\psi}_{X1}^{(d)},\psi_{X1}^{(d)}\rangle-\cos\theta_X\right)^2\xrightarrow[d\to\infty]{P}0.$$

Similarly for $\hat{\psi}_{Y1}^{(d)}$,

$$\left(\langle\hat{\psi}_{Y1}^{(d)},\psi_{Y1}^{(d)}\rangle-\cos\theta_Y\right)^2\xrightarrow[d\to\infty]{P}0.$$

3.4.3.5 Behavior of the first sample canonical correlation coefficient The limiting value of first sample canonical correlation coefficients is found as the first singular value of the matrix $\hat{\mathbf{R}}^{(d)}$ in (3.31) under the limiting operation of $d\to\infty$, which turns out its (1,1)th entry,

$$\left(\hat{\rho}_1-\frac{\langle c_1,d_1\rangle/d^\alpha}{\|c_1\|_2^2/d^\alpha\|c_2\|_2^2/d^\alpha}\right)^2\xrightarrow[d\to\infty]{P}0.$$

Recall from Lemma 2 and 8 that $\langle c_1,d_1\rangle/d^\alpha$ is the limiting value of (1,1) entry of $n\hat{\mathbf{\Sigma}}_{XY}^{(d)}/d^\alpha$ as $d\to\infty$ and that $\|c_1\|_2^2/d^\alpha$ and $\|c_1\|_2^2/d^\alpha$ are the limiting values of the (1,1) entries of $n\hat{\mathbf{\Sigma}}_X^{(d)}/d^\alpha$ and $n\hat{\mathbf{\Sigma}}_Y^{(d)}/d^\alpha$. That is, those values are found in the (d+1,d+1), (1,1) and (1,d+1) positions of the matrix of (3.21) scaled by d^α as d increases,

$$\frac{1}{d^\alpha}\begin{bmatrix}n\hat{\mathbf{\Sigma}}_X^{(d)} & n\hat{\mathbf{\Sigma}}_{XY}^{(d)}\\ \left(n\hat{\mathbf{\Sigma}}_{XY}^{(d)}\right)^T & n\hat{\mathbf{\Sigma}}_Y^{(d)}\end{bmatrix}=\frac{1}{d^\alpha}\begin{bmatrix}\mathbf{X}^{(d)}\\ \mathbf{Y}^{(d)}\end{bmatrix}\begin{bmatrix}\mathbf{X}^{(d)}\\ \mathbf{Y}^{(d)}\end{bmatrix}^T.$$

Referring to (3.16) and (3.17), we switch the (1,2) and (1,d+1) entries, the (2,1) and (d+1,1) entries and the (2,2) and (d+1,d+1) entries of the matrix $\mathbf{\Sigma}_T^{(2d)}$ and denote the resulting matrix by $\acute{\mathbf{\Sigma}}_T^{(2d)}$.

We, also, switch the second and $(d+1)$th rows of the matrix $\mathbf{Z}^{(2d)}$ and denote the resulting matrix by $\acute{\mathbf{Z}}^{(2d)}$. Then, the two data matrices below,

$$\begin{bmatrix}\mathbf{X}^{(d)}\\ \mathbf{Y}^{(d)}\end{bmatrix} = \left(\mathbf{\Sigma}_T^{(2d)}\right)^{\frac{1}{2}} \mathbf{Z}^{(2d)} \text{ and } \left(\acute{\mathbf{\Sigma}}_T^{(2d)}\right)^{\frac{1}{2}} \acute{\mathbf{Z}}^{(2d)},$$

are the same only except that the $(d+1)$th row of the first matrix is the second row of the second one. Following the similar steps as done in the proof of Lemma 2, it can be shown that,

$$\left\| \left(\frac{\acute{\mathbf{\Sigma}}_T^{(2d)}}{d^\alpha}\right)^{\frac{1}{2}} \acute{\mathbf{Z}}^{(2d)} - \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y & \underset{1\times(2d-2)}{\mathbf{0}} \\ \rho\sigma_X\sigma_Y & \sigma_X^2 & \underset{1\times(2d-2)}{\mathbf{0}} \\ \underset{(2d-2)\times 1}{\mathbf{0}} & \underset{(2d-2)\times 1}{\mathbf{0}} & \underset{(2d-2)\times(2d-2)}{\mathbf{0}} \end{bmatrix}^{\frac{1}{2}} \begin{bmatrix} z_{X1\bullet}\\ z_{Y1\bullet}\\ z_{X2\bullet}\\ \vdots\\ z_{Xd\bullet}\\ z_{Y2\bullet}\\ \vdots\\ z_{Yd\bullet}\end{bmatrix} \right\|_2 \xrightarrow[d\to\infty]{P} 0.$$

Then, the limiting values of $\langle c_1, d_1\rangle/d^\alpha$, $\|c_1\|_2^2/d^\alpha$ and $\|c_1\|_2^2/d^\alpha$ are found in the first 2×2 block of the following matrix, as $d\to\infty$,

$$\frac{1}{d^\alpha}\left(\acute{\mathbf{\Sigma}}_T^{(2d)}\right)^{\frac{1}{2}} \acute{\mathbf{Z}}^{(2d)} \left(\acute{\mathbf{Z}}^{(2d)}\right)^T \left(\acute{\mathbf{\Sigma}}_T^{(2d)}\right)^{\frac{1}{2}} = \left(\frac{\acute{\mathbf{\Sigma}}_T^{(2d)}}{d^\alpha}\right)^{\frac{1}{2}} \acute{\mathbf{Z}}^{(2d)} \left(\acute{\mathbf{Z}}^{(2d)}\right)^T \left(\frac{\acute{\mathbf{\Sigma}}_T^{(2d)}}{d^\alpha}\right)^{\frac{1}{2}}.$$

Therefore, the limiting values of $\langle c_1, d_1\rangle/d^\alpha$, $\|c_1\|_2^2/d^\alpha$ and $\|c_1\|_2^2/d^\alpha$ are identified as the (2,2), (1,1) and (2,1) entries of the following matrix, respectively,

$$\begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_X^2 \end{bmatrix}^{\frac{1}{2}} \begin{bmatrix} z_{X1\bullet}\\ z_{Y1\bullet}\end{bmatrix} \begin{bmatrix} z_{X1\bullet}\\ z_{Y1\bullet}\end{bmatrix}^T \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_X^2 \end{bmatrix}^{\frac{1}{2}}.$$

3.4.3.6 Behavior of the rest of sample canonical weight vectors We use the equation (3.34) for the second canonical weight vector $\hat{\psi}_{X2}^{(d)}$. Recall from (3.31) that $a_1^{(d)}$ of $\hat{\eta}_{RXi}^{(d)}$,for $i = 2, 3, \ldots, n$, converges to 0, which implies,

$$\left(\sum_{i=1}^n \frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle e_1^{(d)}, \hat{\xi}_{Xi}^{(d)}\rangle - \sum_{j=2}^n \frac{a_j^{(d)}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}}\langle e_1^{(d)}, \hat{\xi}_{Xj}^{(d)}\rangle\right)^2 \xrightarrow[d\to\infty]{P} 0,$$

$$\left(\sum_{i=1}^n \frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}}\langle e_2^{(d)}, \hat{\xi}_{Xi}^{(d)}\rangle - \sum_{j=2}^n \frac{a_j^{(d)}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}}\langle e_2^{(d)}, \hat{\xi}_{Xj}^{(d)}\rangle\right)^2 \xrightarrow[d\to\infty]{P} 0,$$

$$\left\| \sum_{i=1}^{n} \frac{a_i^{(d)}}{\sqrt{\hat{\lambda}_{Xi}^{(d)}}} \hat{\xi}_{Xi}^{(d)} - \sum_{j=2}^{n} \frac{a_j^{(d)}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}} \hat{\xi}_{Xj}^{(d)} \right\|_2^2 \xrightarrow[d\to\infty]{P} 0.$$

Then, (3.34) becomes,

$$\left(\langle \hat{\psi}_{X2}^{(d)}, \psi_{X1}^{(d)} \rangle - \frac{\cos\theta_X \left(\sum_{j=2}^{n} \frac{da_j^{(d)}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}} \langle e_1^{(d)}, \hat{\xi}_{Xj}^{(d)} \rangle \right) + \sin\theta_X \left(\sum_{j=2}^{n} \frac{da_j^{(d)}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}} \langle e_1^{(d)}, \hat{\xi}_{Xj}^{(d)} \rangle \right)}{\left\| \sum_{j=2}^{n} \frac{da_j^{(d)}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}} \hat{\xi}_{Xj}^{(d)} \right\|_2} \right)^2 \xrightarrow[d\to\infty]{P} 0.$$

We know from11 that $d/\sqrt{\hat{\lambda}_{Xi}^{(d)}}$, for $i = 2, 3, \ldots, n$, is of ($\asymp O_P(1)$) and that $\langle e_1^{(d)}, \hat{\xi}_{Xi}^{(d)} \rangle$ and $\langle e_2^{(d)}, \hat{\xi}_{Xi}^{(d)} \rangle$, for $i = 2, 3, \ldots, n$, both converge in probability to 0. Therefore, we finally have,

$$\left(\langle \hat{\psi}_{Xi}^{(d)}, \psi_{X1}^{(d)} \rangle - 0 \right)^2 \xrightarrow[d\to\infty]{P} 0, \ i = 2, 3, \ldots, n.$$

Results are similar for $\hat{\psi}_{Xi}^{(d)}$ for $i = 3, 4, \ldots, n$. By the similar argument,

$$\left(\langle \hat{\psi}_{Yi}^{(d)}, \psi_{Y1}^{(d)} \rangle - 0 \right)^2 \xrightarrow[d\to\infty]{P} 0, \ i = 2, 3, \ldots, n.$$

3.4.3.7 Behavior of the rest of sample canonical correlation coefficients

The asymptotic ith sample canonical correlation coefficients, for $i = 2, 3, \ldots, n$, is found as the ith singular value of the matrix $\hat{\mathbf{R}}^{(d)}$ in (3.31) under the limiting operation of $d \to \infty$,

$$\hat{\rho}_i \xrightarrow[d\to\infty]{p} 0, \ i = 1, 2, \ldots, n.$$

3.4.4 Case of $\alpha < 1$

We, now, prove Theorem 1 under the condition of $\alpha < 1$ with the spiked model of the population covariance structure of $X^{(d)}$ and $Y^{(d)}$ described in (3.9).

3.4.4.1 Behavior of the cross-covariance matrix

Lemma 13 (CCA HDLSS Asymptotic Lemma 12.)**.** *For $\alpha < 1$, the magnitudes of the sample singular values $\hat{\lambda}_{XYi}^{(d)}$ for $i = 1, 2, .., n$ are of $O_p(d)$ as $d \to \infty$.*

Proof. Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left\| \frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d} \right\|_F^2 =& \frac{\langle c_1, d_1\rangle^2}{d^{2-2\alpha}d^{2\alpha}} + \frac{\langle c_1, d_2\rangle^2}{d^{2-\alpha}d^{\alpha}} + \frac{\langle c_2, d_1\rangle^2}{d^{2-\alpha}d^{\alpha}} + \frac{\langle c_2, d_2\rangle^2}{d^2} \\
&+ \tau_Y^2 \sum_{i=3}^{d} \frac{\langle c_1, z_{Yi\bullet}\rangle^2}{d^{2-\alpha}d^{\alpha}} + \tau_Y^2 \sum_{i=3}^{d} \frac{\langle c_2, z_{Yi\bullet}\rangle^2}{d^2} + \tau_X^2 \sum_{i=3}^{d} \frac{\langle z_{Xi\bullet}, d_1\rangle^2}{d^{2-\alpha}d^{\alpha}} \\
&+ \tau_X^2 \sum_{i=3}^{d} \frac{\langle z_{Xi\bullet}, d_2\rangle^2}{d^2} + \tau_X^2\tau_Y^2 \sum_{i=3}^{d}\sum_{j=3}^{d} \frac{\langle z_{Xi\bullet}, z_{Yj\bullet}\rangle^2}{d^2} \\
\leq& \frac{\langle c_1, d_1\rangle^2}{d^{2-2\alpha}d^{2\alpha}} + \frac{\langle c_1, d_2\rangle^2}{d^{2-\alpha}d^{\alpha}} + \frac{\langle c_2, d_1\rangle^2}{d^{2-\alpha}d^{\alpha}} + \frac{\langle c_2, d_2\rangle^2}{d^2} \\
&+ \tau_Y^2 \sum_{i=3}^{d} \frac{\|c_1\|_2^2\|z_{Yi\bullet}\|_2^2}{d^{2-\alpha}d^{\alpha}} + \tau_Y^2 \sum_{i=3}^{d} \frac{\|c_2\|_2^2\|z_{Yi\bullet}\|_2^2}{d^2} + \tau_X^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2\|d_1\|_2^2}{d^{2-\alpha}d^{\alpha}} \\
&+ \tau_X^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2\|d_2\|_2^2}{d^2} + \tau_X^2\tau_Y^2 \sum_{i=3}^{d}\sum_{j=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2\|z_{Yj\bullet}\|_2^2}{d^2}.
\end{aligned} \tag{3.35}$$

Since $\langle c_1, d_1\rangle^2$ is of $O_P(d^{\alpha})$, $\langle c_1, d_2\rangle^2$ and $\langle c_2, d_1\rangle^2$ are of $O_P(\sqrt{d^{\alpha}})$, and $\langle c_2, d_2\rangle^2$ is of $O_P(1)$, using $\alpha < 1$,

$$\frac{\langle c_1, d_1\rangle^2}{d^{2-2\alpha}d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0, \ \frac{\langle c_1, d_2\rangle^2}{d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0, \ \frac{\langle c_2, d_1\rangle^2}{d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0, \ \frac{\langle c_2, d_2\rangle^2}{d^{2\alpha}} \xrightarrow[d\to\infty]{p} 0.$$

It is not hard to see that $\|z_{Xi\bullet}\|_2^2 \sim \chi_n^2$ and $\|z_{Yi\bullet}\|_2^2 \sim \chi_n^2$. Note that $\|c_1\|_2^2$ and $\|d_1\|_2^2$ are of $O_P(d^{\alpha})$ and that $\|c_2\|_2^2$ and $\|d_2\|_2^2$ are of $O_P(1)$. With the law of large numbers and the condition that $\alpha < 1$,

$$\begin{aligned}
\tau_Y^2 \sum_{i=3}^{d} \frac{\|c_1\|_2^2\|z_{Yi\bullet}\|_2^2}{d^{2-\alpha}d^{\alpha}} &= \frac{\tau_Y^2}{d^{1-\alpha}} \left\| \frac{c_1}{\sqrt{d^{\alpha}}} \right\|_2^2 \sum_{i=3}^{d} \frac{\|z_{Yi\bullet}\|_2^2}{d} \\
&\xrightarrow[d\to\infty]{p} 0 \times O_P(1) \times E(\chi_n^2) = 0, \\
\tau_Y^2 \sum_{i=3}^{d} \frac{\|c_2\|_2^2\|z_{Yi\bullet}\|_2^2}{d^2} &= \frac{\tau_Y^2\|c_2\|_2^2}{d} \sum_{i=3}^{d} \frac{\|z_{Yi\bullet}\|_2^2}{d} \\
&\xrightarrow[d\to\infty]{p} 0 \times E(\chi_n^2) = 0, \\
\tau_X^2 \sum_{i=3}^{d} \frac{\|d_1\|_2^2\|z_{Xi\bullet}\|_2^2}{d^{2-\alpha}d^{\alpha}} &= \frac{\tau_X^2}{d^{1-\alpha}} \left\| \frac{d_1}{\sqrt{d^{\alpha}}} \right\|_2^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2}{d} \\
&\xrightarrow[d\to\infty]{p} 0 \times O_P(1) \times E(\chi_n^2) = 0, \\
\tau_X^2 \sum_{i=3}^{d} \frac{\|d_2\|_2^2\|z_{Xi\bullet}\|_2^2}{d^2} &= \frac{\tau_X^2\|d_2\|_2^2}{d} \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2}{d} \\
&\xrightarrow[d\to\infty]{p} 0 \times E(\chi_n^2) = 0.
\end{aligned}$$

Using $\alpha > 1$ and the law of large numbers,

$$\tau_X^2\tau_Y^2 \sum_{i=3}^{d}\sum_{j=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2\|z_{Yj\bullet}\|_2^2}{d^2} = \tau_X^2\tau_Y^2 \sum_{i=3}^{d} \frac{\|z_{Xi\bullet}\|_2^2}{d} \sum_{j=3}^{d} \frac{\|z_{Yj\bullet}\|_2^2}{d}$$
$$\xrightarrow[d\to\infty]{p} \tau_X^2\tau_Y^2 \times E(\chi_n^2) \times E(\chi_n^2)$$
$$= \tau_X^2\tau_Y^2 n^2.$$

We have,

$$\left\| \frac{\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d} \right\|_F^2 = \sum_{i=1}^{n} \left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d} \right)^2 \xrightarrow[d\to\infty]{p} \leq \tau_X^2\tau_Y^2.$$

Therefore, the magnitudes of the sample singular values $\hat{\lambda}_{XYi}^{(d)}$ for $i = 1, 2, .., n$ are of $O_p(d)$. □

Lemma 14 (CCA HDLSS Asymptotic Lemma 13.)**.** *This lemma improves Lemma 13. For $\alpha < 1$, the sample singular values $n\hat{\lambda}_{XYi}^{(d)}$ for $i = 1, 2, .., n$ are of ($\asymp O_p(d)$) with the following limiting quantity as $d \to \infty$,*

$$\frac{\hat{\lambda}_{XYi}^{(d)}}{d} \xrightarrow[d\to\infty]{P} = \frac{\tau_X\tau_Y}{n}, \ i = 1, 2, .., n.$$

Proof. First we calculate the exact limiting value of the Frobenius norm of $\hat{\mathbf{\Sigma}}_{XY}^{(d)}/d$ as $d \to \infty$. In (3.35), the first 8 terms converges in probability to 0 as $d \to \infty$. The last term of (3.35) expands as,

$$\sum_{i=3}^{d}\sum_{j=3}^{d} \frac{\langle z_{Xi\bullet}, z_{Yj\bullet}\rangle^2}{d^2} = \sum_{i=3}^{d}\sum_{j=3}^{d} \left(\sum_{k=1}^{n} \frac{z_{Xik}z_{Yjk}}{d} \right)^2$$
$$= \sum_{k=1}^{n} \left(\sum_{i=3}^{d} \frac{z_{Xik}^2}{d} \right) \left(\sum_{i=3}^{d} \frac{z_{Yjk}^2}{d} \right)$$
$$+ \sum_{k\neq k'} \left(\sum_{i=3}^{d} \frac{z_{Xik}z_{Yik'}}{d} \right) \left(\sum_{j=3}^{d} \frac{z_{Xjk}z_{Yjk'}}{d} \right).$$

Since $E(z_{yik}z_{xik'}) = 0$ for $k \neq k'$, by the law of large numbers,

$$\sum_{k\neq k'} \left(\sum_{i=3}^{d} \frac{z_{yik}z_{xik'}}{d} \right) \left(\sum_{j=3}^{d} \frac{z_{yjk}z_{xjk'}}{d} \right) \xrightarrow[d\to\infty]{P} 0.$$

Since z_{yik}^2 and z_{xik}^2 follow a Chi-squared distribution with a degree of freedom of 1, by the law of large numberss,

$$\sum_{k=1}^{n} \left(\sum_{i=3}^{d} \frac{z_{yik}^2}{d} \right) \left(\sum_{i=3}^{d} \frac{z_{xjk}^2}{d} \right) \xrightarrow[d\to\infty]{P} n \times 1 \times 1 = n.$$

Hence,

$$\left\| \frac{n\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d} \right\|_F^2 \xrightarrow[d\to\infty]{P} n\tau_X^2\tau_Y^2. \tag{3.36}$$

Now we set a bound for each sample singular value $\hat{\lambda}_{XYi}^{(d)}/d$ for $i = 1, 2, .., n$ as $d \to \infty$. Consider the data matrix $\mathbf{X}^{(d)}$ and $\mathbf{Y}^{(d)}$ in (3.19). Write the singular vectors $\hat{\eta}_{Xi}^{(d)}$ and $\hat{\eta}_{Yi}^{(d)}$ of the sample cross-covariance matrix $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$ of (3.22) by the linear combinations of the sample eigenvectors $\hat{\xi}_{Xj}^{(d)}$ and $\hat{\xi}_{Yj}^{(d)}$,

$$\hat{\eta}_{Xi}^{(d)} = \sum_{j=1}^{n} a_j \hat{\xi}_{Xj}^{(d)},\ \hat{\eta}_{Yi}^{(d)} = \sum_{j=1}^{n} b_j \hat{\xi}_{Yj}^{(d)},\ i = 1, 2, .., n,$$

where $a_1^2 + a_2^2 + \cdots + a_n^2 = b_1^2 + b_2^2 + \cdots + b_n^2 = 1$ to ensure a unit length. Using $Cov^2(X, Y) \leq Var(X)Var(Y)$,

$$\begin{aligned}
\left(\hat{\lambda}_{XYi}^{(d)}\right)^2 &= \frac{1}{n^2}\left\langle \left(\mathbf{X}^{(d)}\right)^T \hat{\eta}_{Xi}^{(d)}, \left(\mathbf{Y}^{(d)}\right)^T \hat{\eta}_{Yi}^{(d)} \right\rangle^2 \\
&\leq \frac{1}{n^2}\left\langle \left(\mathbf{X}^{(d)}\right)^T \hat{\eta}_{Xi}^{(d)}, \left(\mathbf{X}^{(d)}\right)^T \hat{\eta}_{Xi}^{(d)} \right\rangle \left\langle \left(\mathbf{Y}^{(d)}\right)^T \hat{\eta}_{Yi}^{(d)}, \left(\mathbf{Y}^{(d)}\right)^T \hat{\eta}_{Yi}^{(d)} \right\rangle.
\end{aligned}$$

Using the fact that the covariance between two sets of scores of the projections of the observations onto any two different eigenvectors is 0,

$$\begin{aligned}
\frac{1}{n}\left\langle \left(\mathbf{X}^{(d)}\right)^T \hat{\eta}_{Xi}^{(d)}, \left(\mathbf{X}^{(d)}\right)^T \hat{\eta}_{Xi}^{(d)} \right\rangle &= \frac{1}{n}\left\langle \left(\mathbf{X}^{(d)}\right)^T \sum_{j=1}^{n} a_j \hat{\xi}_{Xj}^{(d)}, \left(\mathbf{X}^{(d)}\right)^T \sum_{j=1}^{n} a_j \hat{\xi}_{Xj}^{(d)} \right\rangle \\
&= \sum_{j=j'} \frac{1}{n}\left\langle \left(\mathbf{X}^{(d)}\right)^T a_j \hat{\xi}_{Xj}^{(d)}, \left(\mathbf{X}^{(d)}\right)^T a_{j'} \hat{\xi}_{Xj'}^{(d)} \right\rangle \\
&\quad + \sum_{j\neq j'} \frac{1}{n}\left\langle \left(\mathbf{X}^{(d)}\right)^T a_j \hat{\xi}_{Xj}^{(d)}, \left(\mathbf{X}^{(d)}\right)^T a_{j'} \hat{\xi}_{Xj'}^{(d)} \right\rangle \\
&= \sum_{j=j'} \frac{1}{n}\left\langle \left(\mathbf{X}^{(d)}\right)^T a_j \hat{\xi}_{Xj}^{(d)}, \left(\mathbf{X}^{(d)}\right)^T a_{j'} \hat{\xi}_{Xj'}^{(d)} \right\rangle \\
&= \sum_{j=1}^{n} \frac{a_j^2}{n}\left\langle \left(\mathbf{X}^{(d)}\right)^T \hat{\xi}_{Xj}^{(d)}, \left(\mathbf{X}^{(d)}\right)^T \hat{\xi}_{Xj}^{(d)} \right\rangle \\
&= \sum_{j=1}^{n} a_j^2 \hat{\lambda}_{Xj}^{(d)}.
\end{aligned}$$

Similarly,

$$\frac{1}{n}\left\langle \left(\mathbf{Y}^{(d)}\right)^T \hat{\eta}_{Yi}^{(d)}, \left(\mathbf{Y}^{(d)}\right)^T \hat{\eta}_{Yi}^{(d)} \right\rangle = \sum_{k=1}^{n} b_k^2 \hat{\lambda}_{Yk}^{(d)}.$$

By Lemma 11,

$$\begin{aligned}\left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\right)^2 &\leq \frac{1}{n^2}\frac{\left\langle \left(\mathbf{X}^{(d)}\right)^T \hat{\eta}_{Xi}^{(d)}, \left(\mathbf{X}^{(d)}\right)^T \hat{\eta}_{Xi}^{(d)} \right\rangle}{d}\frac{\left\langle \left(\mathbf{Y}^{(d)}\right)^T \hat{\eta}_{Yi}^{(d)}, \left(\mathbf{Y}^{(d)}\right)^T \hat{\eta}_{Yi}^{(d)} \right\rangle}{d} \\ &= \left(\sum_{j=1}^{n} a_j^2 \frac{\hat{\lambda}_{Xj}^{(d)}}{d}\right)\left(\sum_{k=1}^{n} b_k^2 \frac{\hat{\lambda}_{Yk}^{(d)}}{d}\right) \\ &\xrightarrow[d\to\infty]{P} \frac{\tau_X^2}{n}\frac{\tau_Y^2}{n}\left(\sum_{j=1}^{n} a_j^2\right)\left(\sum_{k=1}^{n} b_k^2\right) \\ &= \frac{\tau_X^2\tau_Y^2}{n^2},\ i=1,2,..,n.\end{aligned}$$

The constraints above and the proven fact of (3.36),

$$\left\| \frac{\hat{\mathbf{\Sigma}}_{XY}^{(d)}}{d} \right\|_F^2 = \sum_{i=1}^{n} \left(\frac{\hat{\lambda}_{XYi}^{(d)}}{d}\right)^2 \xrightarrow[d\to\infty]{P} \frac{1}{n}\tau_X^2\tau_Y^2,$$

implies that,

$$\frac{\hat{\lambda}_{XYi}^{(d)}}{d} \xrightarrow[d\to\infty]{P} \frac{\tau_X\tau_Y}{n},\ i=1,2,..,n.$$

□

3.4.4.2 Behavior of the sample covariance matrices Here, we only include the result about the sample covariance matrix $\hat{\mathbf{\Sigma}}_X^{(d)}$. That of $\hat{\mathbf{\Sigma}}_Y^{(d)}$ is similar.

Lemma 15 (CCA HDLSS Asymptotic Lemma 14.)**.** *The sample eigenvalue $\hat{\lambda}_{Xi}^{(d)}$ and eigenvector $\hat{\xi}_{Xi}^{(d)}$, $i=2,3,..,d$, converge in probability to the following quantities as $d\to\infty$,*

$$\frac{\hat{\lambda}_{Xi}^{(d)}}{d} \xrightarrow[d\to\infty]{P} \frac{\tau_X^2}{n},\ \langle \hat{\xi}_{Xi}^{(d)}, \xi^{(d)} \rangle \xrightarrow[d\to\infty]{P} 0,\ i=1,2,\ldots,n,$$

where $\xi^{(d)}$ is an any given vector in R^d.

Proof. Recall that the underlying random variable $X^{(d)}$ of the sample covariance matrix $\hat{\mathbf{\Sigma}}_X^{(d)}$ has a simple spiked covariance structure of (3.9). Then, the asymptotic behavior of a sample eigenvalue is a direct consequence of Theorem 1 for $\alpha < 1$. The result on the behavior of $\langle \hat{\xi}_{Xi}^{(d)}, \xi^{(d)} \rangle$ is indicated in the proof of Theorem 1 given in [24]. □

3.4.4.3 Behavior of the matrix $\hat{\mathbf{R}}^{(d)}$ Lemma 14 states that the singular values of the sample cross-covariance matrix $\hat{\mathbf{\Sigma}}_{XY}^{(d)}$ become indistinguishable in the limit of $d \to \infty$, in which situation there exist an infinite number of choices for a singular vector set. We choose its singular vectors such that the sample eigenvectors of the sample covariance matrices $\hat{\mathbf{\Sigma}}_{X}^{(d)}$ and $\hat{\mathbf{\Sigma}}_{X}^{(d)}$ are their limiting quantities,

$$\left\|\hat{\eta}_{Xi}^{(d)} - \hat{\xi}_{Xi}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{P} 0, \;\; \left\|\hat{\eta}_{Yj}^{(d)} - \hat{\xi}_{Yj}^{(d)}\right\|_2^2 \xrightarrow[d\to\infty]{P} 0, \;\; i,j = 1,2,\ldots,n.$$

Then, it is easily shown that the matrix $\hat{\mathbf{R}}^{(d)}$ defined in (3.28), reduces to the following limiting form, using the limiting values of the sample singular vectors and eigenvectors given in Lemma 14 and 15,

$$\left\| \hat{\mathbf{R}}^{(d)} - \begin{bmatrix} \hat{\xi}_{X1}^{(d)} & \hat{\xi}_{X2}^{(d)} & \ldots & \hat{\xi}_{Xn}^{(d)} \end{bmatrix} \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \left(\hat{\xi}_{Y1}^{(d)}\right)^T \\ \left(\hat{\xi}_{Y2}^{(d)}\right)^T \\ \vdots \\ \left(\hat{\xi}_{Yn}^{(d)}\right)^T \end{bmatrix} \right\|_F^2 \xrightarrow[d\to\infty]{P} 0. \tag{3.37}$$

3.4.4.4 Behavior of sample canonical weight vectors Use the equation (3.34) for a sample canonical weight vector $\hat{\psi}_{Xi}^{(d)}$ represented as in (3.32) for a given i. We know from Lemma 15 that the magnitude of $d/\sqrt{\hat{\lambda}_{Xi}^{(d)}}$, for $i = 1,2,\ldots,n$, is of ($\asymp O_P(1)$) and that $\langle e_1^{(d)}, \hat{\xi}_{Xi}^{(d)} \rangle$ and $\langle e_2^{(d)}, \hat{\xi}_{Xi}^{(d)} \rangle$, for $i = 1,2,\ldots,n$, converge in probability to 0 as $d \to \infty$, which leads to,

$$\left(\langle \hat{\psi}_{Xi}^{(d)}, \psi_{X1}^{(d)} \rangle - \frac{\cos\theta_X \left(\sum_{j=1}^{n} \frac{da_j^{(d)}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}} \langle e_1^{(d)}, \hat{\xi}_{Xj}^{(d)} \rangle \right) + \sin\theta_X \left(\sum_{j=1}^{n} \frac{da_j^{(d)}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}} \langle e_1^{(d)}, \hat{\xi}_{Xj}^{(d)} \rangle \right)}{\left\| \sum_{j=1}^{n} \frac{da_j^{(d)}}{\sqrt{\hat{\lambda}_{Xj}^{(d)}}} \hat{\xi}_{Xj}^{(d)} \right\|_2} \right)^2 \xrightarrow[d\to\infty]{P} 0.$$

Hence, we have,

$$\left(\langle \hat{\psi}_{Xi}^{(d)}, \psi_{X1}^{(d)} \rangle - 0 \right)^2 \xrightarrow[d\to\infty]{P} 0, \;\; i = 1,2,\ldots,n.$$

Similarly,

$$\left(\langle \hat{\psi}_{Yi}^{(d)}, \psi_{Y1}^{(d)} \rangle - 0 \right)^2 \xrightarrow[d\to\infty]{P} 0, \;\; i = 1,2,\ldots,n.$$

3.4.4.5 Behavior of sample canonical correlation coefficients The asymptotic ith sample canonical correlation coefficients is found as the ith singular value of the matrix (3.37) under the

limiting operation of $d \to \infty$,

$$\hat{\rho}_i \xrightarrow[d\to\infty]{P} 1, \ i = 1, 2, \ldots, n.$$

3.5 SIMULATION

Simulation study in this section aims at verifying the asymptotic behavior of sample canonical correlation coefficients and their corresponding weight vectors given in Theorem 1 as dimension d grows with sample size n fixed. We first state the parameter settings to be used. For the spiked covariance structures of the random variables $X^{(d)}$ and $Y^{(d)}$ described in (3.5) and (3.6), we set $\sigma_X^2 = \tau_X^{(d)} = \sigma_Y^2 = \tau_Y^{(d)} = 1$. The population canonical weight vectors described in (3.7) and population canonical correlation coefficient are set to be,

$$\psi_X^{(d)} = (\cos 0.75\pi)e_1^{(d)} + (\sin 0.75\pi)e_2^{(d)}, \ \psi_Y^{(d)} = (\cos 0.75\pi)e_1^{(d)} + (\sin 0.75\pi)e_2^{(d)}, \ \rho = 0.7.$$

Note that $\langle \psi_X^{(d)}, e_1^{(d)} \rangle = \langle \psi_Y^{(d)}, e_1^{(d)} \rangle = \cos 0.75\pi = 0.7071$, which implies that the angle between $\psi_X^{(d)}$ and $e_1^{(d)}$ is 135°. The population cross-covariance structure of $X^{(d)}$ and $Y^{(d)}$ can be accordingly defined as in (3.9). We perform 100 runs of simulations for each combination of different values of the following three sets,

- Sample size $n \in \{20, 80\}$,
- Dimension $d \in \{200, 500\}$,
- Exponent $\alpha \in \{0.2, 8\}$.

Each case, estimates of the first 5 canonical correlation coefficients $\hat{\rho}_i^{(d)}$ and their corresponding canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and $\hat{\psi}_{Yi}^{(d)}$ are obtained. The estimated vectors $\hat{\psi}_{Xi}^{(d)}$ and $\hat{\psi}_{Yi}^{(d)}$, for $i = 1, 2, \ldots, 5$, are compared to the population canonical weight vector $\psi_X^{(d)}$ using their inner product. Here, we do not include results of $\hat{\psi}_{Yi}^{(d)}$ as they are similar as those of $\hat{\psi}_{Xi}^{(d)}$.

Figure 10 presents the simulation results for a small sample size of $n = 20$. For $\alpha = 0.2$, sample coefficients and vectors are almost of no use as the estimated vectors tend to be as far away as possible from the population direction (implied in the inner products of 0) with always perfect correlation. When α increases to a high strength of 8, the first sample coefficient $\hat{\rho}_1^{(d)}$ approaches to the population direction whereas the rest degenerate to 0 as $d \to \infty$. The first left sample canonical weight vector $\hat{\psi}_{X1}^{(d)}$ converges to the direction $e_1^{(d)}$ (implied in the inner products

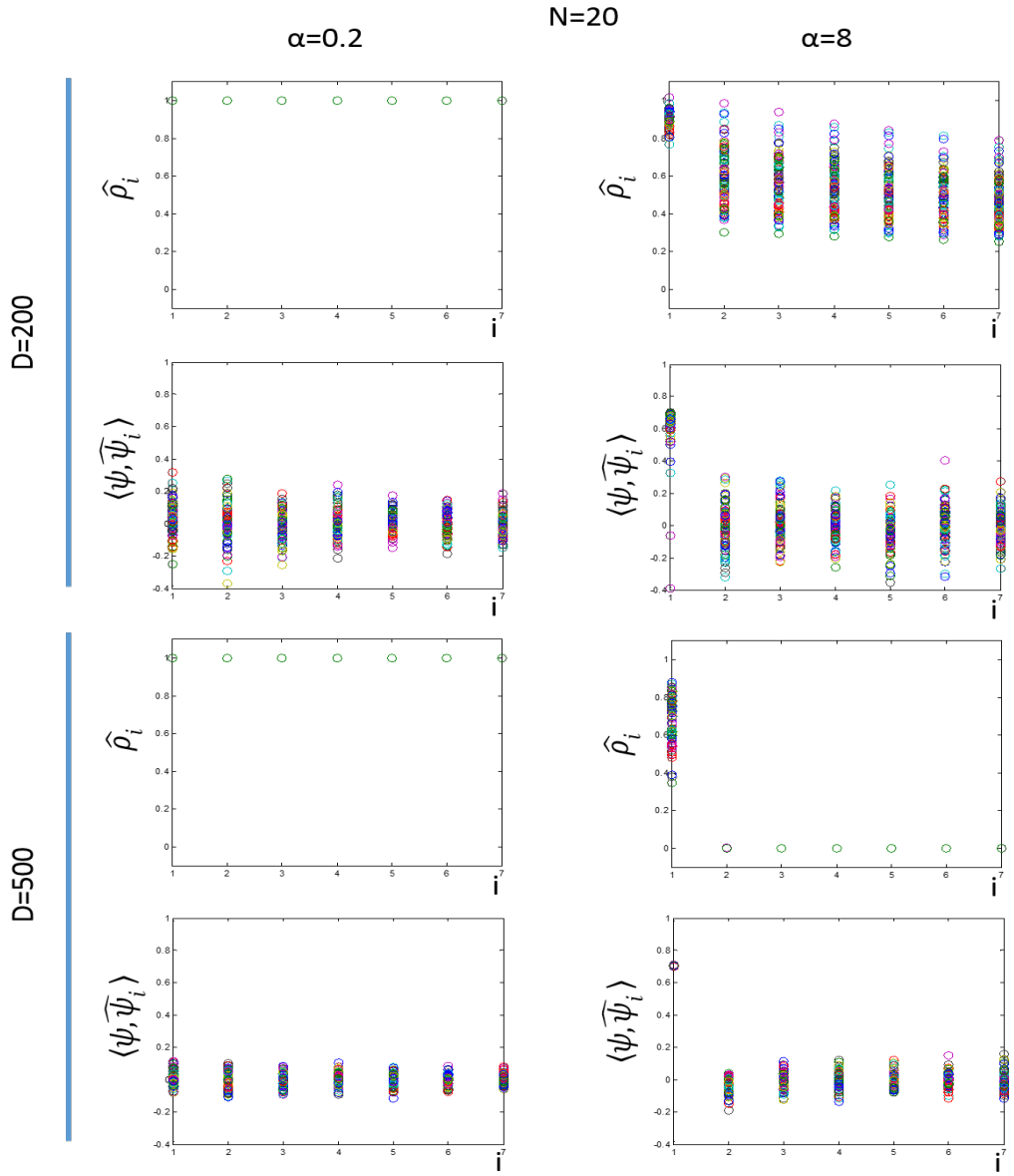

Figure 10. Estimated sample canonical correlation coefficients $\hat{\rho}_i^{(d)}$ and inner products of the sample left canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and the population canonical weight vector $\psi_{Xi}^{(d)}$, for $i = 1, 2, \ldots, 5$, obtained from 100 repetitions of simulations for different settings of dimension d and exponent α with a sample size of $n = 20$.

of $\cos 0.75\pi$) containing dominant variability as $d \to \infty$ and the rest carry no information on the population direction with tending to deviate from it by a highest degree of 90°. Figure 11 illustrates the results for a larger sample size of $n = 80$. For the case of $\alpha = 0.2$, the behavior of $\hat{\rho}_i^{(d)}$ and

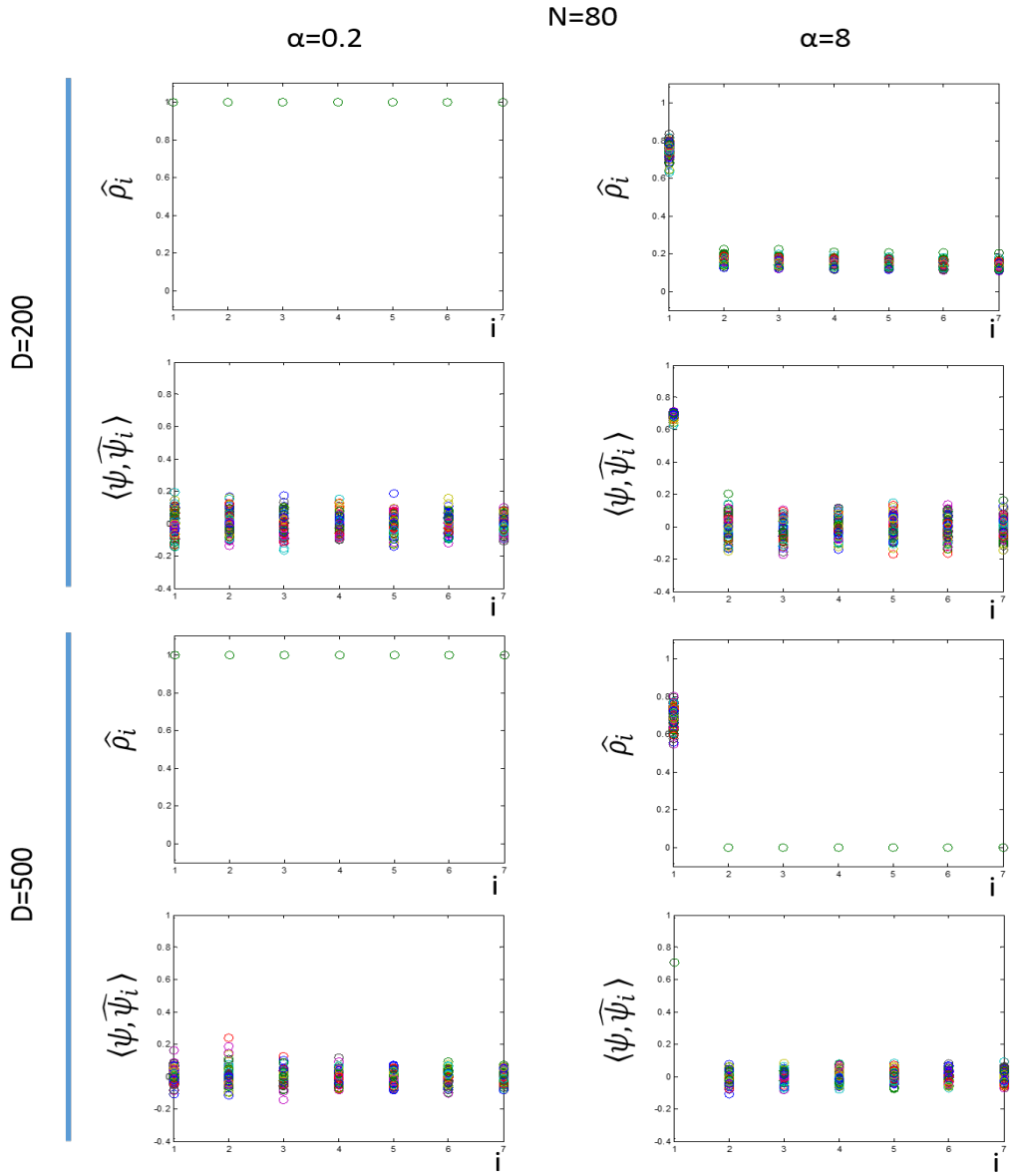

Figure 11. 100 estimated sample canonical correlation coefficients $\hat{\rho}_i^{(d)}$ and inner products of the sample left canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and the population canonical weight vector $\psi_{Xi}^{(d)}$, for $i = 1, 2, \ldots, 5$, obtained from 100 repetitions of simulations for different settings of dimension d and exponent α with a sample size of $n = 80$.

$\hat{\psi}_{Xi}^{(d)}$ is similar as that in a small sample size case. However, for $\alpha = 8$, we see a noticeable decrease in variability of the first sample canonical correlation coefficient $\hat{\rho}_1^{(d)}$ around a true value of 0.7 and of the rest of $\hat{\rho}_i^{(d)}$ around 0. This implies that the usual large sample theory works for $\hat{\rho}_1^{(d)}$.

Diminishing variability is also observed for the sample canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$, where the first sample vector $\hat{\psi}_{Xi}^{(d)}$ becomes almost identical to the largest variance direction $e_1^{(d)}$ and the rest diverge from the population canonical direction $\psi_X^{(d)}$.

3.6 DISCUSSION

A natural question arises about what asymptotic behavior of sample canonical weight vectors we can expect at the boundary case of $\alpha = 1$ as $d \to \infty$ with a sample size fixed at n ? When $\alpha > 1$, we saw that the angle between the first sample canonical weight vector and its population counterpart degenerates to 0 and when $\alpha < 1$, the angle between them diverges by as large as $\pi/2$. We conjecture that, with $\alpha = 1$, the angle formed by the first sample canonical weight vector and its population counterpart converges weakly to some random variable on the support of $[0, \pi/2]$. We leave an investigation of this conjecture for future work.

4.0 SUPERVISED JOINT AND INDIVIDUAL VARIATIONS EXPLAINED

4.1 INTRODUCTION

With a proliferation of technologies that can measure various aspects of a given sample in many sciences, we now often collect multiple data sets of differing domains on a common set of samples. We call those multiple data sets 'multi-block data'. One important aspect we need to address in analyzing multi-block data is to characterize associations among variables in different data sets. The dependency structure between multiple data sets may be utilized to infer interesting patterns of a population that would not be found with separate analyses from individual data sets. On the other hand, dependency structure within variables in a specific data set could impart unique and useful information. In this case, removing between-data-sets dependency structure and revealing the net within-data-set dependency structure will lead to more effective and clearer inference results.

Dependency among variables is often reflected in a direction along which data exhibit meaningful variations. For example, if 3-dimensional vector observations show large variation at the direction of $[1, -2, 0]^T$, then we can say that the first and second variables are associated in a way that, as the first increases by 1, the second tends to decrease by -2. Let $\mathbf{X}_i$, for $i = 1, 2, \ldots, m$, be a $n \times p_i$ matrix containing measurements for the p_i variables of the ith data sets on a common set of n objects. We can find a direction ξ in the $p_1 + p_2 + \cdots + p_m$ dimensional row space of the concatenated matrix $\mathbf{X}$,

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m],$$

such that the variation of scores (of projection of $\mathbf{X}$ onto ξ) is maximized under certain regulation conditions. Accordingly, the $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ parts of the resulting direction ξ can be viewed as associated directions between the data sets and the variation of $\mathbf{X}$ along ξ can be thought of as a joint variation across the data sets. At the same time, there could be a direction ξ_1 in the p_1

dimensional space along which only the data set $\mathbf{X}_1$ shows large variation. This kind of variation can be viewed as an individual variation.

Sometimes, one of data sets has supervision effect over the rest of data sets. To elaborate the case intuitively, consider an illustrative example of high-throughput biomedical data. Let $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$ contain gene expression level, genotype information and DNA methylation rate, respectively. We may obtain an additional data set $\mathbf{Y}$ on the same set of tissues that inherently relates to the underlying joint variations of the multi-block data. Suppose that we may have disease subtype information stored in $\mathbf{Y}$ for all samples. Conceptually, different disease subtypes may explain a large portion of genetic variations in individual data sets. In other words, the data set $\mathbf{Y}$, usually called supervision, potentially drives the joint variations across data sets in the multi-block data. This situation motivates investigations presented in this chapter.

We propose Supervised Joint and Individual Variations Explained (SupJIVE), a general framework for a systematic decomposition of variation in multi-block data according to supervision information available. SupJIVE aims to identify joint variations across multiple data sets and individual variations specific to each data set. It also aims to capture supervision effects that drive those variations. To this end, we combine the two recently proposed methods, Joint and Individual Variation Explained (JIVE) [35] and Supervised Singular Value Decomposition (SupSVD) [32], which are briefly described below.

The SupJIVE method, however, has two major issues: inability to capture partial joint variation structure and rank estimation problem (see Section 4.2.4 for details). To remedy these problems, we propose Generalized SupJIVE (G-SupJIVE). G-SupJIVE algorithm searches for variation structure in each subset of the whole data sets and supervision effect sequentially until the concatenated data matrix $\mathbf{X}$ exausts its rank or there is no supervision effect left in an automated fashion. Therefore, it does not requires any ad-hoc rank estimations. Moreover, G-SupJIVE method allows the supervision matrix $\mathbf{Y}$ to include multiple sets of supervision candidates that are thought of as possible drivers of variations. Then, the method identifies parts of supervision data sets that actually drive a specific variation and also reveals how they drive the variation.

4.1.1 JIVE

JIVE decomposes multi-block data $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m)$ into a sum of three components: a low rank approximation $\mathbf{J}$ capturing joint structure across data sets, a low rank approximations $\mathbf{A}_i$ capturing

individual structure specific to each data set, and a residual noise $\mathbf{E}$,

$$\begin{cases} \mathbf{X}_i = \mathbf{J}_i + \mathbf{A}_i + \mathbf{E}, \ i = 1, 2, \ldots, m, \\ \mathbf{J} = [\mathbf{J}_1, \mathbf{J}_2, \ldots, \mathbf{J}_m] = \mathbf{U}\mathbf{V}^T, \\ \mathbf{A}_i = \mathbf{U}_i \mathbf{V}_i^T, \end{cases} \tag{4.1}$$

where columns of $\mathbf{V}$ and $\mathbf{V}_i$'s contain direction (or laoding) vectors, and $\mathbf{U}$ and $\mathbf{U}_i$'s contain scores. Note that loading vectors in $\mathbf{V}$ contribute to variables across data sets $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ and so, with their scores in $\mathbf{U}$, constitute joint variation structure while $\mathbf{U}_i \mathbf{V}_i^T$ represents individual variation stucture of $\mathbf{X}_i$. The paper [35] proposed to estimate the parameters $\mathbf{V}, \mathbf{V}_i, \mathbf{U}$ and $\mathbf{U}_i$ by iteratively applying singular value decomposition (SVD) to the joint and individual parts of the data $\mathbf{X}$. There is no data set playing the role of supervisor.

4.1.2 SupSVD

When there are only two blocks of data $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}$ is the data set of main interest, and $\mathbf{Y}$ is the supervision data, SupSVD [32] seeks an orthogonal basis of the underlying low rank structure of the main data set $\mathbf{X}$ with respect to which $\mathbf{X}$ is decomposed into three components: one relevant to the supervision information $\mathbf{Y}$, another irrelevant of $\mathbf{Y}$ and the last one a residual noise part $\mathbf{E}$ so that,

$$\mathbf{X} = \mathbf{U}\mathbf{V}^T = (\mathbf{Y}\mathbf{B} + \mathbf{F})\mathbf{V}^T + \mathbf{E}, \tag{4.2}$$

where columns of $\mathbf{V}$ contain direction (or loading) vectors comprising an orthogonal basis of sample space, $\mathbf{B}$ is responsible for conversion of $\mathbf{Y}$ into scores with respect to the loading vectors in $\mathbf{V}$, $\mathbf{F}$ is a random matrix with a certain distributional structure imposed and $\mathbf{E}$ is a random noise. Intuitively, $\mathbf{Y}\mathbf{B}\mathbf{V}^T$ captures the variation that is driven by supervision $\mathbf{Y}$, and $\mathbf{F}\mathbf{V}^T$ captures the variation that is irrelevant of $\mathbf{Y}$. If supervision has no effect on the variation structure of $\mathbf{X}$, then the matrix $\mathbf{B}$ degenerates and SupSVD becomes equivalent to the usual SVD. SupSVD is basically an overparameterized model. In the paper [32], the parameters $\mathbf{B}, \mathbf{F}$ and $\mathbf{V}$ are estimated by modified EM algorithm under certain identifiability conditions and a Gaussian assumption on $\mathbf{F}$ and $\mathbf{E}$.

4.1.3 Motivating real data set [35]

The Cancer Genome Atlas (TCGA) [39] is an ongoing research effort to characterize cancer on a molecular level through the integrative analysis of multiple large scale genomic data sets, funded by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). Following the data analysis conducted in [35], we focus on a set of 234 Glioblastome Multiforme (GBM) tumor samples. GBM is a common and fatal form of malignant brain tumor. GBM samples are not homogeneous. Verhaak et al. [46] classified the GBM samples into four subtypes: Neural, Mesenchymal, Proneural and Classical. These subtypes have distinct expression characteristics, copy number alterations and gene mutations. In addition, there were clinical differences across subtypes in response to chemical therapy.

While the relation of copy number aberrations and somatic mutations to gene expressions has been established in [38, 05], the role of microRNA (miRNA) in GBM biology has not been well understood [41]. Current biological ideas, however, suggest that miRNAs might function mainly as negative regulators that decrease gene expression levels. We demonstrate our new methods on the integrated analysis of miRNA and gene expression data. For each tumor sample, there are measures of intensity for 534 miRNAs and 23,293 genes. These data are publicly available as a supplemental materials of [35].

Given the biological relation between gene expression and miRNA, it is reasonable to expect shared patterns in the two sets of measurements. we refer to such shared patterns as joint variation. We also expect the gene expression data to have its own variation that is unrelated to the miRNA and vice versa. This individual variation can be of biological interest. This individual patterns can interfere with finding the important joint patterns, just as joint variations can obscure the important signal specific to each data set. JIVE [35] is proposed to separate these joint and individual variations.

Given each tumor cell subtype's unique characteristics and different response to aggressive therapy, it is reasonable to expect that subtypes are responsible for some of joint and individual variations. We refer to the factors driving specific patterns as supervisions and the variations driven by them as supervised variations. The variations unrelated to any of supervisions are referred to as unsupervised variations. SupSVD [32] is a statistical framework proposed to separate supervised and unsupervised variations given supervision information.

The major goal of the integrative analysis of the GBM data set is clear given its data structure:

separate joint and individual variations and, for each of them, identify its driving subtypes. Note that JIVE and SupSVD each address only one of those two tasks.

Each of our proposed methods can achieve both of the major goals of the integrative multi-block analysis. As SupJIVE turns out to have a few issues, we analyze GBM data using G-SupJIVE. The first new method, SupJIVE, is discussed in Section 4.2, and the Generalized Sup-JIVE is in Section 4.3 and 4.4. The analysis of GBM data is presented in Section 4.5.

4.2 SUPJIVE

This section introduces a new model and estimation procedures to identify joint and individual variations and supervision effects driving them. Since we are not going to use this algorithm in practice due to the issues explained later in Section 4.2.4, we only demonstrate its performance using simulated data in Section 4.2.2. Recently, an integrated approach, called Supervised Integrated Factor Analysis for Multi-View Data (SIFA), to decompose multi-block data into joint and individual variations is proposed by [31], which uses a similar model but with different estimation approaches.

4.2.1 Population model

Adopting the SupSVD model (4.2) into the JIVE model (4.1), the formal model for SupJIVE is,

$$\begin{cases} \mathbf{X} &= [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m], \\ \mathbf{X}_i &= \mathbf{J}_i + \mathbf{A}_i + \mathbf{E}_i, \\ \mathbf{J} &= [\mathbf{J}_1, \mathbf{J}_2, \ldots, \mathbf{J}_k], \\ \mathbf{J} &= \mathbf{U}\mathbf{V}^T = (\mathbf{Y}\mathbf{B} + \mathbf{F})\mathbf{V}^T, \\ \mathbf{A}_i &= \mathbf{U}_i\mathbf{V}_i^T = (\mathbf{Y}_i\mathbf{B}_i + \mathbf{F}_i)\mathbf{V}_i^T. \end{cases} \tag{4.3}$$

where $\mathbf{U}\mathbf{V}^T$ (or $\mathbf{U}_i\mathbf{V}_i^T$) is the rank r (or r_i) decomposition of $\mathbf{J}$ (or $\mathbf{A}_i$); $\mathbf{B}$ (or $\mathbf{B}_i$) is a $q \times r$ (or $q_i \times r_i$) conversion matrix; and $\mathbf{F}$ (or $\mathbf{F}_i$) is a $n \times r$ (or $n \times r_i$) random matrix. We here assume that a joint variation $\mathbf{J}$ has its supervision $\mathbf{Y}$ (of dimension q) and further assume that individual variations $\mathbf{A}_i$ is possibly driven by their own supervisions $\mathbf{Y}_i$ (of dimension q_i). To fit the model

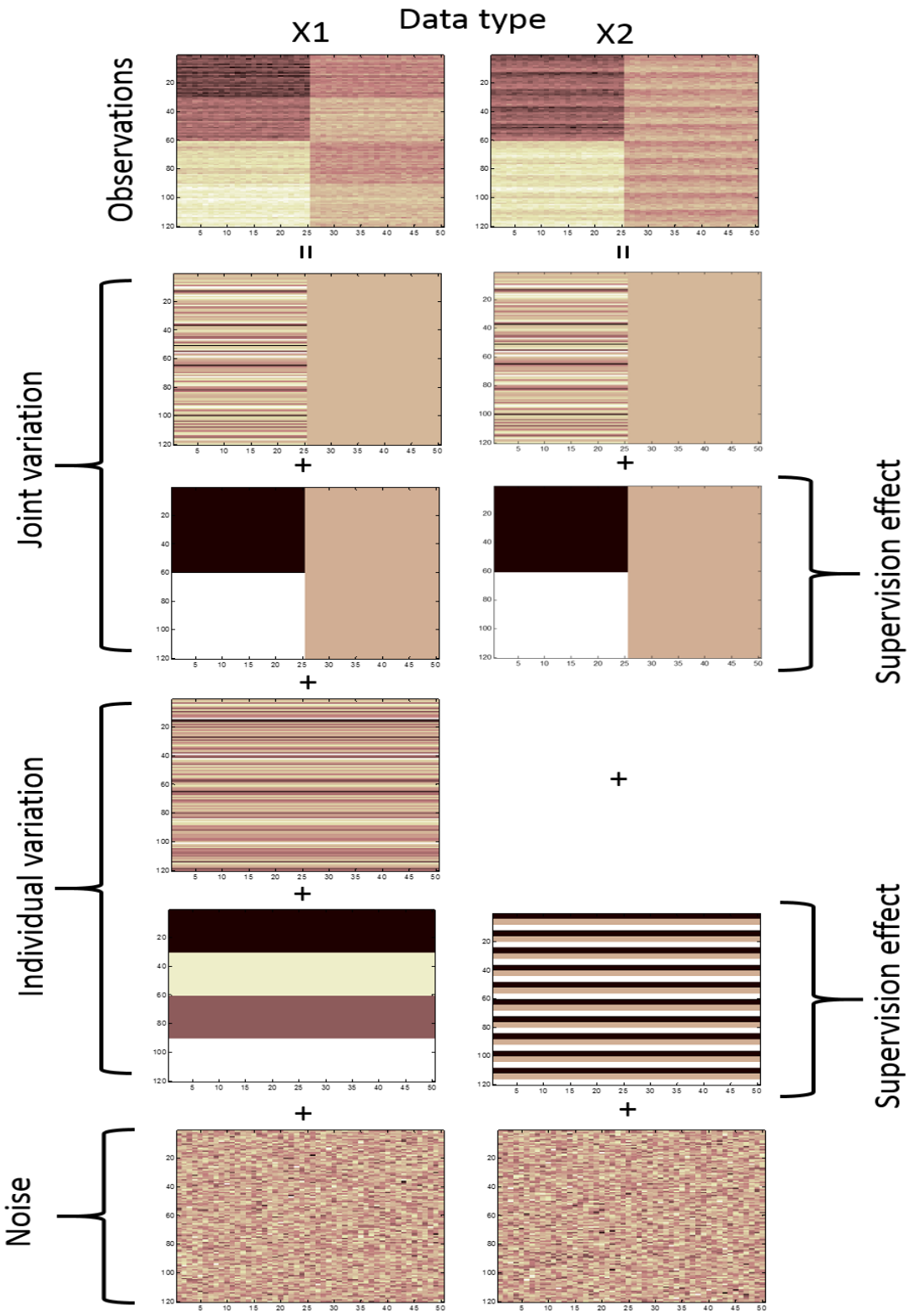

Figure 12. Population structure of an illustrative example.

into likelihood framework, we impose some distributional assumptions on the error matrices. We assume entries of $\mathbf{E}_i$ are i.i.d. from $N(0, \sigma_{e_i}^2)$. Moreover, we assume the random matrix $\mathbf{F}$ (or $\mathbf{F}_i$) has i.i.d. rows from multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma}_f)$ (or $N(\mathbf{0}, \mathbf{\Sigma}_{f_i})$). For identifiability consideration in the SupJIVE part, we require columns of $\mathbf{V}$ (or $\mathbf{V}_i$) to be orthonormal and $\mathbf{\Sigma}_f$ (or $\mathbf{\Sigma}_{f_i}$) to be diagonal with positive distinct values. Furthermore, referring to the JIVE model, we require the column space of $\mathbf{J}$ to be orthogonal with the column space of $\mathbf{A}_i$.

4.2.2 Illustrative example

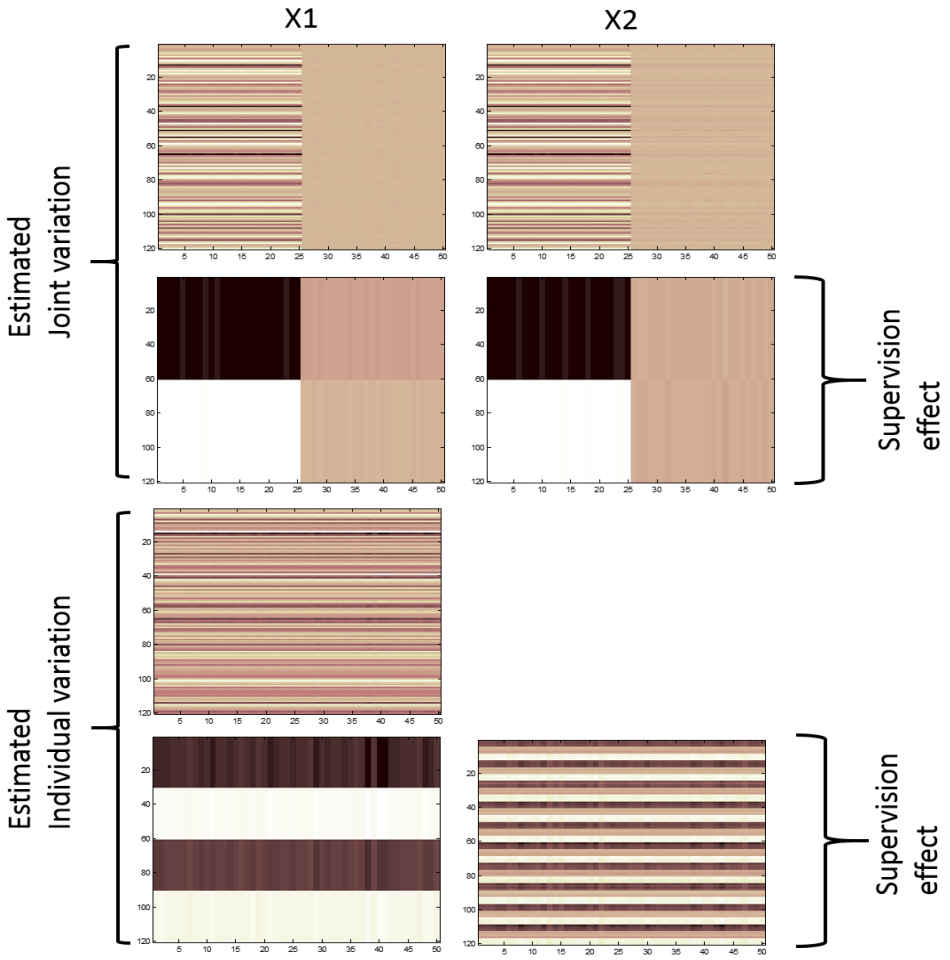

Figure 13. Estimates for the illustrative example.

To illustrate the type of data analysis SupJIVE aims to conduct, we generate two matrices, $\mathbf{X}_1$ and $\mathbf{X}_2$, with patterns corresponding to joint and individual variations which are partially

associated with different groupings. The simulated data are depicted in Figure 12. Both $\mathbf{X}_1$ and $\mathbf{X}_2$ are of dimension 120×50, that is, each has 50 variables measured for the same 120 objects. Joint variation is two-fold. First, 120 independent standard normal variables are added to half of the columns in $\mathbf{X}_1$ and in $\mathbf{X}_2$. The second joint variation is associated with two supervision groups. Those rows corresponding to two different groups have respective -7 and 7 added to their columns in a similar manner as for the first joint variation. Individual variation in $\mathbf{X}_1$ is also two-fold. On top of 120 independent standard normal variables added to the columns of $\mathbf{X}_1$, -3, 1, -2 and 2 are added to the columns of the rows corresponding to 4 different groups, respectively. The second data set $\mathbf{X}_2$ has three group effects but groups are shuffled in the repetitive manner as seen in Figure 12. Finally, independent standard normal variables are added to elements of $\mathbf{X}_1$ and $\mathbf{X}_2$. Both joint and individual variations are visually obscured.

Figure 13 shows SupJIVE estimates for joint and individual variations, each divided into supervised and unsupervised parts. It clearly separates variations associated and unassociated with corresponding groups in each of joint and individual structures.

4.2.3 Estimation

To estimate the SupJIVE model parameters, we adopt an iterative procedure similar with JIVE. Assume the ranks for joint structure and individual structures are known. For fixed individual structures $\mathbf{A}_i$, $i = 1, \cdots, m$, we calculate $\mathbf{X}^\star = (\mathbf{X}_1 - \mathbf{A}_1, \mathbf{X}_2 - \mathbf{A}_2, \cdots, \mathbf{X}_m - \mathbf{A}_m)$ and apply SupSVD algorithm to $\mathbf{X}^\star$ to get a prediction for $\mathbf{J}$. Then, for the fixed joint structure $\mathbf{J}$, we apply SupSVD algorithm to the projection of $\mathbf{X}_i - \mathbf{J}_i$ onto the orthogonal space of col($\mathbf{U}$) and find the individual structure $\mathbf{A}_i$ for $i = 1, 2, \cdots, m$. Iterate the two steps until convergence. For details of SupSVD and JIVE algorithms, refer to their individual papers. We summarize the estimation procedure in the next page. In practice, one needs to estimate the ranks for the joint structure and all individual structures. We use the permutation-test based rank estimation method described in JIVE [35].

4.2.4 Potential issues of SupJIVE

Before explaining potential issues of the algorithm, we classify joint variations into two types. First, the full joint variation is referred to as a pattern shared in the whole data sets. Second, the partial joint variation is a pattern shared in the more than one but not the whole data sets. Graphical

1: Initial step;

(1) Apply SupSVD to $\mathbf{X}$ to get estimates $\mathbf{V}^{[1]}$, $\mathbf{B}^{[1]}$, $\mathbf{F}^{[1]}$, $\mathbf{U}^{[1]}$ and $\mathbf{J}^{[1]}$.

(2) Apply SupSVD to $\mathbf{P}^{\perp}_{\mathbf{U}^{[1]}}(\mathbf{X}_i - \mathbf{J}_i^{[1]})$, where $\mathbf{P}^{\perp}_{\mathbf{U}^{[1]}} = \mathbf{I} - \mathbf{U}^{[1]}\left((\mathbf{U}^{[1]})^T\mathbf{U}^{[1]}\right)^{-1}(\mathbf{U}^{[1]})^T$ to get estimates $\mathbf{V}_i^{[1]}$, $\mathbf{B}_i^{[1]}$, $\mathbf{F}_i^{[1]}$ and $\mathbf{A}_i^{[1]}$ for $i = 1, 2, \cdots, m$.

(3) Store $\mathbf{X}^{[1]} = [\mathbf{J}_1^{[1]} + \mathbf{A}_1^{[1]}, \mathbf{J}_2^{[1]} + \mathbf{A}_2^{[1]}, \cdots, \mathbf{J}_m^{[1]} + \mathbf{A}_m^{[1]}]$.

2: sth step;

(1) Fix $\mathbf{A}_i^{[s-1]}$ and calculate $\mathbf{X}^{*[s]} = [\mathbf{X}_1 - \mathbf{A}_1^{[s-1]}, \mathbf{X}_2 - \mathbf{A}_2^{[s-1]}, \cdots, \mathbf{X}_m - \mathbf{A}_m^{[s-1]}]$.

(2) Apply SupSVD to $\mathbf{X}^{*[s]}$ to get $\mathbf{V}^{[s]}$, $\mathbf{B}^{[s]}$, $\mathbf{F}^{[s]}$, $\mathbf{U}^{[s]}$ and $\mathbf{J}^{[s]}$.

(3) Fix $\mathbf{J}^{[s]}$ and estimate $\mathbf{A}_i^{[s]}$ through SupSVD of $\mathbf{P}^{\perp}_{\mathbf{U}^{[s]}}(\mathbf{X}_i - \mathbf{J}_i^{[s]})$, where $\mathbf{P}^{\perp}_{\mathbf{U}^{[s]}} = \mathbf{I} - \mathbf{U}^{[s]}\left((\mathbf{U}^{[s]})^T\mathbf{U}^{[s]}\right)^{-1}(\mathbf{U}^{[s]})^T$ to get estimates $\mathbf{V}_i^{[s]}$, $\mathbf{B}_i^{[s]}$, $\mathbf{F}_i^{[s]}$ and $\mathbf{A}_i^{[s]}$ for $i = 1, 2, \cdots, m$.

(4) Store $\mathbf{X}^{[s]} = [\mathbf{J}_1^{[s]} + \mathbf{A}_1^{[s]}, \mathbf{J}_2^{[s]} + \mathbf{A}_2^{[s]}, \cdots, \mathbf{J}_m^{[s]} + \mathbf{A}_m^{[s]}]$.

3: Repeat until $|\mathbf{X}^{[s]} - \mathbf{X}^{[s-1]}| < e$ for some predetermined tolerance e.

description for a 3-source data set is shown in Figure 14. Now we describe problems SupJIVE suffer that render it rather impractical,

1. Need to estimate ranks for joint and individual variations prior to an analysis. There has been no universally good answers to rank selection problem. There exist several methods to calculate ranks for a given matrix such as a permutation-test based method used in JIVE [35] and a simple scree-plot method commonly used in PCA, but these methods suffer from either heavy computations or subjectivity. Failure in calculating ranks may lead to a lower performance of the SupJIVE algorithm. For example, uncaught dimensions in joint variation may turn into an individual variation component.

2. Need to select supervision specific to each of joint and individual variations. Sometimes we do not know which supervision might drive variation across all data sets or in each of them.

3. Current SupJIVE model (4.3) considers a full joint variation but it not capable of correctly modeling partial joint variations.

We could include all individual, partial and full joint variations in the SupJIVE model. However, exponentially increasing workload of calculating rank and selecting supervisions for each variation would make using this model algorithm rather impractical as the number of data sets increases.

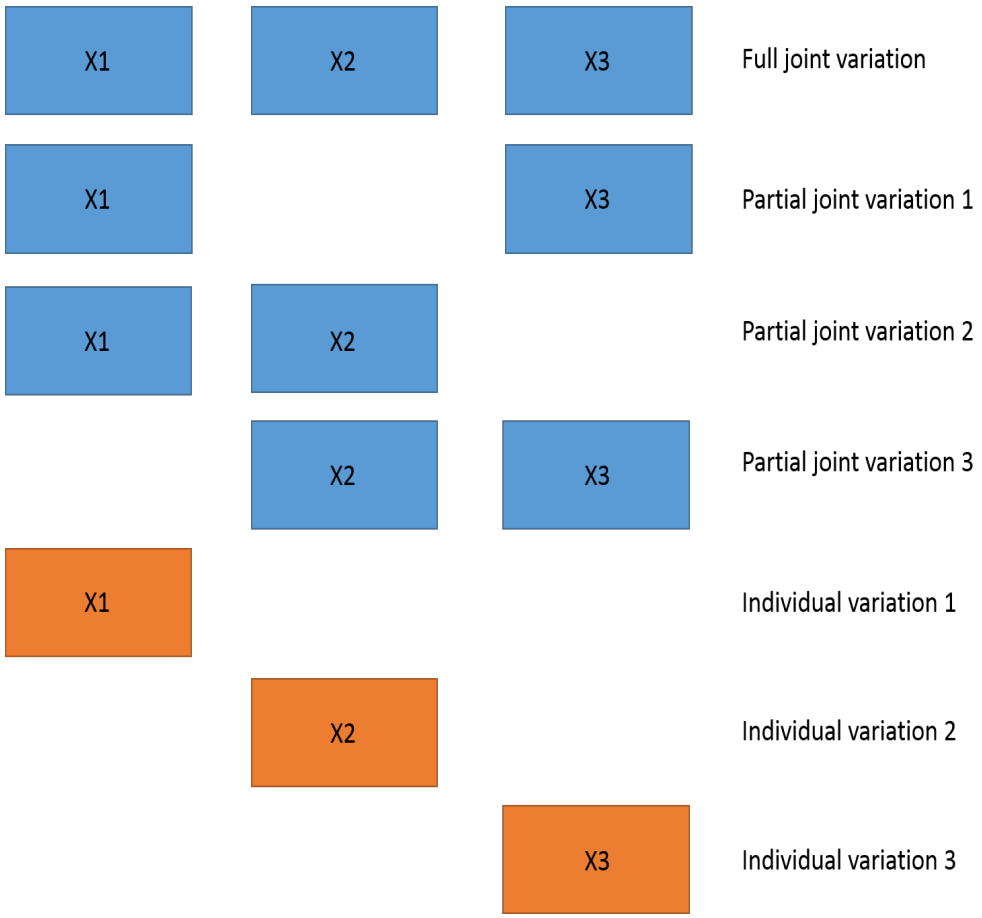

Figure 14. Joint and individual variations for three data sets.

4.3 GENERALIZED SUPJIVE

In this section, we propose a new algorithm, called Generalized Supervised Joint and Individual Variations Explained (G-SupJIVE), that overcomes the issues of SupJIVE described in the previous section.

The G-SupJIVE model much more flexible than any of SupJIVE 4.2, JIVE [35], SupSVD [32] or SIFA [31]. In particular, there is no restriction on the direction vectors of $\mathbf{V}$ so that any of individual, partial or full joint variation components is freely modeled.

To estimate the parameters of G-SupJIVE model, we propose a layer-by-layer algorithm estimating a pair of a variation direction and its supervision effect at a time until the multi-source data set exhausts its rank. The main merit of this algorithm is that it automatically 1) chooses

whether the variation estimated at each layer is full, partial or individual, 2) selects its supervision from multiple supervision sets and 3) stops when rank is exhausted.

4.3.1 Population model

The G-SupJIVE model is a fully automated data-driven model for the integrated analysis of multi-block data. Specifically, the population model is,

$$\begin{cases} \mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m], \\ \mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_k], \\ \mathbf{X} = (\mathbf{YB} + \mathbf{F})\mathbf{V}^T + \mathbf{E}. \end{cases} \tag{4.4}$$

Here $\mathbf{X}_i$, $i = 1, 2, .., m$, is a $n \times p_i$ component data matrix collected from possibly a distinct source; $\mathbf{Y}_j$, $j = 1, 2, .., k$, is a $n \times q_j$ matrix that represents different possible supervision information; the $(p_1 + p_2 + \cdots + p_m) \times r$ matrix, $\mathbf{V}$, contains r directions of variations arranged column-wise; the $(q_1 + q_2 + \cdots + q_k) \times r$ matrix, $\mathbf{B}$, is a conversion matrix that translates supervision information in $\mathbf{Y}$ into scores for column vectors in $\mathbf{V}$; the $n \times (q_1 + q_2 + \cdots + q_m)$ matrix $\mathbf{F}$ contributes to the variability unrelated to supervision. Finally, the $n \times (p_1 + p_2 + \cdots + p_m)$ matrix, $\mathbf{E}$, includes noise.

Unlike the SupJIVE models (4.3), the G-SupJIVE model (4.4) does not separate joint and individual variations a priori. We desire estimates of variation directions (columns of $\mathbf{V}$) to automatically catch block-wise structures, i.e., full, partial or individual variations and also desire supervision effects related to each variation direction to be selected from the candidate supervision collection $\mathbf{Y}$ in a data-driven fashion. To this end, we impose the group-wise sparsity condition on columns of $\mathbf{B}$ and $\mathbf{V}$ corresponding to the blocks given by $\mathbf{Y}_j$'s and $\mathbf{X}_i$'s, respectively. Without loss of generality, we assume that both $\mathbf{X}$ and $\mathbf{Y}$ are column-centered so that the model does not have intercepts. The random matrices $\mathbf{E}$ and $\mathbf{F}$ are assumed to be independent with each other. Each entry of the error matrix $\mathbf{E}$ is independent and identically distributed (i.i.d.) with mean zero and variance σ_ϵ^2. Each row of $\mathbf{F}$ is i.i.d. with mean zero and covariance matrix $\mathbf{\Sigma}_f$.

For the model (4.4) to be identifiable in terms of parameters $\mathbf{V}, \mathbf{B}, \mathbf{\Sigma}_f$ and σ_ϵ^2, we adopt the following constraints from the SupSVD model [32] as the model (4.4) generalizes SupSVD to multi-source data,

1. The matrix $\mathbf{V}$ has orthonormal columns.
2. The matrix $\mathbf{\Sigma}_f$ is diagonal with r distinct positive entries.

3. The columns of $\mathbf{V}$ are sorted in the descending order in terms of the variances of $\mathbf{YB} + \mathbf{F}$ and the first entry of each column is positive.
4. The supervision data matrix $\mathbf{Y}$ has linearly independent columns.

Under these conditions, the G-SupJIVE is identifiable.

4.3.2 Illustrative example

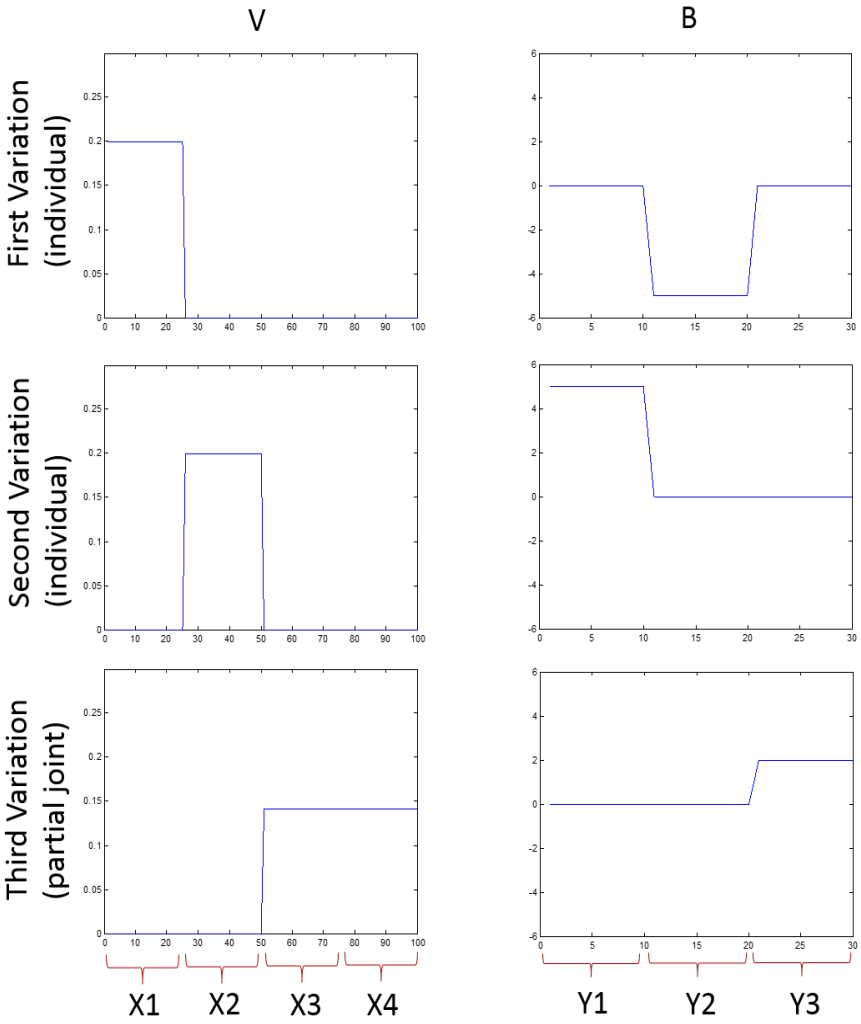

Figure 15. Popuation $\mathbf{V}$ and $\mathbf{B}$

To illustrate what G-SupJIVE does, we provide an intuitive example. The data we simulate has the following structure,

$$\underset{120\times100}{\mathbf{X}} = [\underset{120\times25}{\mathbf{X}_1}, \underset{120\times25}{\mathbf{X}_2}, \underset{120\times25}{\mathbf{X}_3}, \underset{120\times25}{\mathbf{X}_4}] = \underset{120\times3}{\mathbf{U}}\ \underset{3\times100}{\mathbf{V}^T} + \underset{120\times100}{\mathbf{E}},$$

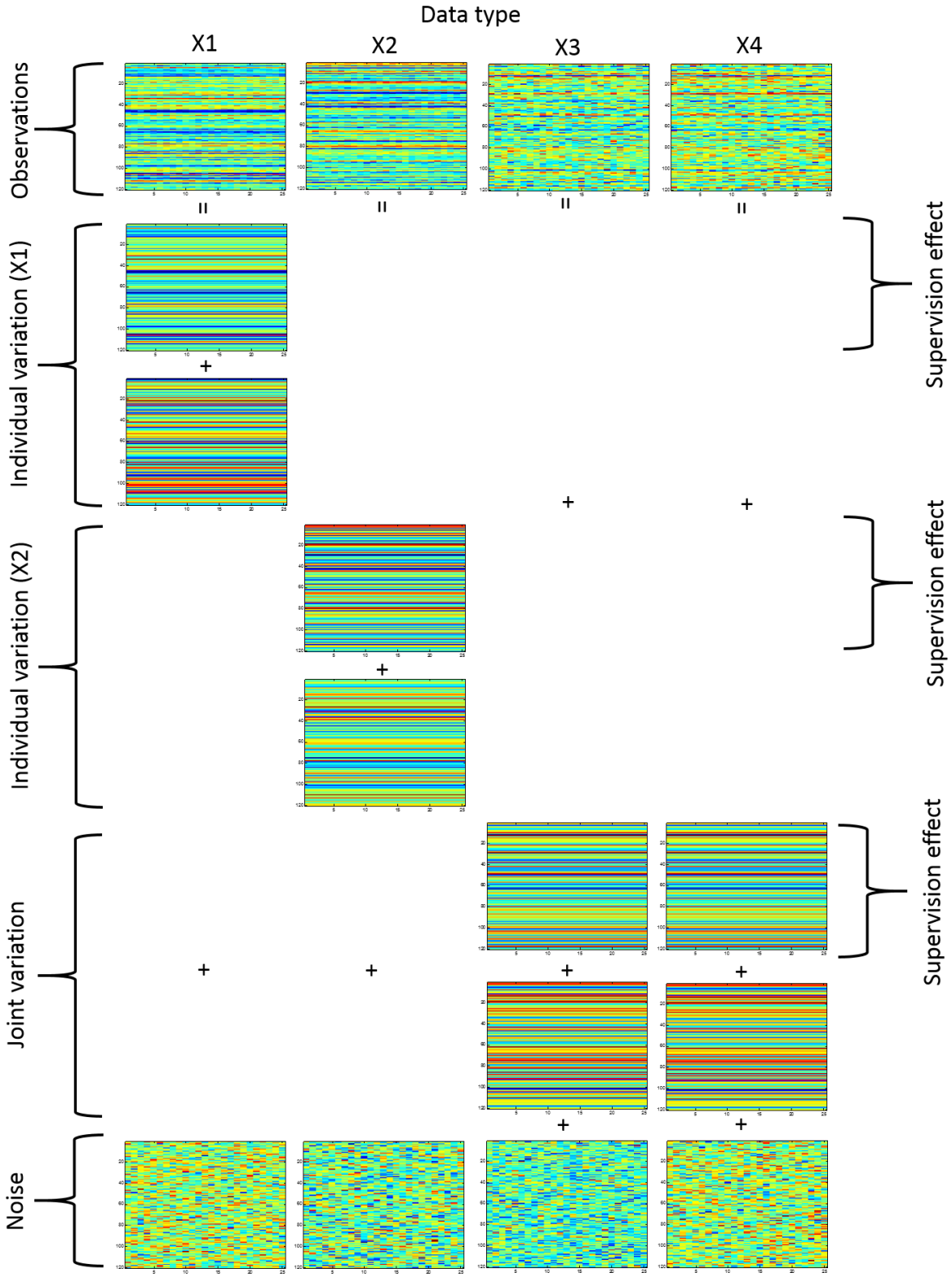

Figure 16. Heatmap of the simulate data.

$$\underset{120\times 30}{\mathbf{Y}} = [\underset{120\times 10}{\mathbf{Y}_1}, \underset{120\times 10}{\mathbf{Y}_2}, \underset{120\times 10}{\mathbf{Y}_3}],$$

$$\underset{120\times 3}{\mathbf{U}} = \underset{120\times 30}{\mathbf{Y}} \underset{30\times 3}{\mathbf{B}} + \underset{120\times 3}{\mathbf{F}}.$$

The multi-source data set $\mathbf{X}$ has 120 observations and consists of 4 subsets $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$, each having 25 measurements. The supervision set $\mathbf{Y}$ collects 3 supervision candidates $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$, where each column of $\mathbf{Y}_i$ is filled with 120-dimensional normal random vector with mean $\mathbf{0}$ and diagonal covariance matrix $\sigma_i^2\mathbf{I}$ for $\sigma_i^2 = 2, 1.5, 1$. The rows of the matrix $\mathbf{F}$, which stands for unsupervised effects, are filled with 3-dimensional standard normal with mean $\mathbf{0}$ and diagonal covariance matrix with entries of 6, 4 and 2. The noise $\mathbf{E}$ of element-wise independent standard normal is added.

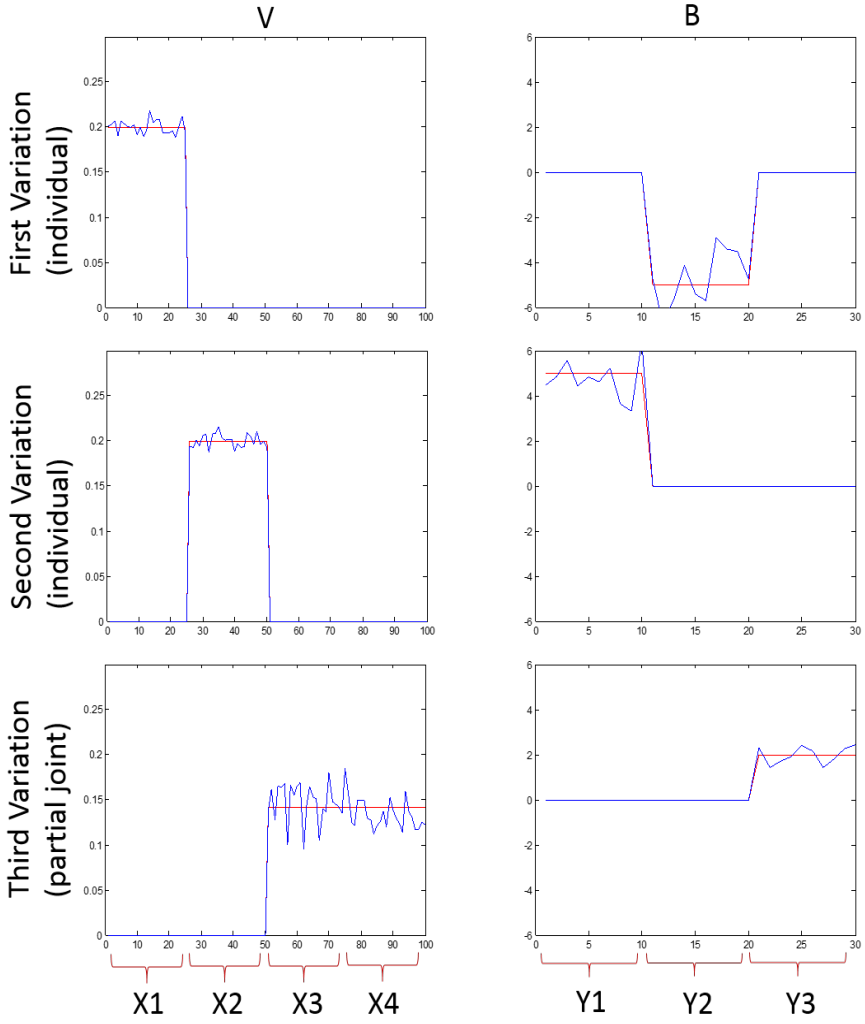

Figure 17. Estimation of the parameters.

The populations $\mathbf{V}$ and $\mathbf{B}$ are depicted in Figure 15, where each row represents a pair of columns

of $\mathbf{V}$ and $\mathbf{B}$. The matrix $\mathbf{V}$ provides three variation directions, resulting in the data set $\mathbf{X}$'s intrinsic rank being three. The first variation (first column of $\mathbf{V}$) is an individual variation only specific to the data set $\mathbf{X}_1$ and its supervision effect stems from the second supervision candidate set $\mathbf{Y}_2$ as the first column of $\mathbf{B}$ indicates. The second variation has a similar interpretation. The third variation is a partial joint variation that covers the third and last data sets $\mathbf{X}_3$ and $\mathbf{X}_4$. Its supervision comes from the third supervision candidate set $\mathbf{Y}_3$.

The heatmap of the simulated data and their decomposed variations are depicted in Figure 16. The structure of the data is much more complicated than and not as visually clear as that in Figure 12 since their supervisions are continuous variables on the contrary to categorical supervision in Figure 12.

As a preprocessing, we column-center the data by subtracting the mean within each column to remove baseline differences between data sets. To circumvent cases where 'the largest data set wins', we scale each data set by its total variation, i.e., each data set's Frobenius norm. The G-SupJIVE model is then fit by the algorithm we discuss later in Section 4.3.3. Estimation results are shown in Figure 17, where estimates in blue and parameters in red are overlaid. Overall, the G-SupJIVE algorithm effectively captures the variation patterns and supervisions that drive them.

4.3.3 Estimation

We adopt a sequential approach that estimates parameters layer by layer. In specific, we estimate the first column $\mathbf{v}_1$ of $\mathbf{V}$ in the model (4.4) and then move to the subspace orthogonal to $\mathbf{v}_1$ to estimate the next column of $\mathbf{V}$. In each layer, we obtain a penalized maximum likelihood using a group Lasso penalty [55] in estimations of $\mathbf{V}$ and $\mathbf{B}$. The groups for the penalty are naturally defined from the structure of the multi-block data set and the supervision candidate data set.

A description of the G-SupJIVE estimation procedure is summarized in the next page. Here included are the steps for the first layer. The prime advantages of the proposed algorithm are two-fold,

1. Adaptively choose the individual, partial and full joint variations and corresponding supervision effects: no need to specify the types of variation.
2. Automatically stops when the rank of $\mathbf{X}$ is exhausted: no pre-calculation of ranks.

Now we give a detailed explanation of the proposed algorithm. Consider the following rank 1

:Repeat the following steps for ith layer to get estimates $\hat{\mathbf{v}}_i, \hat{\mathbf{b}}_i, \hat{\sigma}^2_{f_i}, \hat{\sigma}^2_\epsilon$ until $\hat{\mathbf{v}}_i = \mathbf{0}$ or $\hat{\mathbf{b}}_i = \mathbf{0}$. Assume the data matrix $\mathbf{X}$ is of rank 1 and form a penalized likelihood Q as a function of parameters $\mathbf{v}_1, \mathbf{b}_1, \sigma^2_{f_1}, \sigma^2_\epsilon$.

1: Initial step;

(1) Apply rank-one SVD to $\mathbf{X}$, i.e. $\mathbf{X} \approx \lambda \mathbf{u}\mathbf{v}^T$ to get estimates $\mathbf{v}_1^{[0]}$.

(2) Get an estimate $\sigma_\epsilon^{2[0]}$ from residuals of $\mathbf{X} - \lambda \mathbf{u}(\mathbf{v}_1^{[0]})^T$.

(3) Get an estimate $\mathbf{b}_1^{[0]}$ from regression $\lambda \mathbf{u} = \mathbf{Y}\mathbf{b}$.

(4) Get an estimate $\sigma_{f_1}^{2[0]}$ from residuals of $\lambda \mathbf{u} - \mathbf{Y}\mathbf{b}_1^{[0]}$.

(5) Calcualte $Q^{[0]}$ with arguments $\mathbf{v}_1^{[0]}, \mathbf{b}_1^{[0]}, \sigma_{f_1}^{2[0]}, \sigma_\epsilon^{2[0]}$.

2: sth step;

(1) Maximize $Q^{[s-1]}$ as a function of $\mathbf{v}_1$ with a group Lasso penalty to $\mathbf{v}_1$ to get a maximizer $\mathbf{v}_1^{[s]}$.

(2) Let $Q_{v_1}^{[s-1]}$ be $Q^{[s-1]}$ where $\mathbf{v}_1^{[s-1]}$ is updated with $\mathbf{v}_1^{[s]}$. Maximize $Q_v^{[s-1]}$ as a function of $\mathbf{b}_1$ with a group Lasso penalty to $\mathbf{b}_1$ to get a maximizer $\mathbf{b}_1^{[s]}$.

(3) Let $Q_{v_1,b_1}^{[s-1]}$ be $Q^{[s-1]}$ where $\mathbf{v}_1^{[s-1]}, \mathbf{b}_1^{[s-1]}$ are updated with $\mathbf{v}_1^{[s]}, \mathbf{b}_1^{[s]}$. Maximize $Q_{v,b}^{[s-1]}$ as a function of $\sigma^2_{f_1}, \sigma^2_\epsilon$ to get a maximizer $\sigma_{f_1}^{2[s]}, \sigma_\epsilon^{2[s]}$.

(4) Let $Q^{[s]}$ be $Q^{[s-1]}$ where all previous arguments are replaced with new updates $\mathbf{v}_1^{[s]}, \mathbf{b}_1^{[s]}, \sigma_{f_1}^{2[s]}, \sigma_\epsilon^{2[s]}$.

3: Repeat until $Q^{[s]} - Q[s-1] < e$ for some predetermined tolerance e to get final estimates $\hat{\mathbf{v}}_1, \hat{\mathbf{b}}_1, \hat{\sigma}^2_{f_1}$.

4: For the next layer, find the projections onto the complement subspace $\mathbf{X}^{[1]} = \mathbf{X} - \mathbf{X}\hat{\mathbf{v}}_1$ and apply the previous steps 1 - 3 to $\mathbf{X}^{[1]}$ to estimate $\mathbf{v}_2, \mathbf{b}_2, \sigma^2_{f_2}$ and update the estimate of σ^2_ϵ.

model for $\mathbf{X}$,

$$\underset{n\times p}{\mathbf{X}} = (\underset{n\times q}{\mathbf{Y}}\ \underset{q\times 1}{\mathbf{b}_1} + \underset{n\times 1}{\mathbf{f}_1})\ \underset{1\times p}{\mathbf{v}_1^T} + \underset{n\times p}{\mathbf{E}}.$$

We assume that the matrix $\mathbf{X}$ consists of m data sets, each with dimensions p_i ($p_1 + p_2 + \cdots + p_m = p$). Accordingly, $\mathbf{v}_1^T$ is partitioned with the corresponding blocks and we write $\mathbf{v}_1^T = [\mathbf{v}_{1,G_1}^T, \mathbf{v}_{1,G_2}^T, \ldots, \mathbf{v}_{1,G_k}^T]$ where $\mathbf{v}_{1,G_i}$ is a vector of variables belonging to the group i. We slao assume that the matrix $\mathbf{Y}$ consists of k supervision candidate data sets, each with dimensions q_j ($q_1+q_2+\cdots+q_k = q$). Accordingly, $\mathbf{b}_1^T$ is partitioned with the corresponding blocks and we write $\mathbf{b}_1^T = [\mathbf{b}_{1,G_1}^T, \mathbf{b}_{1,G_2}^T, \ldots, \mathbf{b}_{1,G_m}^T]$ where $\mathbf{b}_{1,G_j}$ is a vector of variables belonging to the group j. Then

each observation row $\mathbf{x}_i$ of $\mathbf{X}$ are i.i.d. from a multivariate normal with the following parameters,

$$E(\mathbf{x}_i) = \mathbf{y}_i \mathbf{b}_1 \mathbf{v}_1^T, \quad \mathbf{\Sigma}_{\mathbf{x}_i} = \sigma_e^2 \mathbf{I}_p + \sigma_{f_1}^2 \mathbf{v}_1 \mathbf{v_1}^T.$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are i rows of $\mathbf{X}$ and $\mathbf{Y}$ respectively. Then the log likelihood function is,

$$\begin{aligned} &\log \mathrm{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \mid \mathbf{b}_1, \mathbf{v}_1, \sigma_e^2, \sigma_{f_1}^2) \\ &\quad = -\frac{n}{2} \log\left(2\pi \det(\mathbf{\Sigma}_{\mathbf{x}_1})\right) - \frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{x}_i - \mathbf{y}_i \mathbf{b}_1 \mathbf{v}_1^T\right) \mathbf{\Sigma}_{\mathbf{x}_1}^{-1} \left(\mathbf{x}_i - \mathbf{y}_i \mathbf{b}_1 \mathbf{v}_1^T\right)^T. \end{aligned} \tag{4.5}$$

We state two lemmas that will be used to make the log likelihood function (4.5) more tractable for optimization purpose.

Lemma 16 (Determinant identity.)**.** *[17] If* $\mathbf{A}$ *is an invertible square matrix and* $\mathbf{u}$ *and* $\mathbf{v}$ *are column vectors, then,*

$$Det(\mathbf{A} + \mathbf{u}\mathbf{v}^T) = Det(\mathbf{A})(1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{u}).$$

Lemma 17 (Inverse matrix identity.)**.** *[17]. If* $\mathbf{A}$ *and* $\mathbf{A} + \mathbf{B}$ *are invertible and* $\mathbf{B}$ *has a rank 1, then let* $g = trace(\mathbf{B}\mathbf{A}^{-1})$*, then* $g \neq -1$ *and*

$$(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \frac{1}{1+g} \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}.$$

By Lemma 16,

$$\begin{aligned} \det(\mathbf{\Sigma}_{\mathbf{x}_1}) &= \det(\sigma_e^2 \mathbf{I}_p + \sigma_{f_1}^2 \mathbf{v}_1 \mathbf{v_1}^T) \\ &= \det(\sigma_e^2 \mathbf{I}_p) \det(1 + \mathbf{v}_1^T (\sigma_{f_1}^2 \mathbf{I}_p)^{-1} \mathbf{v}_1) \\ &= (\sigma_e^2)^{p-1} (\sigma_{f_1}^2 + \sigma_e^2). \end{aligned}$$

By Lemma 17,

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{x}_1}^{-1} &= \frac{1}{\sigma_e^2} \mathbf{I}_p + \left(\frac{1}{\sigma_{f_1}^2 + \sigma_e^2} - \frac{1}{\sigma_e^2} \right) \mathbf{v}_1 \mathbf{v_1}^T \\ &= \frac{1}{\sigma_e^2} \mathbf{I}_p - \frac{\sigma_{f_1}^2}{\sigma_e^2 (\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{v}_1 \mathbf{v_1}^T. \end{aligned}$$

So the summands in the summation part in the log likelihood function (4.5) become,

$$\begin{aligned} &\left(\mathbf{x}_i - \mathbf{y}_i \mathbf{b}_1 \mathbf{v}_1^T\right) \mathbf{\Sigma}_{\mathbf{x}_1}^{-1} \left(\mathbf{x}_i - \mathbf{y}_i \mathbf{b}_1 \mathbf{v}_1^T\right)^T \\ &\quad = \frac{1}{\sigma_e^2} \left(\mathbf{x}_i - \mathbf{y}_i \mathbf{b}_1 \mathbf{v}_1^T\right) \left(\mathbf{x}_i - \mathbf{y}_i \mathbf{b}_1 \mathbf{v}_1^T\right)^T - \frac{\sigma_{f_1}^2}{\sigma_e^2 (\sigma_{f_1}^2 + \sigma_e^2)} \left(\mathbf{x}_i \mathbf{v}_1 - \mathbf{y}_i \mathbf{b}_1\right) \left(\mathbf{x}_i \mathbf{v}_1 - \mathbf{y}_i \mathbf{b}_1\right)^T. \end{aligned}$$

Hence the summation part in (4.5) becomes,

$$\sum_{i=1}^{n} \left(\mathbf{x}_i - \mathbf{y}_i \mathbf{b}_1 \mathbf{v}_1^T\right) \mathbf{\Sigma}_{\mathbf{x}_1}^{-1} \left(\mathbf{x}_i - \mathbf{y}_i \mathbf{b}_1 \mathbf{v}_1^T\right)^T = \frac{1}{\sigma_e^2} \left\|\mathbf{X} - \mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2 - \frac{\sigma_{f_1}^2}{\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \left\|\mathbf{X}\mathbf{v}_1 - \mathbf{Y}\mathbf{b}_1\right\|_2^2 .$$

Finally the log likelihood (4.5) becomes,

$$\begin{aligned}
&\log \mathrm{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \mid \mathbf{b}_1, \mathbf{v}_1, \sigma_e^2, \sigma_{f_1}^2) \\
&= -\frac{n}{2} \log(2\pi \, (\sigma_e^2)^{p-1} (\sigma_{f_1}^2 + \sigma_e^2)) - \frac{1}{2\sigma_e^2} \left\|\mathbf{X} - \mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2 + \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \left\|\mathbf{X}\mathbf{v}_1 - \mathbf{Y}\mathbf{b}_1\right\|_2^2 .
\end{aligned}$$

Imposing group Lasso penalties to the likelihood function above, we maximize the following likelihood function Q for some tuning parameters $\lambda > 0$ and $\gamma > 0$,

$$\begin{aligned}
&Q(\mathbf{v}_1, \mathbf{b}_1, \sigma_{f_1}^2, \sigma_\epsilon^2 | \lambda, \gamma) \\
&= \log \mathrm{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \mid \mathbf{b}_1, \mathbf{v}_1, \sigma_e^2, \sigma_{f_1}^2) - \sum_{i=1}^{m} \lambda \|\mathbf{v}_{1,G_i}\|_2 - \sum_{j=1}^{k} \gamma \|\mathbf{b}_{1,G_j}\|_2 \\
&= -\frac{n}{2} \log(2\pi \, (\sigma_e^2)^{p-1} (\sigma_{f_1}^2 + \sigma_e^2)) - \frac{1}{2\sigma_e^2} \left\|\mathbf{X} - \mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2 \\
&\quad + \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \left\|\mathbf{X}\mathbf{v}_1 - \mathbf{Y}\mathbf{b}_1\right\|_2^2 - \sum_{i=1}^{m} \lambda \|\mathbf{v}_{1,G_i}\|_2 - \sum_{j=1}^{k} \gamma \|\mathbf{b}_{1,G_j}\|_2 .
\end{aligned} \tag{4.6}$$

Since the maximizer of 4.6 has no closed form, we resort to an iterative algorithm described next. The algorithm is presented in the following 4 steps (① - ④) that are consistent with steps in the G-SupJIVE algorithm.

① Initial estimates for $\mathbf{b}_1, \mathbf{v}_1, \sigma_{f_1}^2, \sigma_e^2$

Find the rank-1 approximation of $\mathbf{X}$ via a singular value decomposition

$$\mathbf{X} \approx \mathbf{u}\mathbf{v}^T,$$

where $\mathbf{u}$ is the vector of the product of the first left singular vector and the first singular value. We set $\mathbf{v}_1^{[0]} = \mathbf{v}$. Treat $\mathbf{X} - \mathbf{u}\mathbf{v}^T$ as a random matrix with i.i.d. entries from normal distribution with mean 0 and variance σ_e^2 and set the initial value $\sigma_e^{2[0]}$ for σ_e^2 as the sample variance of the entries. Then we regress $\mathbf{u}$ on $\mathbf{Y}$ by assuming that the residuals $\mathbf{f}_1$ are i.i.d. normal random variables with mean 0 and variance $\sigma_{f_1}^2$.

$$\mathbf{u} = \mathbf{Y}\mathbf{b} + \mathbf{f}_1.$$

Then we set the initial values $\mathbf{b}_1^{[0]}$ and $\sigma_{f_1}^{2[0]}$ as,

$$\mathbf{b}_1^{[0]} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{u}, \quad \sigma_{f_1}^{2[0]} = \frac{\|\mathbf{u} - (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{u}\|_2^2}{n-q}.$$

For reference, we compute the value of the penalized likelihood at the 0th iteration and denote it by $Q^{[0]}$,

$$Q^{[0]} = \text{logP}(\mathbf{X} \mid \mathbf{b}_1^{[0]}, \mathbf{v}_1^{[0]}, \sigma_e^{2[0]}, \sigma_{f_1}^{2[0]}) - \sum_{i=1}^{m} \lambda \left\| \mathbf{v}_{1,G_i}^{[0]} \right\|_2 - \sum_{j=1}^{k} \gamma \left\| \mathbf{b}_{1,G_j}^{[0]} \right\|_2.$$

② The sth step

Let $\mathbf{b}_1^{[s-1]}, \mathbf{v}_1^{[s-1]}, \sigma_{f_1}^{2[s-1]}$ and $\sigma_e^{2[s-1]}$ be the updates from the previous iteration. To update $\mathbf{b}_1^{[s-1]}$, we maximize the penalized likelihood 4.6 as a function of $\mathbf{b}_1$ given $\mathbf{v}_1^{[s-1]}, \sigma_e^{2[s-1]}$ and $\sigma_{f_1}^{2[s-1]}$. We show in Section 4.8.1 that this work is equivalent to minimizing the following,

$$A^{[s]}(\mathbf{b}_1) = \left\| \sqrt{\frac{1}{2\sigma_e^{2[s-1]}}} \mathbf{\Phi}_1 - \mathbf{\Phi}_2 \mathbf{b}_1 \right\|_2^2 + \sum_{j=1}^{k} \gamma \left\| \mathbf{b}_{1,G_j} \right\|_2, \tag{4.7}$$

where

$$\mathbf{\Phi}_1 = \begin{pmatrix} \mathbf{X}_{\cdot 1} \\ \mathbf{X}_{\cdot 2} \\ \vdots \\ \mathbf{X}_{\cdot p} \end{pmatrix}, \quad \mathbf{\Phi}_2^{[s-1]} = \begin{pmatrix} \sqrt{\frac{\mathbf{v}_{1,1}^{[s-1]}}{2\sigma_e^{2[s-1]}}} \mathbf{Y} \\ \sqrt{\frac{\mathbf{v}_{1,2}^{[s-1]}}{2\sigma_e^{2[s-1]}}} \mathbf{Y} \\ \vdots \\ \sqrt{\frac{\mathbf{v}_{1,p}^{[s-1]}}{2\sigma_e^{2[s-1]}}} \mathbf{Y} \end{pmatrix}$$

Here $\mathbf{X}_{\cdot 1}$ is the ith column of $\mathbf{X}$. The form 4.7 is a regression problem with a group Lasso penalty. The minimizer $\hat{\mathbf{b}}_1$ of 4.7 becomes the next iterate $\mathbf{b}_1^{[s]}$. There are a couple of group Lasso implementations available. We chose the SLEP package [33] for its extensive coverage of various other penalties.

Now $\mathbf{v}_1^{[s-1]}$ needs to be updated. We maximize the penalized likelihood 4.6 as a function of $\mathbf{v}_1$ given a new update $\mathbf{b}_1^{[s]}$ and the previous updates $\sigma_e^{2[s-1]}, \sigma_{f_1}^{2[s-1]}$. We show in Section 4.8.2 that this work is equivalent to minimizing the following under the condition that the supervision effect

exists,

$$B^{[s]}(\mathbf{v}_1) = \left\| \left(\frac{1}{2\sigma_e^{2[s-1]}} \mathbf{\Psi}_2^T \mathbf{\Psi}_2 - \frac{\sigma_{f_1}^{2[s-1]}}{2\sigma_e^{2[s-1]} \left(\sigma_{f_1}^{2[s-1]} + \sigma_e^{2[s-1]} \right)} \mathbf{X}^T \mathbf{X} \right)^{-\frac{1}{2}} \right.$$
$$\times \left(\frac{1}{2\sigma_e^{2[s-1]}} \mathbf{\Psi}_2^T \mathbf{\Psi}_1 - \frac{\sigma_{f_1}^{2[s-1]}}{2\sigma_e^{2[s-1]} \left(\sigma_{f_1}^{2[s-1]} + \sigma_e^{2[s-1]} \right)} \mathbf{X}^T \mathbf{Y} \mathbf{b}_1^{[s]} \right)$$
$$\left. - \left(\frac{1}{2\sigma_e^{2[s-1]}} \mathbf{\Psi}_2 \mathbf{\Psi}_2^T - \frac{\sigma_{f_1}^{2[s-1]}}{2\sigma_e^{2[s-1]} \left(\sigma_{f_1}^{2[s-1]} + \sigma_e^{2[s-1]} \right)} \mathbf{X}^T \mathbf{X} \right)^{\frac{1}{2}} \mathbf{v} \right\|_2^2 + \sum_{i=1}^{m} \lambda \left\| \mathbf{v}_{1,G_i}^{[0]} \right\|_2,$$

where

$$\mathbf{\Psi}_1 = \begin{pmatrix} \mathbf{X}_{\cdot 1} \\ \mathbf{X}_{\cdot 2} \\ \vdots \\ \mathbf{X}_{\cdot p} \end{pmatrix}, \quad \mathbf{\Psi}_2 = \begin{pmatrix} \mathbf{Y}\mathbf{b}_1^{[s]} & 0 & \dots & 0 \\ 0 & \mathbf{Y}\mathbf{b}_1^{[s]} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{Y}\mathbf{b}_1^{[s]} \end{pmatrix}.$$

Similar as in the updating step 4.7, the estimate $\mathbf{v}_1^{[s]}$ is obtained by the above regression problem with a group Lasso penalty.

From 4.6, it is natural to update $\sigma_e^{2[s]}$ and $\sigma_{f_1}^{2[s]}$ for σ_e^2 and $\sigma_{f_1}^2$ by,

$$\sigma_{f_1}^{2[s]} = \frac{\left\| \mathbf{Y}\mathbf{v}_1^{[s]} - \mathbf{X}\mathbf{b}_1^{[s]} \right\|_2^2}{n}, \quad \sigma_e^{2[s]} = \frac{\left\| \mathbf{Y} - \left(\mathbf{Y}\mathbf{v}_1^{[s]} \right) \left(\mathbf{v}_1^{[s]} \right)^T \right\|_F^2}{np}.$$

For comparison, the penalized likelihood at the mth iterate is,

$$Q^{[s]} = \log \mathrm{P}(\mathbf{X} \mid \mathbf{b}_1^{[s]}, \mathbf{v}_1^{[s]}, \sigma_e^{2[s]}, \sigma_{f_1}^{2[s]}) - \sum_{i=1}^{m} \lambda \left\| \mathbf{v}_{1,G_i}^{[s]} \right\|_2 - \sum_{j=1}^{k} \gamma \left\| \mathbf{b}_{1,G_j}^{[s]} \right\|_2$$

③ Repeat until $Q^{[s]} - Q^{[s-1]} < e$ for a prescribed e to get final estimates $\hat{\mathbf{b}}_1$, $\hat{\mathbf{v}}_1$ and $\hat{\sigma}_{f_1}^2$ for the first layer.

④ To estimate $\mathbf{b}_2$, $\mathbf{v}_2, \sigma_{f_2}^2$ and to update $\hat{\sigma}_e^2$ for the second layer, locate a subspace complement to the plane spanned by $\hat{\mathbf{v}}_1$,

$$\mathbf{X}^{[1]} = \mathbf{X} - \left\langle \mathbf{X}, \frac{\hat{\mathbf{v}}_1}{\|\hat{\mathbf{v}}_1\|_2^2} \right\rangle \frac{\hat{\mathbf{v}}_1}{\|\hat{\mathbf{v}}_1\|_2^2}.$$

and apply steps ① - ③.

The above procedures ① - ③ continues to be applied to $\mathbf{X}^{[i-1]}$ to get estimates $\hat{\mathbf{b}}_i$, $\hat{\mathbf{v}}_i$ and $\hat{\sigma}^2_{f_i}$ for the ith layer and a update $\hat{\sigma}^2_e$. The algorithm stops if $\hat{\mathbf{v}}_i = \mathbf{0}$ or no supervision effect is detected, i.e. $\hat{\mathbf{b}}_i = \mathbf{0}$.

4.4 STOPPING RULE OF G-SUPJIVE ALGORITHM

One of the major edges of G-SupJIVE over other competing methods, JIVE [35], SupJIVE [32] and SIFA [31], is that it does not require a pre-calculation of rank for each type of variation; individual, partial joint and full joint variations. The algorithm sequentially estimates a direction of variation at a time regardless of its type until the multi-block data set exhausts its rank related to the supervision effect. This section explains our method's stopping mechanism in detail. We describe the criterion by which our proposed algorithm stops in Lemma 18.

Lemma 18 (G-SupJIVE algorithm stopping rule.)**.** *The sufficient condition that G-SupJIVE algorithm stops for the ith layer is, for a given group Lasso penalty* $\lambda > 0$*,*

$$\|\mathbf{Y}\hat{\mathbf{b}}_i\|_2 < \lambda, \;\; or \;\; \|\mathbf{X}^{[i]}\|_F < \lambda.$$

We now interpret Lemma 18. G-SupJIVE employs a lay-by-layer estimation scheme. It estimates the ith direction of variation $\hat{\mathbf{v}}_i$ and then moves to its complement data subspace $\mathbf{X}^{[i]} - \mathbf{X}^{[i]}\hat{\mathbf{v}}_i$ to estimates the next one, $\hat{\mathbf{v}}_{i+1}$. Note that the size of $\|\mathbf{X}^{[i]}\|_F$ is necessarily decreasing as the number of layer estimated increases. Therefore, for a sufficiently large λ, the algorithm is guaranteed to stop at the ith layer for some i. Moreover, the algorithm stops when the supervision effect, represented by $\hat{\mathbf{b}}_\mathbf{i}$, becomes weak. For most cases of simulations and real data analyses, G-SupJIVE algorithm stops due to the lack of supervision effect, which triggers a call for violation of the positive definite constraint described in (4.17). A proof of Lemma 18 is provided in Section 4.8.3.

In practice, we choose the tuning parameters λ using BIC (similarly for γ),

$$\text{BIC} = -2\log\left(B^{[s-1]}(\mathbf{v}_1)\right) + \kappa\log(n),$$

where κ is the number of non-zero components in the vector $\mathbf{v}_1$ and n is the number of observations in $\mathbf{X}$. In all of our experiments, the number of layers estimated by the G-SupJIVE algorithm, if not correct, exceeds the true rank of the concatenated data by no more than one or two.

4.5 REAL DATA ANALYSIS

We apply G-SupJIVE to the GBM data described in Section 4.1.3. Since their dimensions (23,293 for gene expression and 534 for miRNA) are much larger than sample size of 234, we first reduce their dimensions to 30 and 10 using scores from singular value decomposition (SVD) accounting for over 70 percent of the total variation in each data set. The two component data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are each column-centered and scaled by their Frobenius norms. We model the data using G-SupJIVE as follows,

$$\begin{aligned}
\underset{234\times 40}{\mathbf{X}} &= \left[\underset{234\times 30}{\mathbf{X}_1}, \underset{234\times 10}{\mathbf{X}_2} \right] = \underset{234\times r}{\mathbf{U}} \underset{r\times 40}{\mathbf{V}^T} + \underset{234\times 40}{\mathbf{E}} \\
\underset{234\times 5}{\mathbf{Y}} &= \left[\underset{234\times 1}{\mathbf{Y}_1}, \underset{234\times 1}{\mathbf{Y}_2}, \underset{234\times 1}{\mathbf{Y}_3}, \underset{234\times 1}{\mathbf{Y}_4}, \underset{234\times 1}{\mathbf{Y}_5} \right] \\
\underset{234\times r}{\mathbf{U}} &= \underset{234\times 5}{\mathbf{Y}} \underset{5\times r}{\mathbf{B}} + \underset{234\times r}{\mathbf{F}}
\end{aligned}$$

Each component of supervision sets $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mathbf{Y}_4, \mathbf{Y}_5$ corresponds to the binary variable indicating whether observation belong to one of subtypes of Neural, Mesenchymal, Proneural, Classical and Unclassified (1 if yes and 0 if no). Since each supervision candidate set is a column vector, group Lasso penalty imposed on columns of $\mathbf{B}$ reduces to Lasso penalty. We fit the G-SupJIVE model (4.4) using the algorithm discussed in Section 4.3.3. The algorithm stops at the 17th layer, which means that the intrinsic rank of $\mathbf{X}$ is determined at $r = 17$.

Figure 18 shows parameter estimates for the first four layers in $\mathbf{V}$ and $\mathbf{B}$ and the jitter plots of projection scores of the GBM data onto estimated variation directions, i.e. $\mathbf{X}\mathbf{v}_i$ for $i = 1, 2, 3, 4$. Different colors in the estimates of $\mathbf{B}$ represents different cancer subtypes. Figure 18 suggests that the first variation direction $\mathbf{v}_1$ in $\mathbf{V}$ represents a joint variation defined over both miRNA and gene expression level, where the variation is driven by red, green and cyan-colored subtypes as indicated in the first supervision effect $\mathbf{b}_1$ in $\mathbf{B}$. More specifically, red subtype makes negative contribution and green and cyan subtypes make positive contribution to the variation. This estimated supervision effect indicates that the first estimated variation direction is well discriminating among red and green/cyan subtypes, but not the other subtypes. As a matter of fact, red data points are well separated from green and cyan ones in the jitter plot for the first layer. The second layer estimate $\mathbf{v}_2$ in $\mathbf{V}$, on the other hand, represents an individual variation defined only on gene expression level, which is supervised by a cyan-colored subtype with negative contribution and a black-colored subtype with positive contribution as indicated in the second supervision effect $\mathbf{b}_2$ in $\mathbf{B}$. The cor-

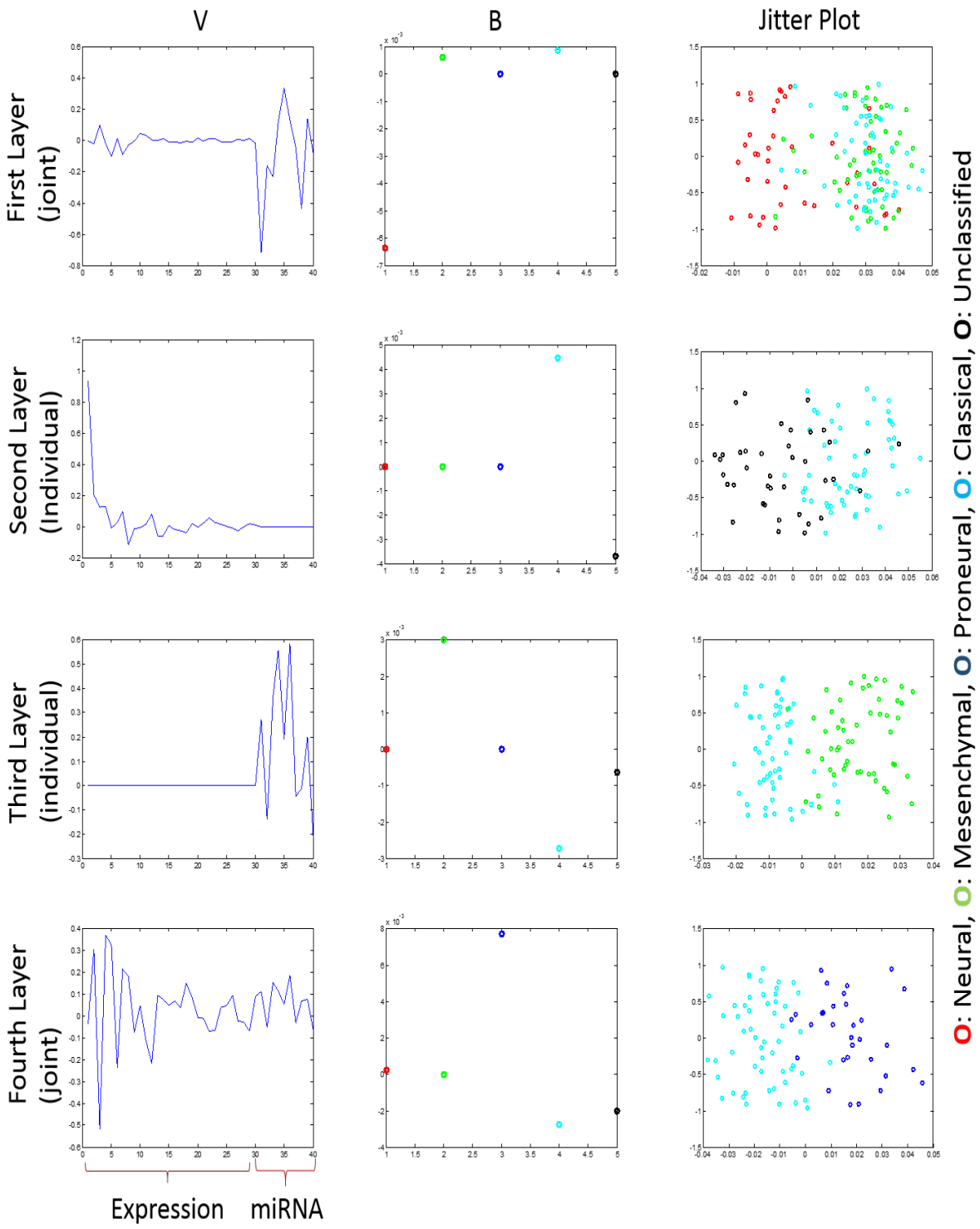

Figure 18. Estimates of **V** and **B** and their discriminating ability.

responding jitter plot clearly shows the second direction's ability of discriminating cyan and black data clusters. Components of the rest of the estimated layers have a similar interpretation.

The benefits of the G-SupJIVE method is clear in this example. It identifies the major joint or individual variation directions in data automatically, reveals supervision that drives each of them, and suggests how the supervision works in driving variations.

4.6 COMPARISON WITH OTHER METHODS

The purpose of this section is to show how G-SupJIVE generalizes existing methods in terms of factorizing multi-block data. More specifically, we aim to demonstrate statistical capacity that our proposed method possesses but other competing do not. First, it would be beneficial to conceptually compare the five competing models, SVD, JIVE [35], SupSVD [32], SIFA [31] and G-SupJIVE.

			SVD	JIVE	SupSVD	SIFA	G-SupJIVE
Data sets	Single	Individual variation	**O**	**O**	**O**	**O**	**O**
	Multiple	Individual variation		**O**		**O**	**O**
		Partial joint variation					**O**
		Full joint variation		**O**		**O**	**O**
Supervision sets	Single				**O**	**O**	**O**
	Multiple						**O**

Table 4. Comparison table indicating a set of functionalities each method is able to perform.

The comparison in Table 4 indicates a set of functionalities each method is able to perform. SVD only factorizes a single data set into layers in the order of magnitude of variability. It is not designed to incorporate multiple data sets or supervision information. JIVE generalizes SVD in such a way that multiple data sets are decomposed into individual and joint variations. SupSVD, on the other hand, can incorporate supervision information so that a single data set is decomposed into supervised and unsupervised variations. SIFA is an effort to combine the advantages of SupSVD and JIVE but is not able to handle partial joint variations or multiple supervision data sets. These

generalization relations are summarized in Figure 19. The method located at the end of an arrow generalizes the one at the start of the arrow. G-SupJIVE so far is the most generalized framework for an integrative decomposition of a multi-block data set. We note that the proposed estimating algorithm of G-SupJIVE is not simple extension of JIVE, SupJIVe or SIFA.

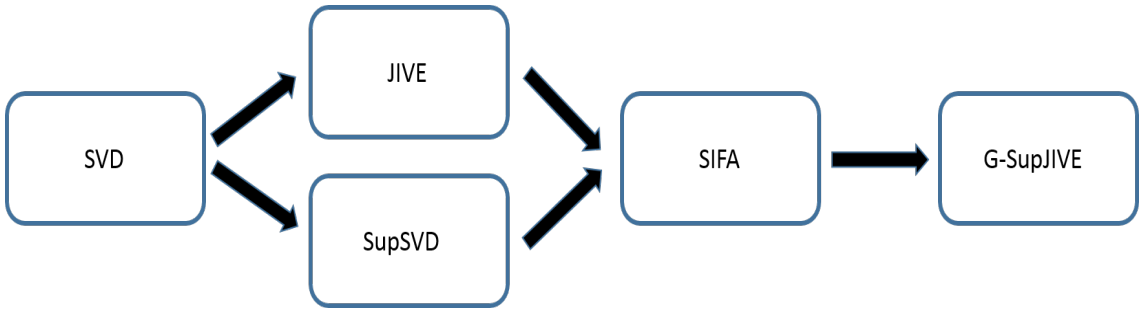

Figure 19. Conceptual diagram of the generalization relationship among different methods. The end point of each arrow generalizes its starting point.

To compare these methods quantitatively, we here present an illustrative example, which is more general than the previous example in Section 4.3.2. We use the following setting,

$$\underset{120\times100}{\mathbf{X}} = [\underset{120\times25}{\mathbf{X}_1}, \underset{120\times25}{\mathbf{X}_2}, \underset{120\times25}{\mathbf{X}_3}, \underset{120\times25}{\mathbf{X}_4}] = \underset{120\times4}{\mathbf{U}} \underset{4\times100}{\mathbf{V}^T} + \underset{120\times100}{\mathbf{E}}$$

$$\underset{120\times40}{\mathbf{Y}} = [\underset{120\times10}{\mathbf{Y}_1}, \underset{120\times10}{\mathbf{Y}_2}, \underset{120\times10}{\mathbf{Y}_3}, \underset{120\times10}{\mathbf{Y}_4}]$$

$$\underset{120\times4}{\mathbf{U}} = \underset{120\times40}{\mathbf{Y}} \underset{40\times4}{\mathbf{B}} + \underset{120\times4}{\mathbf{F}}$$

The multi-source data set $\mathbf{X}$ has 120 observations and consists of 4 subsets $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$, each with 25 measurements. We assumed four components in V consisting of two individual, one partial joint and one full joint variation direction as depicted in Figure 20. The intrinsic rank of $\mathbf{X}$ is four. The supervision set $\mathbf{Y}$ now consists of 4 supervision candidates $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mathbf{Y}_4$ where each column of $\mathbf{Y}_i$ is filled with 120-dimensional normal random vector with mean $\mathbf{0}$ and diagonal covariance matrix $\sigma_i^2\mathbf{I}$ for $\sigma_i^2 = 2.5, 2, 1.5, 1$. Each row of the maxtix $\mathbf{F}$ is independently generated from 4-dimensional normal random vector with mean $\mathbf{0}$ and diagonal covariance structure with entries of 8, 6, 4 and 2. As shown in the first row of Figure 20, the first column of $\mathbf{B}$ picks up the second supervision candidate and provides variations in the direction represented by the first column of V. The roles of the rest of columns of B are interpreted in a similar way. Each entry of the noise matrix $\mathbf{X}$ is filled up with a value from i.i.d. standard normal random variable. The

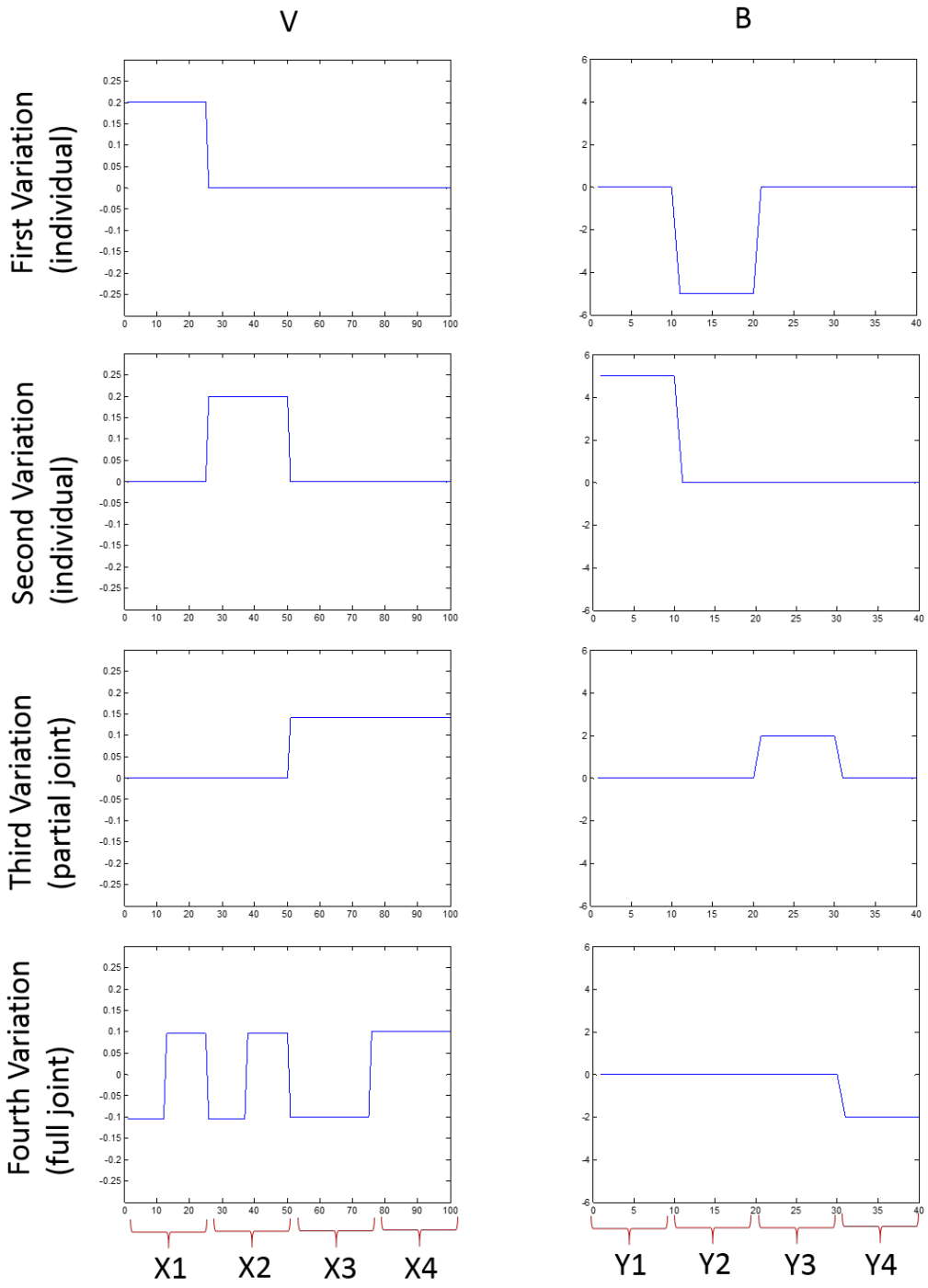

Figure 20. Parameters used in the comparison study.

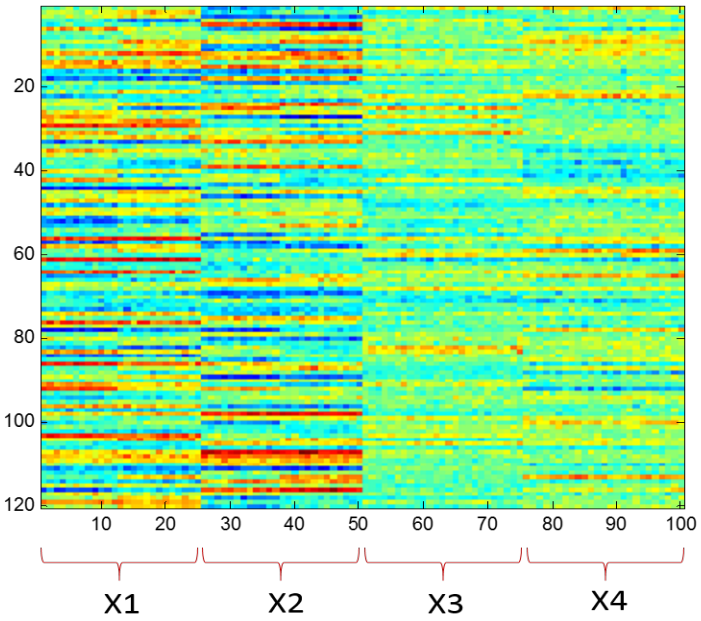

Figure 21. Heat map of the simulated data.

Heat map of this simulated data is shown in Figure 21. For this data set, we fit G-SupJIVE, SVD, JIVE SupSVD and SIFA. The estimation results are visually summarized in Figures 22, 23, 24, 25.

Estimation results for G-SupJIVE are shown in Figure 22, where estimates are represented by blue curves and parameters are by red curves. G-SupJIVE algorithm stops at the fourth iteration, which implies that it correctly estimates the intrinsic rank of $\mathbf{X}$. Overall, G-SupJIVE effectively captures the variation patterns and supervisions that drive them. Figure 23 shows variation direction estimates from SVD and JIVE. As expected, SVD is not able to correctly differentiate three different types of variations. JIVE relatively performs well except that the partial joint variation is estimated by two separate individual variations since JIVE decomposes variations into joint and individual only. These two methods both do not provide any information on which supervision drives which major variation. Estimates of variation directions and their corresponding supervision effects from the application of SupSVD to the example are shown in Figure 24. SupSVD performs poor since it is, like SVD, designed for a single data set, not for multiple data sets. SupSVD is not able to correctly distinguish among three different types of variations but it provides supervision effect information though not precise. To our surprise, SIFA performs not properly as observed in

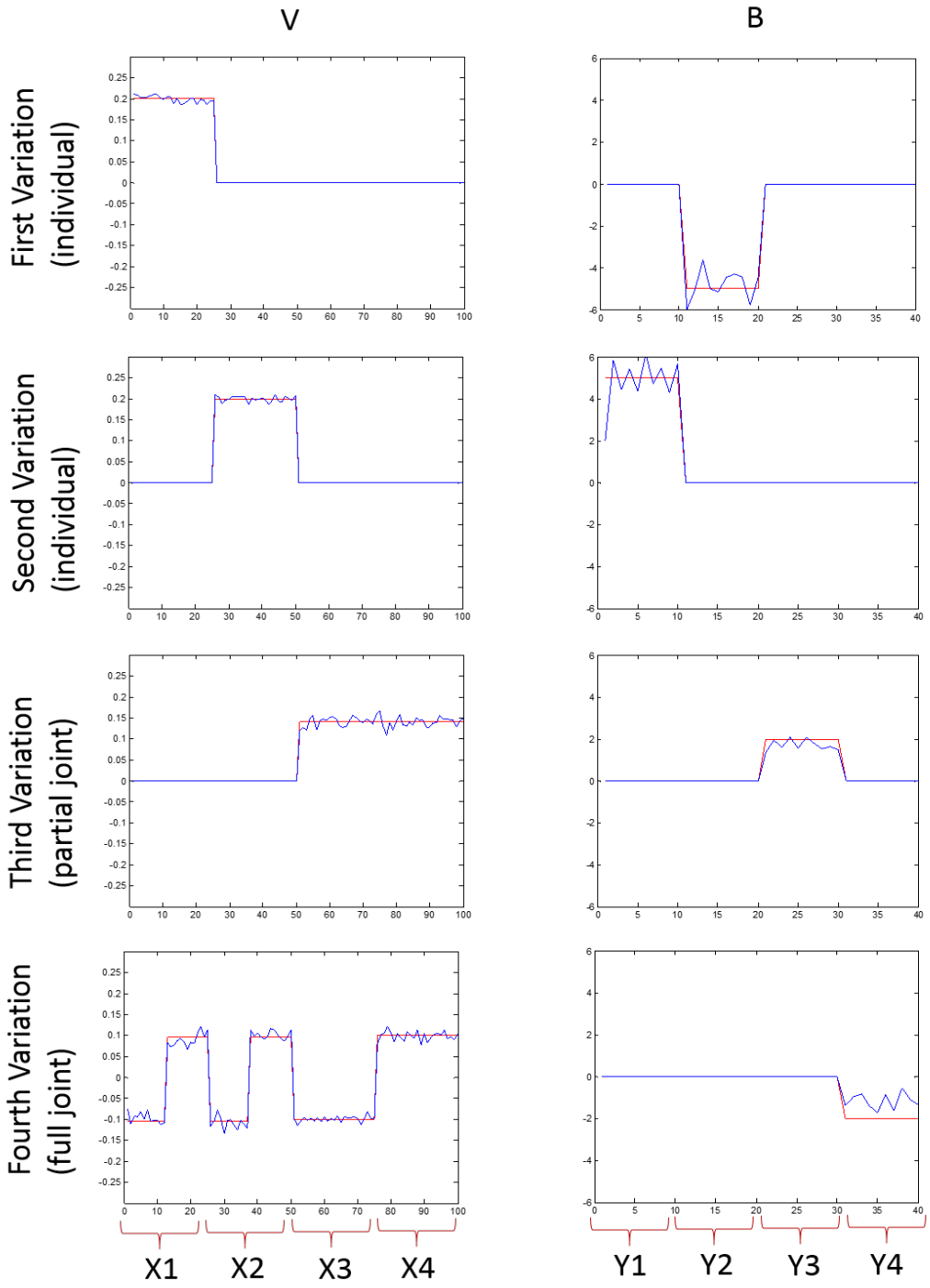

Figure 22. G-SupJIVE estimates overlaid with the parameters.

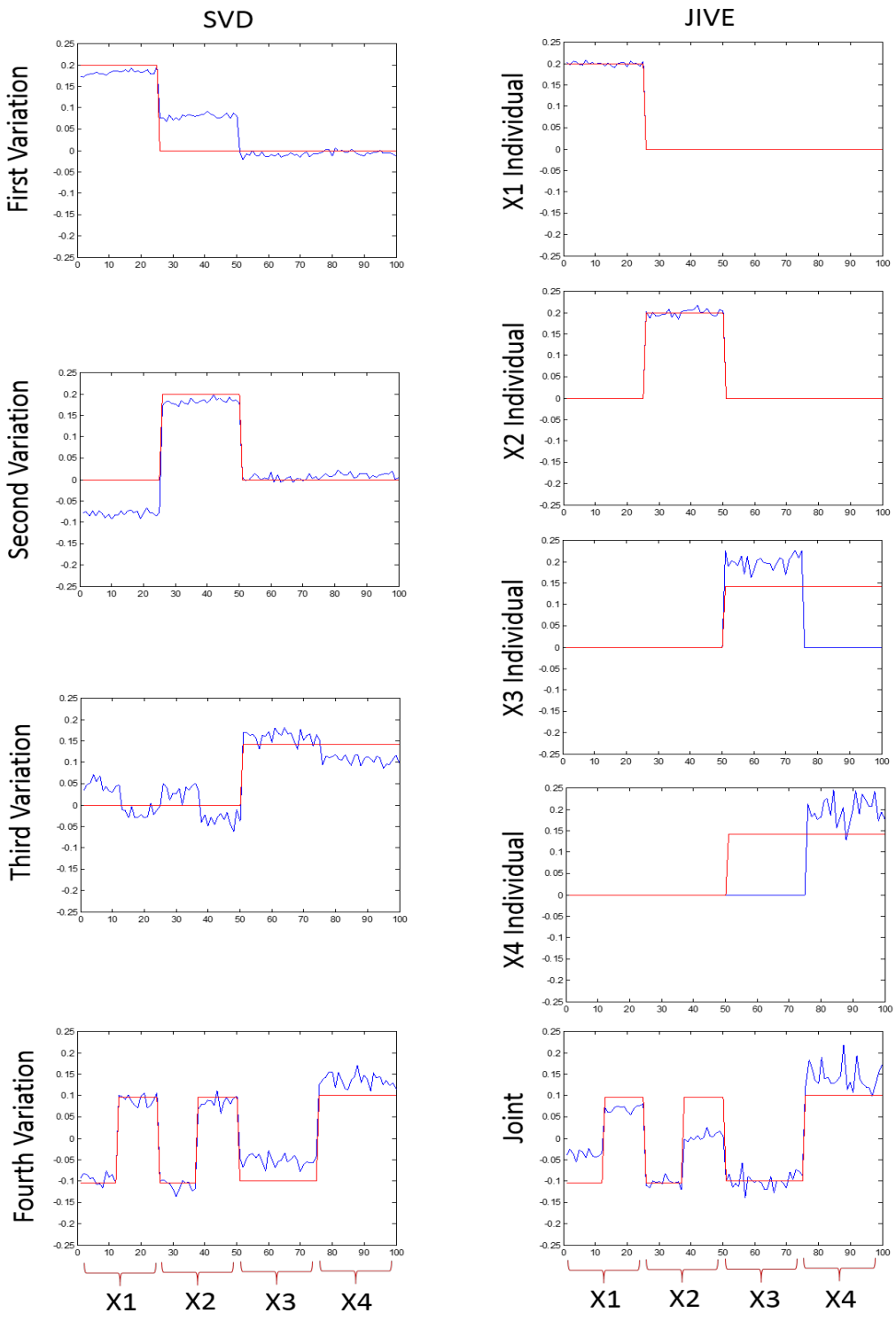

Figure 23. SVD and JIVE estimates overlaid with the parameters.

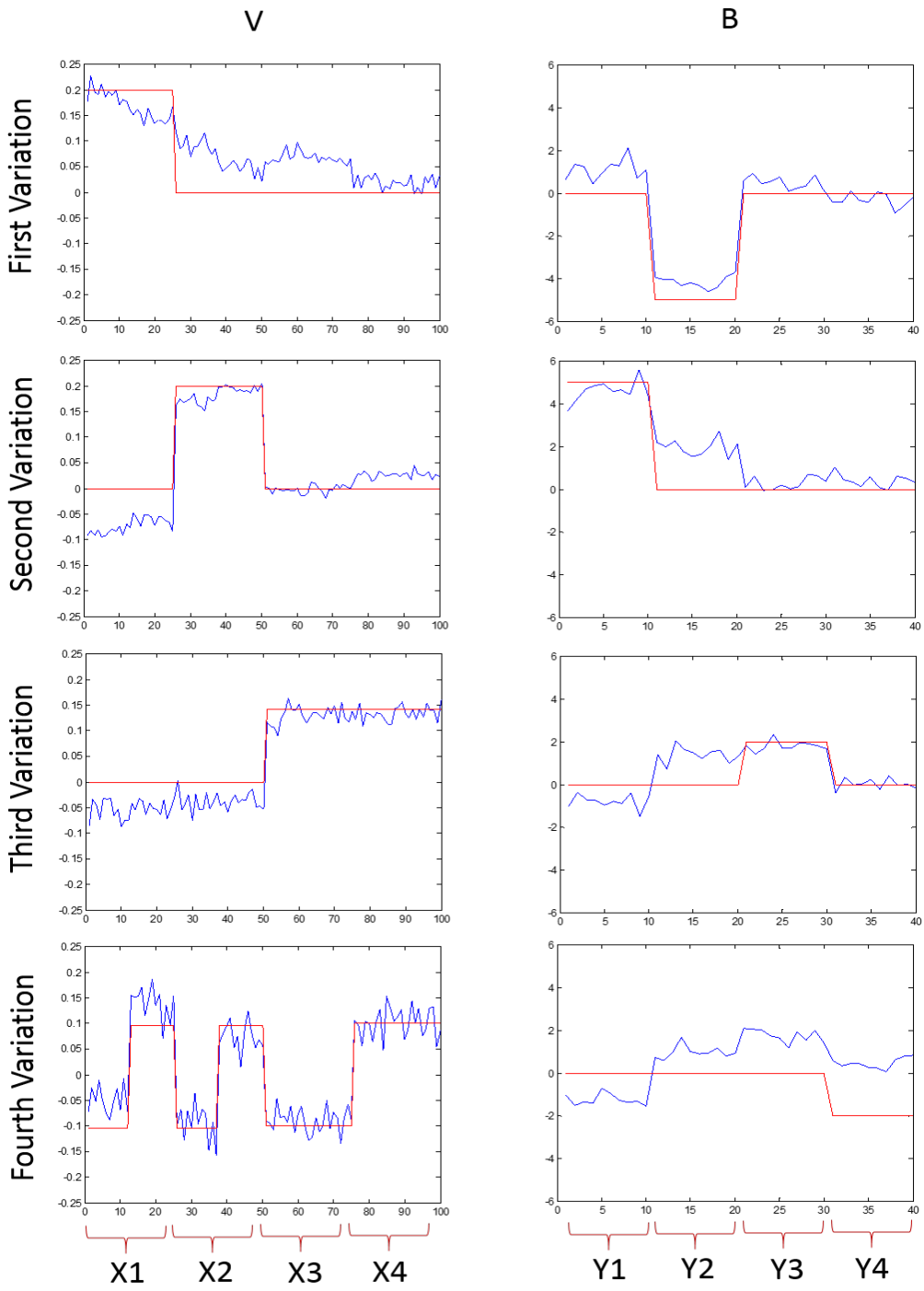

Figure 24. SupSVD estimates overlaid with the parameters.

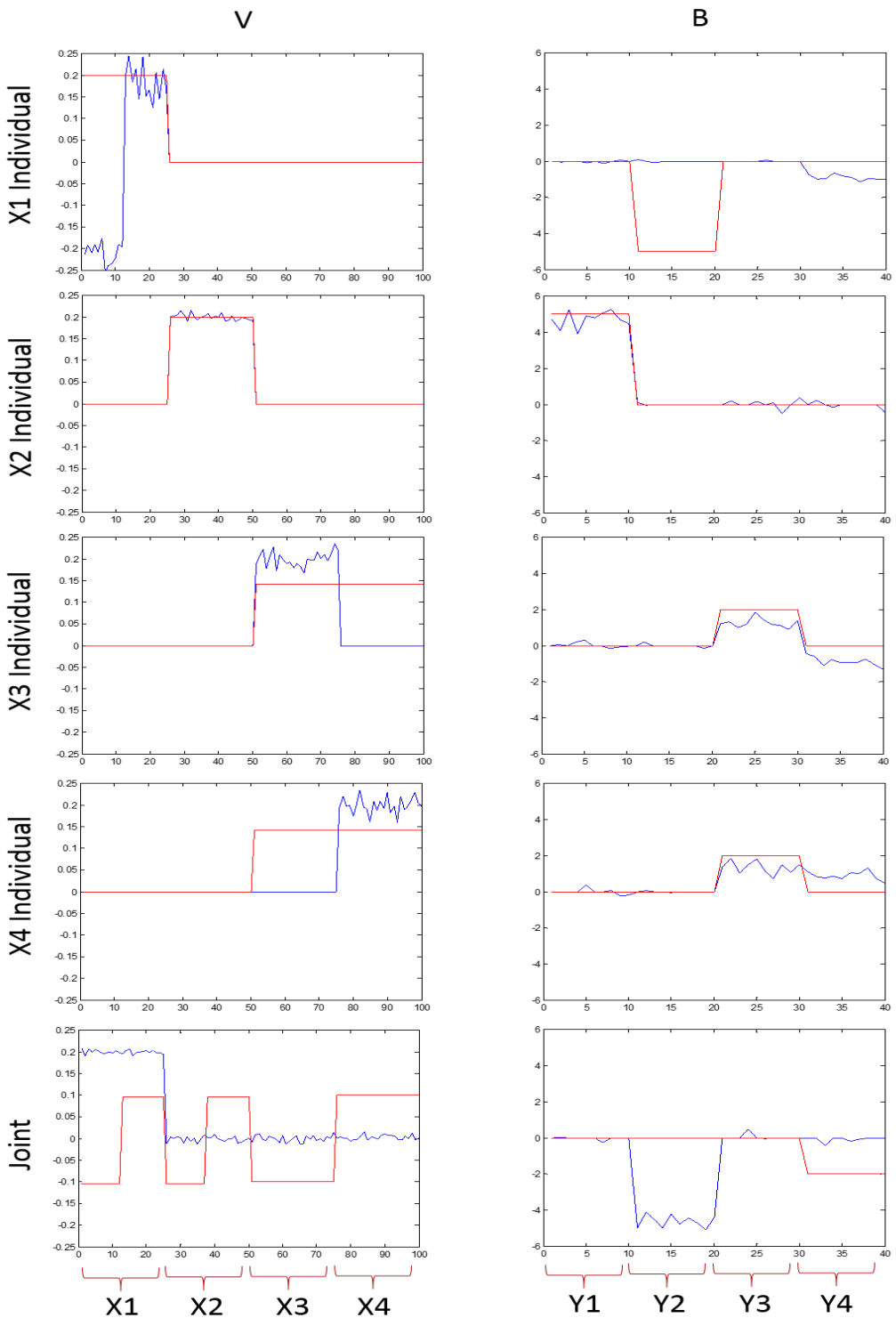

Figure 25. SIFA estimates overlaid with the parameters.

Figure 25. This is because SIFA is not robust to the model misspecification.

4.7 SIMULATION

In this section, we conduct comprehensive simulation studies to demonstrate the advantage of the proposed method over existing ones. We evaluate the accuracy of the parameter estimation to compare G-SupJIVE with SVD, JIVE, SupSVD and SIFA.

4.7.1 Simulation eetting

The simulated multi-block data set consists of four primary data sets $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ and $\mathbf{X}_4$ on the same set of subjects with sample size 120 and dimension 25 for each data set throughout all simulation settings. We consider two different settings where, in Setting 1, the generative models are G-SupJIVE including a partial joint variation and, in Setting 2, SIFA (or G-SupJIVE not including a partial joint variation).

- Setting 1 (G-SupJIVE): 2 individual, 1 partial and 1 full joint variations

$$\begin{aligned}
\underset{120\times100}{\mathbf{X}} &= [\underset{120\times25}{\mathbf{X}_1}, \underset{120\times25}{\mathbf{X}_2}, \underset{120\times25}{\mathbf{X}_3}, \underset{120\times25}{\mathbf{X}_4}] = \underset{120\times4}{\mathbf{U}}\ \underset{4\times100}{\mathbf{V}^T} + \underset{120\times100}{\mathbf{E}} \\
\underset{120\times40}{\mathbf{Y}} &= [\underset{120\times10}{\mathbf{Y}_1}, \underset{120\times10}{\mathbf{Y}_2}, \underset{120\times10}{\mathbf{Y}_3}, \underset{120\times10}{\mathbf{Y}_4}] \\
\underset{120\times4}{\mathbf{U}} &= \underset{120\times40}{\mathbf{Y}}\ \underset{40\times4}{\mathbf{B}} + \underset{120\times4}{\mathbf{F}}
\end{aligned}$$

  This setting generates a multi-block data set of its intrinsic rank of 4. The supervision matrix $\mathbf{Y}$ contains 4 candidate supervision sets $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mathbf{Y}_4$. Each column of $\mathbf{Y}_i$ is filled with 120-dimensional normal random vector with mean $\mathbf{0}$ and diagonal covariance matrix $\sigma_i^2\mathbf{I}$ for $\sigma_i^2 = 2.5, 2, 1.5, 1$. Each row of the matrix $\mathbf{F}$ comes from i.i.d. 4-dimensional normal random vector with mean $\mathbf{0}$ and diagonal covariance matrix with entries of 8, 6, 4 and 2. Entries of the measurement error $\mathbf{E}$ is filled with i.i.d. standard normal random samples. Components of the supervision effect matrix $\mathbf{B}$ and variation matrix $\mathbf{V}$ are chosen as in Figure 20.

- Setting 2 (SIFA): 4 individual and 1 full joint variations

$$\begin{aligned}
\underset{120\times100}{\mathbf{X}} &= [\underset{120\times25}{\mathbf{X}_1}, \underset{120\times25}{\mathbf{X}_2}, \underset{120\times25}{\mathbf{X}_3}, \underset{120\times25}{\mathbf{X}_4}] = \underset{120\times5}{\mathbf{U}}\ \underset{5\times100}{\mathbf{V}^T} + \underset{120\times100}{\mathbf{E}}, \\
\underset{120\times40}{\mathbf{Y}} &= [\underset{120\times10}{\mathbf{Y}_1}, \underset{120\times10}{\mathbf{Y}_2}, \underset{120\times10}{\mathbf{Y}_3}, \underset{120\times10}{\mathbf{Y}_4}, \underset{120\times10}{\mathbf{Y}_3}],
\end{aligned}$$

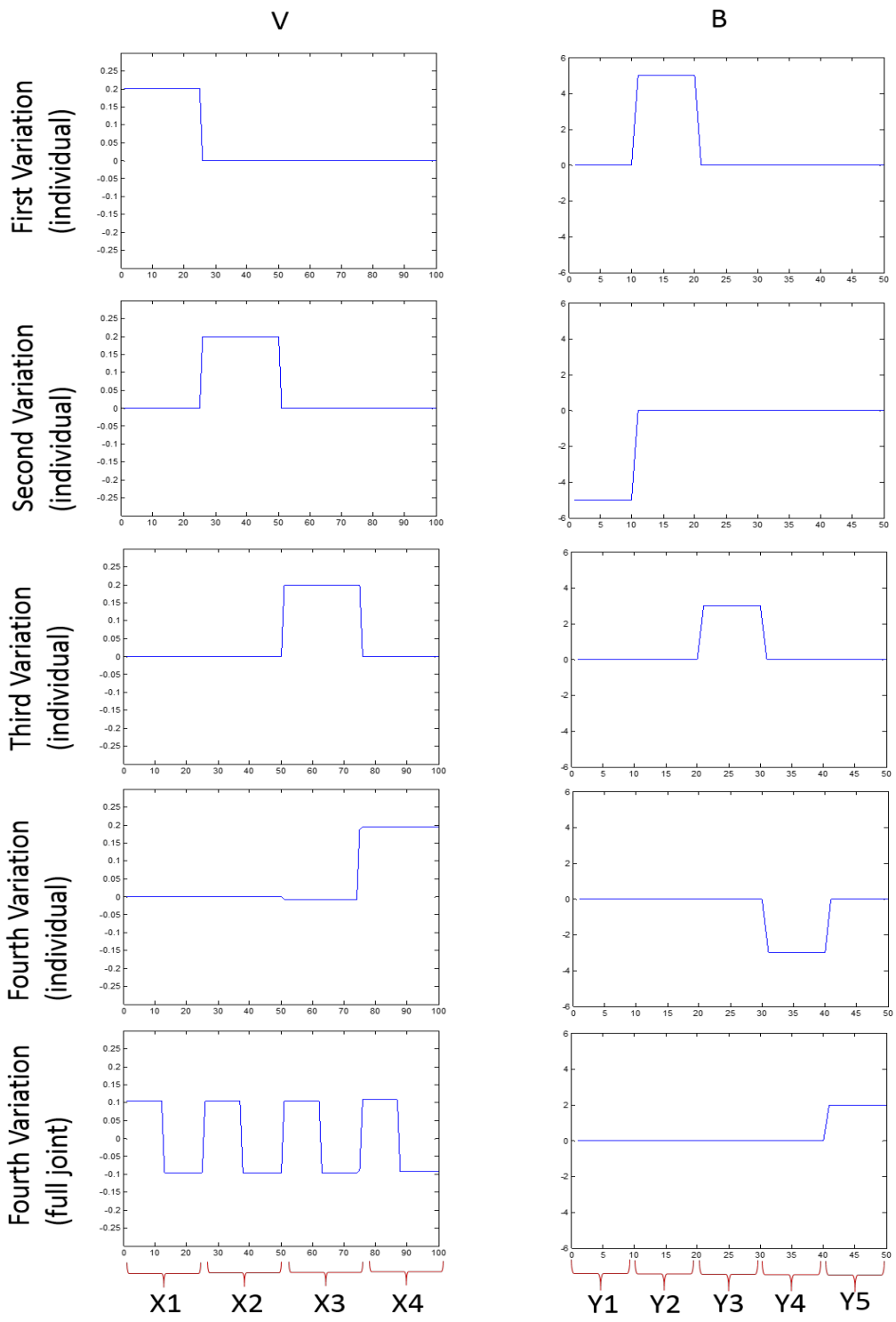

Figure 26. Parameters for simulation setting 2

$$\underset{120\times 5}{\mathbf{U}} = \underset{120\times 40}{\mathbf{Y}}\ \underset{40\times 5}{\mathbf{B}} + \underset{120\times 5}{\mathbf{F}}.$$

This setting generates a multi-block data set of its intrinsic rank of 5 and differs from the setting 1 in that a partial joint variation is excluded. The supervision matrix $\mathbf{Y}$ contains 5 candidate supervision sets $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mathbf{Y}_4, \mathbf{Y}_5$. Each column of $\mathbf{Y}_i$ is filled with 120-dimensional normal random vector with mean $\mathbf{0}$ and diagonal covariance matrix $\sigma_i^2\mathbf{I}$ for $\sigma_i^2 = 3, 2.5, 2, 1.5, 1$. Each row of the matrix $\mathbf{F}$ comes from i.i.d. 4-dimensional normal random vector with mean $\mathbf{0}$ and diagonal covariance matrix with entries of 8, 6.5, 5, 4.5 and 2. Entries of the measurement error $\mathbf{E}$ is filled with i.i.d. standard normal random samples. Components of the supervision effect matrix $\mathbf{B}$ and variation matrix $\mathbf{V}$ are displayed as in Figure 26.

4.7.2 Simulation results

We generate 100 simulated data sets from each setting and fit SVD, JIVE, SupSVD, SIFA and G-SupJIVE to them. For SVD and SupSVD, the concatenated data set $\mathbf{X}$ is considered as a single data set and, for SupSVD and SIFA, the concatenated supervision set $\mathbf{Y}$ is fed as if it is a single supervision set. To avoid ambiguity, we fit each model with the true ranks. SVD and SupSVD are fitted with rank 4 or 5 depending on setting. If a method (JIVE and SIFA) does not assume a partial joint variation, the rank of a partial joint variation is assigned to the rank of each of individual sets on which the partial joint variation is defined. To compare the quality of the estimated loading vectors for $\mathbf{B}$ and $\mathbf{V}$ among the five methods, we evaluate the angle formed between each vector estimate for $\mathbf{B}$ and $\mathbf{V}$ and its counterpart parameter measured by degree (°). Note that since JIVE and SupSVD do not provide estimates related to a partial joint variation direction, corresponding evaluation results are missing in Table 5.

The result of simulations is summarized in Table 5. Poor performance of SVD, either in Setting 1 or 2, is expected as it is not for multi-block data nor for data with supervision effect. JIVE, on the other hand, exhibits a good performance in catching individual variations both in Setting 1 and 2. However, it is not estimating well the full joint variation for either Setting 1 or 2. Intuitively, JIVE conducts SVD within each of individual and joint data sets given its rank. It does not take any distributional assumption and simply find strong variations. In our settings, individual variations are relatively strong compared to the joint signal, which actually is the weakest. This may explain the reason why JIVE is good in revealing individual variations even without supervision being taken into account. SupSVD shows suboptimal performance in estimating variation directions. SupSVD

			SVD	JIVE	SupSVD	SIFA	G-SupJIVE
Setting 1 (G-SupJIVE)	**V**	$\angle(\mathbf{V}_{X_1}, \hat{\mathbf{V}}_{X_1})$	36.815(4.012)	**1.763(0.343)**	27.498(15.210)	41.782(0.298)	1.769(0.242)
		$\angle(\mathbf{V}_{X_2}, \hat{\mathbf{V}}_{X_2})$	37.248(3.895)	1.896(0.191)	39.786(13.658)	**1.859(0.212)**	1.880(0.275)
		$\angle(\mathbf{V}_P, \hat{\mathbf{V}}_P)$	44.078(4.680)		44.987(24.055)		**2.463(1.614)**
		$\angle(\mathbf{V}_F, \hat{\mathbf{V}}_F)$	43.998(4.709)	22.556(0.579)	33.640(29.723)	75.649(4.314)	**3.971(3.905)**
	B	$\angle(\mathbf{B}_{\mathbf{V}_{X_1}}, \hat{\mathbf{B}}_{\mathbf{V}_{X_1}})$			27.208(14.617)	43.937(4.312)	**7.309(1.290)**
		$\angle(\mathbf{B}_{\mathbf{V}_{X_2}}, \hat{\mathbf{B}}_{\mathbf{V}_{X_2}})$			32.130(14.928)	13.713(5.417)	**9.387(1.491)**
		$\angle(\mathbf{B}_{\mathbf{V}_P}, \hat{\mathbf{B}}_{\mathbf{V}_P})$			59.527(17.961)		**11.320(1.225)**
		$\angle(\mathbf{B}_{\mathbf{V}_F}, \hat{\mathbf{B}}_{\mathbf{V}_F})$			39.249(26.475)	63.643(1.697)	**9.168(1.188)**
Setting 2 (SIFA)	**V**	$\angle(\mathbf{V}_{X_1}, \hat{\mathbf{V}}_{X_1})$	37.343(5.895)	1.591(0.215)	35.355(19.771)	**1.587(0.223)**	1.704(0.237)
		$\angle(\mathbf{V}_{X_2}, \hat{\mathbf{V}}_{X_2})$	37.560(5.867)	**1.542(0.165)**	52.728(18.704)	1.547(0.181)	2.843(0.433)
		$\angle(\mathbf{V}_{X_3}, \hat{\mathbf{V}}_{X_3})$	41.304(12.187)	**1.591(0.215)**	50.127(17.476)	2.717(0.348)	1.700(0.612)
		$\angle(\mathbf{V}_{X_4}, \hat{\mathbf{V}}_{X_4})$	51.269(12.193)	1.701(0.229)	45.013(19.663)	**1.403(0.088)**	2.16(0.128)
		$\angle(\mathbf{V}_F, \hat{\mathbf{V}}_F)$	20.560(0.695)	28.983(1.680)	23.240(10.867)	69.349(0.461)	**8.766(0.847)**
	B	$\angle(\mathbf{B}_{\mathbf{V}_{X_1}}, \hat{\mathbf{B}}_{\mathbf{V}_{X_1}})$			32.946(20.078)	21.827(2.575)	**12.443(0.949)**
		$\angle(\mathbf{B}_{\mathbf{V}_{X_2}}, \hat{\mathbf{B}}_{\mathbf{V}_{X_2}})$			47.082(20.226)	23.015(4.083)	**9.606(0.738)**
		$\angle(\mathbf{B}_{\mathbf{V}_{X_3}}, \hat{\mathbf{B}}_{\mathbf{V}_{X_3}})$			57.593(15.593)	**13.401(1.767)**	15.119(0.965)
		$\angle(\mathbf{B}_{\mathbf{V}_{X_4}}, \hat{\mathbf{B}}_{\mathbf{V}_{X_4}})$			47.634(18.849)	**13.429(1.755)**	17.910(0.888)
		$\angle(\mathbf{B}_{\mathbf{V}_F}, \hat{\mathbf{B}}_{\mathbf{V}_F})$			36.630(12.902)	12.857(1.971)	**11.350(1.423)**

Table 5. Simulation results under Setting 1 and 2 (each with 100 simulation runs), where $\mathbf{V}_{X_i}, \mathbf{V}_P, \mathbf{V}_F$ are population directions for individual variation in data set $\mathbf{X}_i$, partial and full joint variations respectively; $\mathbf{B}_{\mathbf{V}_{X_i}}, \mathbf{B}_{\mathbf{V}_P}, \mathbf{B}_{\mathbf{V}_F}$ are population supervision effects for variation directions $\mathbf{V}_{X_i}, \mathbf{V}_P, \mathbf{V}_F$ respectively; symbols with $\hat{}$ are estimate counterparts of symbols without $\hat{}$; $\angle$ is the angle between two arguments. The mean and standard deviation of the angle between estimate and its population counterpart measured by degree for each method are shown in the table. The best results are highlighted in bold.

improves upon SVD in a way that it provides supervision effects though not precise. Like SVD, SupSVD cannot handle effectively the multi-block data. SIFA's low performance in Setting 1 is due to the model misspecification: SIFA is not capable of incorporating any partial joint distribution and multiple supervision sets. However, SIFA is good at estimating both variations and supervision effects for the data generated from SIFA model except for the full joint variation estimation. This may be due to the smaller size of variation of the joint component. It is clear from Table 5 that G-SupJIVE performs much better than other methods for both Setting 1 and 2. G-SupJIVE algorithm stops at the 4th or 5th layer estimation for setting 1 and at the 5th or 6th for setting 2.

4.8 TECHNICAL DETAILS

4.8.1 Details of transformation of objective function of $\mathbf{b}_1$

The minus log likelihood (4.6) as a function $\mathbf{b}_1$ given $\mathbf{v}_1, \sigma_e^2$, and $\sigma_{f_1}^2$ modulo the constant terms with respect to $\mathbf{b}_1$ (the index $[s-1]$ is omitted) is,

$$\frac{1}{2\sigma_e^2}\left\|\mathbf{X}-\mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}\left\|\mathbf{X}\mathbf{v}_1-\mathbf{Y}\mathbf{b}_1\right\|_2^2 + \sum_{j=1}^{k}\gamma\left\|\mathbf{b}_{1,G_j}\right\|_2. \tag{4.8}$$

We show that minimizing the function (4.8) is equivalent to minimizing the following objective function $A(\mathbf{b}_1)$,

$$A(\mathbf{b}_1) = \left\|\mathbf{\Phi}_1 - \mathbf{\Phi}_2\mathbf{b}_1\right\|_2^2 + \sum_{j=1}^{k}\gamma\left\|\mathbf{b}_{1,G_j}\right\|_2, \tag{4.9}$$

where

$$\mathbf{\Phi}_1 = \begin{pmatrix} \sqrt{\frac{1}{2\sigma_e^2}}\mathbf{X}_{\cdot 1} \\ \sqrt{\frac{1}{2\sigma_e^2}}\mathbf{X}_{\cdot 2} \\ \vdots \\ \sqrt{\frac{1}{2\sigma_e^2}}\mathbf{X}_{\cdot p} \end{pmatrix}, \quad \mathbf{\Phi}_2 = \begin{pmatrix} \sqrt{\frac{\mathbf{v}_{1,1}}{2\sigma_e^2}}\mathbf{Y} \\ \sqrt{\frac{\mathbf{v}_{1,2}}{2\sigma_e^2}}\mathbf{Y} \\ \vdots \\ \sqrt{\frac{\mathbf{v}_{1,p}}{2\sigma_e^2}}\mathbf{Y} \end{pmatrix}.$$

It is easy to see that.

$$\frac{1}{2\sigma_e^2} \geq \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}. \tag{4.10}$$

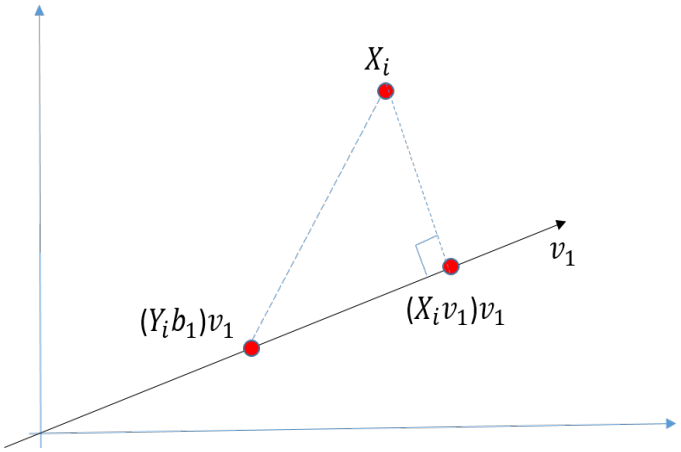

Figure 27. Geometric relation between the two Frobenius norms, (4.11) and (4.12)

By separating each column of the matrix $\mathbf{X} - \mathbf{Y}\mathbf{b}_1 (\mathbf{v}_1)^T$ in the first term of the equation (4.8) and stacking one on another, it can be shown that its Frobenius norm is equivalent to a vector norm,

$$\left\|\mathbf{X} - \mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2 = \|\mathbf{\Phi}_1 - \mathbf{\Phi}_2\mathbf{b}_1\|_2^2 . \tag{4.11}$$

By the Unitary invariant property of Frobenius norm,

$$\|\mathbf{X}\mathbf{v}_1 - \mathbf{Y}\mathbf{b}_1\|_2^2 = \left\|(\mathbf{X}\mathbf{v}_1 - \mathbf{Y}\mathbf{b}_1)\mathbf{v}_1^T\right\|_F^2 . \tag{4.12}$$

The Figure 27 shows the geometric relation between the two Frobenius norms, (4.11) and (4.12). Consider the ith rows of $\mathbf{X}$ and $\mathbf{Y}$ and denote them by $\mathbf{X}_i$ and $\mathbf{Y}_i$. Note that the vector $\mathbf{v}_1$ is fixed with a unit norm so that $\mathbf{X}_i\mathbf{v}_1$ is a projection of $\mathbf{X}_i$ onto $\mathbf{v}_1$. Then $\mathbf{X}_i - \mathbf{Y}_i\mathbf{b}_1\mathbf{v}_1^T$ and $(\mathbf{X}_i\mathbf{v}_1 - \mathbf{Y}_i\mathbf{b}_1)\mathbf{v}_1^T$, respectively, represent the hypothenus and the bottom of a right triangle formed by $\mathbf{X}_i, (\mathbf{Y}_i\mathbf{b}_1)\mathbf{v}_1$ and $(\mathbf{X}_i\mathbf{v}_1)\mathbf{v}_1$. By the Pythogorean theorem,

$$\left\|\mathbf{X}_i - \mathbf{Y}_i\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2 \geq \left\|(\mathbf{X}_i\mathbf{v}_1 - \mathbf{Y}_i\mathbf{b}_1)\mathbf{v}_1^T\right\|_F^2, \text{ for each } i,$$

which leads to,

$$\left\|\mathbf{X} - \mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2 \geq \left\|(\mathbf{X}\mathbf{v}_1 - \mathbf{Y}\mathbf{b}_1)\mathbf{v}_1^T\right\|_F^2 . \tag{4.13}$$

A tedious calculation of partial derivatives of (4.11) and (4.12) with respect to each component $\mathbf{b}_{1,i}$ of $\mathbf{b}_1$ using $\mathbf{v}_1$ being of a unit length shows,

$$\frac{\partial \left\|\mathbf{X} - \mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2}{\partial \mathbf{b}_{1,i}} = \frac{\partial \left\|(\mathbf{X}\mathbf{v}_1 - \mathbf{Y}\mathbf{b}_1)\mathbf{v}_1^T\right\|_F^2}{\partial \mathbf{b}_{1,i}}, \text{ for each } i. \tag{4.14}$$

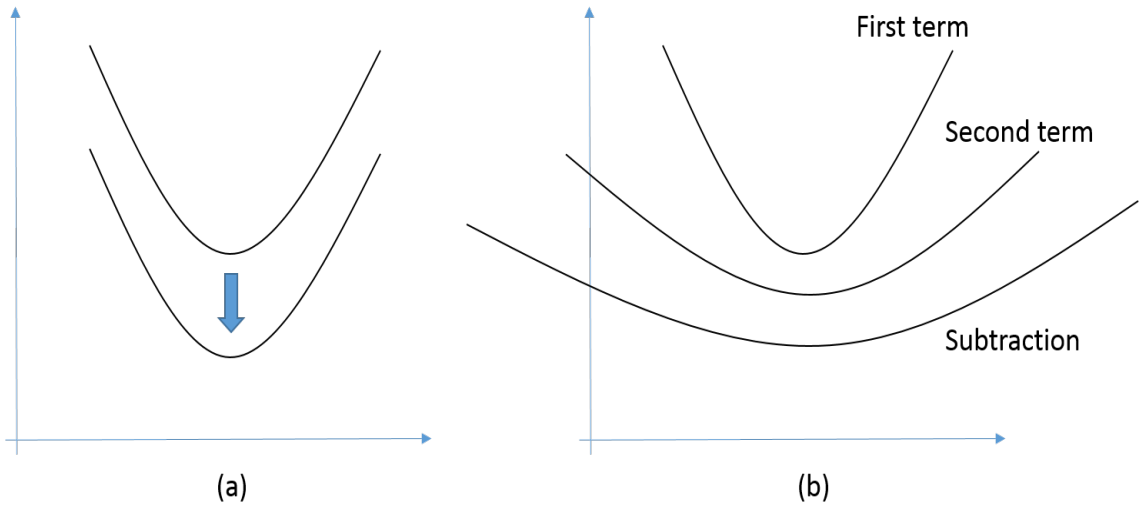

Figure 28. Description of changes of objective functions

In specific,

$$\frac{\partial \left\|\mathbf{X}-\mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2}{\partial \mathbf{b}_{1,i}} = \frac{\partial \left\|(\mathbf{X}\mathbf{v}_1-\mathbf{Y}\mathbf{b}_1)\mathbf{v}_1^T\right\|_F^2}{\partial \mathbf{b}_{1,i}}$$
$$= \sum_{e=1}^{n}\sum_{f=1}^{p}\left(\mathbf{X}_{e,f}-\left(\sum_{g=1}^{m}\mathbf{Y}_{e,g}\mathbf{b}_{1,g}\right)\mathbf{v}_{1,f}\right)\mathbf{Y}_{e,i}\mathbf{v}_{1,f}, \text{ for each } i.$$

The inequality (4.13) and the equality (4.14) imply that (4.12) is a vertical downward translation of (4.11) and positive as shown in (a) of Figure 28. Moreover, (4.10) indicates that the first term of (4.8) is a vertical downward translation of the second term of (4.11) with being flattened as shown in (b) of Figure (28). As a result, as shown in (b) of Figure 28, the subtraction of the second term from the first in (4.8) is still a positive quadratic form with its minimum being attained at the same point where the first term does.

4.8.2 Details of transformation of objective function of $\mathbf{v}_1$

The minus log likelihood (4.6) as a function $\mathbf{v}_1$ given $\mathbf{b}_1, \sigma_e^2$, and $\sigma_{f_1}^2$ modulo the constant terms with respect to $\mathbf{v}_1$ (the index $[s]$ and $[s-1]$ are omitted) is,

$$B(\mathbf{v}_1) = \frac{1}{2\sigma_e^2}\left\|\mathbf{X}-\mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T\right\|_F^2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}\left\|\mathbf{Y}\mathbf{b}_1-\mathbf{X}\mathbf{v}_1\right\|_2^2 + \sum_{i=1}^{m}\lambda\left\|\mathbf{v}_{1,G_i}\right\|_2, \tag{4.15}$$

We show that minimizing the function (4.15) is equivalent to minimizing the following objective function $B(\mathbf{v}_1)$ under the condition where there exists supervision effect,

$$B^*(\mathbf{v}_1) = \left\| \left(\frac{1}{2\sigma_e^2}\mathbf{\Psi}_2^T\mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}\mathbf{X}^T\mathbf{X} \right)^{-\frac{1}{2}} \left(\frac{1}{2\sigma_e^2}\mathbf{\Psi}_2^T\mathbf{\Psi}_1 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}\mathbf{X}^T\mathbf{Y}\mathbf{b} \right) - \left(\frac{1}{2\sigma_e^2}\mathbf{\Psi}_2^T\mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}\mathbf{X}^T\mathbf{X} \right)^{\frac{1}{2}} \mathbf{v}_1 \right\|_2^2 + \sum_{i=1}^{m} \lambda \left\| \mathbf{v}_{1,G_i} \right\|_2, \tag{4.16}$$

where

$$\mathbf{\Psi}_1 = \begin{pmatrix} \mathbf{X}_{\cdot 1} \\ \mathbf{X}_{\cdot 2} \\ \vdots \\ \mathbf{X}_{\cdot p} \end{pmatrix}, \quad \mathbf{\Psi}_2 = \begin{pmatrix} \mathbf{Y}\mathbf{b}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Y}\mathbf{b}_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{Y}\mathbf{b}_1 \end{pmatrix}.$$

By seperating each column of the matrix $\mathbf{X} - \mathbf{Y}\mathbf{b}_1 (\mathbf{v}_1)^T$ in the first term of the equation (4.15) and stacking one on another, it can be shown that its Frobenius norm is equivalent to a vector norm below,

$$\left\| \mathbf{X} - \mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T \right\|_F^2 = \left\| \mathbf{\Psi}_1 - \mathbf{\Psi}_2\mathbf{v}_1 \right\|_2^2.$$

Expansion of each of the first two terms in (4.15) shows,

$$\begin{aligned} \left\| \mathbf{X} - \mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T \right\|_F^2 &= \left\| \mathbf{\Psi}_1 - \mathbf{\Psi}_2\mathbf{v}_1 \right\|_2^2 \\ &= (\mathbf{\Psi}_1^T - \mathbf{v}_1^T\mathbf{\Psi}_2^T)(\mathbf{\Psi}_1 - \mathbf{\Psi}_2\mathbf{v}_1) \\ &= \mathbf{\Psi}_1^T\mathbf{\Psi}_1 - 2\mathbf{v}_1^T\mathbf{\Psi}_2^T\mathbf{\Psi}_1 + \mathbf{v}_1^T\mathbf{\Psi}_2^T\mathbf{\Psi}_2, \\ \left\| \mathbf{Y}\mathbf{b}_1 - \mathbf{X}\mathbf{v}_1 \right\|_2^2 &= (\mathbf{b}^T\mathbf{Y}^T - \mathbf{v}_1^T\mathbf{X}^T)(\mathbf{Y}\mathbf{b} - \mathbf{X}\mathbf{v}_1) \\ &= \mathbf{b}^T\mathbf{Y}^T\mathbf{Y}\mathbf{b} - 2\mathbf{v}_1^T\mathbf{X}^T\mathbf{Y}\mathbf{b} + \mathbf{v}_1^T\mathbf{X}^T\mathbf{X}\mathbf{v}_1. \end{aligned}$$

Now,

$$\begin{aligned} &\frac{1}{2\sigma_e^2}\left\| \mathbf{X} - \mathbf{Y}\mathbf{b}_1\mathbf{v}_1^T \right\|_F^2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}\left\| \mathbf{Y}\mathbf{b}_1 - \mathbf{X}\mathbf{v}_1 \right\|_2^2 \\ &= \mathbf{v}_1^T \left(\frac{1}{2\sigma_e^2}\mathbf{\Psi}_2^T\mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}\mathbf{X}^T\mathbf{X} \right)\mathbf{v}_1 - 2\mathbf{v}_1^T \left(\frac{1}{2\sigma_e^2}\mathbf{\Psi}_2^T\mathbf{\Psi}_1 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}\mathbf{X}^T\mathbf{Y}\mathbf{b}_1 \right) \\ &\quad + \left(\frac{1}{2\sigma_e^2}\mathbf{\Psi}_1^T\mathbf{\Psi}_1 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2+\sigma_e^2)}\mathbf{b}_1^T\mathbf{Y}^T\mathbf{Y}\mathbf{b}_1 \right) \end{aligned}$$

$$= \left\| \left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{X} \right)^{-\frac{1}{2}} \left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_1 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{Y} \mathbf{b}_1 \right) \right.$$
$$\left. - \left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{X} \right)^{\frac{1}{2}} \mathbf{v}_1 \right\|_2^2 + \text{constant}.$$

Therefore minimizing the log likelihood (4.15) is equivalent to minimizing the function (4.16) under the condition that,

$$\Lambda_i(\mathbf{T}) = \Lambda_i \left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{X} \right) > 0, \text{ for all } i, \tag{4.17}$$

where $\Lambda_i(\mathbf{A})$ is the ith largest eigenvalue of a matrix $\mathbf{A}$. Since $\mathbf{\Psi}_2$ is full rank, the matrix $\mathbf{T}$ is also full rank. Eigenvalue analysis of $\mathbf{T}$ shows that $\mathbf{T}$ is not positive definite if there is no supervision effect, i.e., $\mathbf{b}_1 = 0$.

4.8.3 Details of Stopping Rule Lemma

Since the rank of the multi-block data set $\mathbf{X}$ is directly related to the number of variation directions $\hat{\mathbf{v}}_i$, we focus on the estimation step of $\hat{\mathbf{v}}_i$ described in (4.16). According to [18], given a lasso penalty parameter λ, we must have $\mathbf{v}_{1,G_i} = \mathbf{0}$ if,

$$\|\mathbf{Z}_i^T \mathbf{r}_i\|_2 < \lambda, \tag{4.18}$$

where $\mathbf{Z}_i$ is a column slice of the matrix,

$$\left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{X} \right)^{\frac{1}{2}},$$

corresponding the the ith group and $\mathbf{r}_i$ is the ith partial residual defined as,

$$\mathbf{r}_i = \left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{X} \right)^{-\frac{1}{2}} \left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_1 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{Y} \mathbf{b} \right) - \sum_{j \neq i} \mathbf{Z}_j \mathbf{v}_{1,G_j}.$$

Since we are looking for the condition such that $\mathbf{v}_{1,G_i} = \mathbf{0}$ for all $i = 1, 2, .., m$, the previous condition (4.18) becomes,

$$\|\mathbf{Z}^T \mathbf{r}\|_2 < \lambda,$$

where,

$$\mathbf{Z} = \left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{X} \right)^{\frac{1}{2}},$$

$$\mathbf{r} = \left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_2 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{X} \right)^{-\frac{1}{2}} \left(\frac{1}{2\sigma_e^2} \mathbf{\Psi}_2^T \mathbf{\Psi}_1 - \frac{\sigma_{f_1}^2}{2\sigma_e^2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{Y}\mathbf{b} \right).$$

Noting that the square root term and the inverse square root term are canceled and that $\mathbf{\Psi}_2^T \mathbf{\Psi}_1 = \mathbf{X}^T \mathbf{Y}\mathbf{b}$ and using Cauchy-Schwarz inequality, we have $\mathbf{v}_1 = \mathbf{0}$ if,

$$\left\| \frac{1}{2(\sigma_{f_1}^2 + \sigma_e^2)} \mathbf{X}^T \mathbf{Y}\mathbf{b} \right\|_2 < \sqrt{\frac{1}{2(\sigma_{f_1}^2 + \sigma_e^2)}} \left\| \mathbf{X} \right\|_2 \left\| \mathbf{Y}\mathbf{b} \right\|_2 < \lambda.$$

BIBLIOGRAPHY

[01] J. Ahn, J.S. Marron, K. Muller, and Y. Chi. The high dimension, low sample size geometric represetnation holds under mild conditions. *Biometrika*, 94(3):1–7, 2007.

[02] Z. Bai and Y. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21(3):1275–1294, 1993.

[03] F. Bookstein. The measurement of biological shape and shape change. *Systematic Zoology*, 29(1):102–104, 1980.

[04] F. Bookstein. *Morphometric Tools for Landmark Data: Geometry and Biology*. 1997.

[05] M Bredel, D. M. Scholtens, G. R., C. Bredel, J. P. handler, J. J. Ren-frow, A. K. Yadav, H. Vogel, A. C. Scheck, R. Tibshirani, and B. I. Si-kic. A network model of a cooperative genetic landscape in brain tumors. *JAMA*, 302:261–275, 2009.

[06] D. Chen and H-G Müller. Nonlinear manifold representations for functional data. *The Annals of Statistics*, 40(1):1–29, 2012.

[07] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1979.

[08] Carl de Boor. *A Practical Guide to Splines*. Springer, 2001.

[09] T. Gasser and A. Kneip. Searching for structure in curve samples. *Journal of the American Statistical Association*, 90(432):1179–1188, 1995.

[10] T. Gasser, W. Köhler, H-G Müller, A. Kneip, R. Largo, L. Molinari, and A. Prader. Velocity and acceleration of height growth using kernel estimation. *Annals of Human Biology*, 11(5):397–411, 1984.

[11] T. Gasser, H-G Müller, W. Köhler, L. Molinari, and A. Prader. Nonparametric regression analysis of growth curves. *The Annals of Statistics*, 12(1):210–229, 1984.

[12] D. Gervini. Warped functional regression. *Biometrika*, 102(1):1–14, 2015.

[13] D. Gervini and T. Gasser. Self-modeling warping functions. *Journal of the Royal Statistical Society*, 66(4):959–971, 2004.

[14] P. Hadjipantelis, J. Aston, H-G Müller, and J. Evans. Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of mandarin chinese. *Journal of the American Statistical Association*, 110(510):545–559, 2015.

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

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

[17] D. Harvile. Matrix algebra from a statistician's perspective. *Springer*, 2000.

[18] T. Hastie, R. Tibishirani, and M. Wainwright. Statistical learning with sparcity. *CRC Press*, 2015.

[19] G. He, H-G Müller, and J. Wang. Functional canonical analysis for square-integrable stochastic processes. *Journal of Multivariate Analysis*, 85(1):54–77, 2003.

[20] H. Hotelling. Relations between two sets of bariants. *Biometrika*, 28:321–377, 1936.

[21] G.M. James. Curve alignment by moments. *Annals of Applied Statistics*, 1(2):480–501, 2007.

[22] I. Johnstone and A. Lu. Sparse principal components analysis. *Technical report, Stanford University*, 2009.

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

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

[25] A. Kneip and T. Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20(3):1266–1305, 1992.

[26] A. Kneip and J.O. Ramsay. Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483):1155–1165, 2008.

[27] S. Kurtek, W. Wu, G.E. Christensen, and A. Srivastava. Segmentation, alignment and statistical analysis of biosignals with application to disease classification. *Journal of Applied Statistics*, 40(6):1270–1288, 2013.

[28] M. Lee. Continuum direction vectors in high dimensional low sample size data. *Dissertation, University of North Carolina at Chapel Hill*, pages 54–87, 2007.

[29] S.E. Leurgans, R.A. Moyeed, and B.W. Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society*, 55(3):725–740, 1993.

[30] E. Levina and P. Bickel. The earth movers distance is the mallows distance: Some insights from statistics. *Computer Vision*, 2:251–256, 2001.

[31] G. Li and S. Jung. Supervised integrated factor analysis for multi-view data. *Submitted to Biometrics*, 2016.

[32] G. li, D. Yang, A. Nobel, and H. Shen. Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis*, 2015.

[33] J. Liu, S. Ji, and J. Ye. http://www.yelab.net/software/slep/. Retreived on Sep.28.2000.

[34] X. Liu and H-G Müller. Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association*, 99(467):687–699, 2004.

[35] E. Lock, K. Hoadley, J.S. Marron, and A. Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, 2013.

[36] X. Lu and J.S. Marron. Principal nested spheres for time warped functional data analysis. *arXiv*, 2013.

[37] J.S. Marron, J.O. Ramsay, L.M. Sangalli, and A. Srivastava. Functional data analysis of amplitude and phase variation. *MOX-Report No.27/2015*, 2015.

[38] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008.

[39] National Institutes of Health. http://cancergenome.nih.gov/. Retreived on Sep.28.2000.

[40] D. Paul. Asymptotics of the leading sample eigenvalues for a spiked covariance model. *Technical report, Stanford University*, 2005.

[41] M. E. Peter. Targeting of mrnas by multiple mirnas: The next step. *Oncogene*, 29:2161–2164, 2010.

[42] J.O. Ramsay, V.L. Gracco K.G. Munhall, and D.J. Ostry. Functional data analyses of lip motion. *Acoustical Society of America*, 99(6):3718–3727, 1996.

[43] J.O. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society*, 60(2):351–363, 1998.

[44] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, second edition, 2005.

[45] R.D.Tuddenham and M.M. Snyder. Physical growth of california boys and girls from birth to eighteen years. *Univ. of Calif. Publications in Child Development*, 1(2):183–364, 1954.

[46] Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, and Hayes DN. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell*, 17(1):98–110, 2010.

[47] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. *Proceedings of the 1998 IEEE International Conference on Computer Vision*, 1998.

[48] D. Samarov. The analysis and advanced extensions of canonical correlation analysis. *Dissertation, University of North Carolina at Chapel Hill*, pages 121–176, 2009.

[49] L.M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. K-mean alignment for curve clustering. *Computational Statistics and Data Analysis*, 54(5):1219–1233, 2010.

[50] B.W. Silverman. Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society*, 57(4):673–689, 1995.

[51] A. Srivastava, W. Wu, S. Kurtek, E. Klassen, and J.S. Marron. Registration of functional data using fisher-rao metric. *arXiv*, 2011.

[52] R. Tang and H-G Müller. Pairwise curve synchronization for functional data. *Biometrika*, pages 875–889, 2008.

[53] J.D. Tucker, W. Wu, and A. Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics and Data Analysis*, 60:50–66, 2013.

[54] D. Witten and R. Tibishirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

[55] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68(1):49–67, 2006.