이학박사학위논문

Dimension Reduction Methods for Multi-source and Manifold-valued Data

2023년 2월

서울대학교 대학원

통계학과

최서원

Dimension Reduction Methods for Multi-source and Manifold-valued Data

지도교수 정성규

이 논문을 이학박사 학위논문으로 제출함
2022년 12월

서울대학교 대학원
통계학과
최서원

최서원의 이학박사 학위논문을 인준함
2022년 12월

위원장 임채영 (인)

부위원장 정성규 (인)

위원 오희석 (인)

위원 김용대 (인)

위원 김병원 (인)

Dimension Reduction Methods for Multi-source and Manifold-valued Data

By

SeoWon Choi

A Thesis
Submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
February, 2023

ABSTRACT

Dimension Reduction Methods for Multi-source and Manifold-valued Data

SeoWon Choi
The Department of Statistics
The Graduate School
Seoul National University

In this thesis, we discuss dimension reduction methods in non-Euclidean space. Due to non-zero curvature, we cannot make use of the traditional techniques like Pythagorean theorem in building a statistical method in non-Euclidean space. To capture the structure of data set, it is necessary to understand the geometric nature of a given non-Euclidean space. We propose the following two dimension reduction methods, generalizing popular multivariate data analysis methods, factor analysis and PCA, to non-Euclidean settings.

In Chapter 2, we propose Principal Structure Identification (PSI) for multi-source dataset.

Analysis of multi-source dataset, where data on the same objects are collected from multiple sources, is of rising importance in many fields, most notably in multi-omics biology. We propose a novel framework and algorithms for integrative decomposition of such multi-source data, to identify and sort out common factor scores in terms of whether the scores are relevant to all data sources (fully joint), to some data sources (partially joint), or to a single data source.

The key difference between our proposal and existing approaches is that we utilize raw source-wise factor score subspaces in the identification of the partially-joint block-wise association structure. To identify common score subspaces, which may be partially joint to some of data sources, from noisy observations, our proposed algorithm sequentially computes one-dimensional flag means among source-wise score subspaces, then collects the subspaces that are close to the mean.

In Chapter 3, we propose Penalized Principal Nested Spheres (PenPNS) for dataset on the hypersphere surface.

Analysis of Principal Nested Spheres (PNS) (Jung, 2012) is a flexible dimension reduction method for dataset on the hypersphere, *e.g.* directional data (Fisher,1993; Fisher et al.,1993; Mardia and Jupp,2000) and shape data (Kendall,1984; Dryden and Mardia, 1998). In PNS, the dimension reduction is an iterative procedure for discarding unimportant dimensions. It is specifically designed to capture a certain type of non-geodesic variation by fitting a small sphere.

However, PNS suffers from overfitting, a phenomenon where data points are fitted with a small sphere even though they are generated from along a great sphere. We consider two types of overfitting phenomena. (1) When the estimated radius $\widehat{r}$ is over $\pi/2$, the estimated axis $\widehat{v}$ is flipped to $-\widehat{v}$ and we take the radius as $\pi - \widehat{r}$, since the radius as parameter ranges from 0 to $\pi/2$. Then the distribution of the estimated radius becomes a folded version of the distribution between the true axis and the data points. Thus the expectation of the estimated radius is less than $\pi/2$. (2) When data points are generated along a great sphere but within a short interval, the data point cloud has a disc shape and is usually fitted by a small sphere with a very small radius.

PenPNS is an improvement of PNS that overcomes the overfitting phenomena in small sphere fitting. To deal with the first type of overfitting phenomenon, PenPNS regularize radius in estimation, where the value of the penalty term grows larger as radius decreases departing from $\pi/2$. For the second type, PenPNS gives a penalizing term on the cross-validation error in choosing tuning parmeter. The penalizing term, called Index of Dispersion, has a larger value in the case a disc-shaped distribution is fitted with a small radius. In Simulation Study and Real Data analysis, we demonstrate that PenPNS successfully mitigates the overfitting phenomena.

Keywords: Dimension reduction, Non-Euclidean space, Multi-source dataset, Dataset on the hypersphere, Multi-omics dataset.
Student Number: 2017 – 38449

Contents

3 Penalized Principal Nested Spheres 47

List of Tables

List of Figures

Chapter 1

Introduction

Recent days, statistical methods for dataset in non-Euclidean space have gained much attention, in need of analyzing unconventional datasets, such as directional data (Fisher,1993; Fisher et al.,1993; Mardia and Jupp,2000) and shape data (Kendall,1984; Dryden and Mardia,1998), or exploring the geometric interpretations of existing methods, such as PCA (Hotelling,1993). However, due to non-zero curvature, applying even elementary statistical methods to non-Euclidean space is a challenging task. The difficulties come from two ways: (1) identify the mathematical entity of the dataset, (2) design an estimate that explains the data set well, but also reflects well the geometric nature of the space where the data points dwell in.

For example, defining even a basic statistical concept in non-Euclidean space can be puzzling, *e.g.* defining “mean” on the sphere surface, S^2. We may raise the following questions: Is there any concrete way of defining “mean” on the sphere, such as “summing up data points” on the sphere surface or “divide the sum by the number of points"? What would be the "geometrical entity" of that mean? Does it even reside on the same original sphere surface as data points? In building up an optimization problem to find mean, how can we measure a distance between two points on S^2?

In this thesis, we consider two types of dataset.

1. Multi-source dataset (*see* Chapter 2)

2. Dataset on the hypersphere surface (*see* Chapter 3).

Before jumping into methodologies, we need to provide mathematical frameworks for these unconventional types of dataset. In the next sections in this chapter, we give mathematical preliminaries for the thesis.

We want to clarify the relation between our dataset and non-Euclidean geometry. The non-Euclidean spaces upon which our data sets reside are described as manifolds. Specifically, we explain (1) Grassmann manifold for multi-source dataset and (2) the hypersphere S^d for dataset on the hypersphere surface.

We also provide related geometrical concepts therein, *e.g.* (1) a concept of "distance" between two points on each manifold or (2) an analogue of "mean" for data points in non-Euclidean space. We utilize geometrical concepts in building statistical methods later.

We then present the proposed statistical methods to analyze our datasets in the next chapters:

1. Partially-joint Structure Indentification (PNS) for multi-source dataset (*see* Chapter 2)

2. Penalized Principal Nested Sphere (PenPNS) for dataset on the hypersphere surface (*see* Chapter 3).

A detailed introduction to each method is deferred to the chapter of its own, *see* Section 2.1 and Section 3.1 respectively.

1.1 Mathematical Backgrounds for Multi-source Dataset

By the *multi-source data*, we mean data obtained for multiple groups of variables on the same set of subjects. In other words, there are K data blocks, $X_k \in \mathbb{R}^{p_k \times n}$, for $k = 1, \ldots, K$, observed for matched n subjects. Assuming zero mean, each data block is decomposed into

$$X_k = Z_k + E_k = U_k V_k^T + E_k, \quad k = 1, \ldots, K, \tag{1.1}$$

in which the low-rank "signal" matrix Z_k (say, of rank r_k) is factored into a loading matrix U_k and a score matrix V_k. The score matrix $V_k \in \mathbb{R}^{n \times r_k}$ is an orthonormal matrix and can be viewed as a representation of a r_k-dimensional linear subspace in $\mathbb{R}^n$. We denote the score subspace spanned by the columns of V_k as $[V_k]$ and say that V_k represents the subspace $[V_k]$.

In the language of matrix manifolds, each score subspace $[V_k]$ is described as an element of the Grassmann manifold. Formally, the p-dimensional *Grassmann manifold* in $\mathbb{R}^n$, $\mathrm{Gr}(n, p)$, is defined as the set of all p-dimensional linear subspaces of the n-dimensional vector space $\mathbb{R}^n$.

The problem here is that score subspace $[V_k]$ have different dimensions each other. In this section, we want to define a notion of distance among linear subspaces of different dimensions (for short, *subspace distance*). In other words, it is a distance $d(A, B)$ with $A \in \mathrm{Gr}(n, p_1)$ and $B \in \mathrm{Gr}(n, p_2)$ which allows $p_1 \neq p_2$. We here adopt the framework of Ye and Lim (2016). In fact, the framework is a generalization of distance among linear subspaces of the same dimension to the one of different dimensions.

The Framework of Ye and Lim (2016)

Let $p_1 \leq p_2$. Suppose we have a distance d' on $\mathrm{Gr}(n, p_2)$, that is, among linear subspaces of the same dimension p_2. A distance d between $[A] \in \mathrm{Gr}(n, p_1)$ and $[B] \in \mathrm{Gr}(n, p_2)$ is derived as

$$d([A],[B]) = \min\left\{d'([X],[B]) : [X] \in \Omega([A])\right\},$$

where $\Omega([A]) = \{[X] \in \mathrm{Gr}(n, p_2) : [A] \subset [X]\}$. For more details, *see* Theorem 7 and 12 in Ye and Lim (2016).

In Chapter 2, we use the *Frobenius norm distance* which is defined as

$$d([A],[B]) = \frac{1}{\sqrt{2}}\|AA^T - BB^T\|_F,$$

where $\|\cdot\|_F^2$ is the Frobenius norm and A and B are orthonormal matrices which represent $[A] \in \mathrm{Gr}(n, p_1)$ and $[B] \in \mathrm{Gr}(n, p_2)$. (To put it precisely, we generalize the Frobenius norm distance among linear subspaces of the same dimension to of different dimensions using Ye and Lim framework.)

Using the concept of subspace distance, we can define an analogue of "mean" for a set of linear subspaces. One of the examples is the one-dimensional mean among linear subspaces of different dimensions (for short, *mean direction*).

Suppose linear subspaces $[V_k]$ of dimension r_k in $\mathbb{R}^n$ for $k = 1, \dots, K$ are given. The mean direction $\hat{w}$ minimizes the sum of the squares of subspace distances d between a candidate w and subspaces $[\widehat{V}_k]$ for $k \in S_i$,

$$\hat{w} = \underset{w^T w = 1}{\arg\min} \sum_{k \in S_i} d\left([w], [\widehat{V}_k]\right)^2 . \tag{1.2}$$

When we use the Frobenius norm distance for d, the mean direction $\hat{w}$ can be computed easily using the truncated SVD, *see* Section A.1.2 in the supplementary materials.

The concept of subspace distance among linear subspaces of different dimensions will play a crucial role in Section 2.3.1. The concept of mean direction, or one-dimensional average among linear subspaces of different dimensions will be used in Section 2.3.1, notably (2.4).

1.2 Mathematical Backgrounds for Dataset on Hypersphere

Our aim is to regard the hypersphere as a Riemannian manifold. We begin by presenting a list of terminology.

A manifold of dimension d is a Hausdoff space with a countable basis. Each point p of a d-dimensional manifold M has a neighborhood that is homeomorphic to an open subset of real space $\mathbb{R}^d$. Thus one can attach to every point p a real vector space and we call this real vector space a *tangent space*, T_pM.

The set of all local homeomorphisms (U, ϕ) is called *atlas*, such that $U \subset M$ is open and $\phi : U \to \phi(U) \subset \mathbb{R}^d$. If a manifold M is equipped with a differentiable atlas, M is called a smooth manifold. If a smooth manifold M is equipped with smoothly varying inner product $\langle\, , \rangle_p$ on tangent bundle $TM = \cup_{p \in M} T_p M$, M is called a Riemannian manifold.

To define a distance on a Riemannian manifold, we need to define a length of a curve. Let a smooth map $\gamma : [a, b] \subset \mathbb{R} \to M$ be a *curve* in M. Indeed, for any $t \in [a, b]$, $\dot{\gamma}(t) = \mathrm{d}\gamma(\mathrm{d}/\mathrm{d}t)$ is a tangent vector in $T_{\gamma(t)}M$. For a curve γ such that $\dot{\gamma}(t) \neq 0$ for all $t \in [a, b]$, the length of γ is defined as

$$\ell_{[a,b]}(\gamma) = \int_a^b \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle_{\gamma(t)}}\, \mathrm{d}t.$$

Then, the *geodesic distance* between two points $p, q \in M$ is defined as the shortest length of all curves joining p and q. The shortest curve joining two points on M is called a *geodesic*.

For each $p \in M$, a geodesic at p with direction $v \in T_p M$ is a curve $\gamma : [0, 1] \to M$ such that $\gamma(0) = p$, $\dot{\gamma}(0) = v$ and $\|\dot{\gamma}(t)\| = \|v\|$ for $t \in [0, 1]$. Such a geodesic exists uniquely by the Picard's existence theorem. Then the *exponential map* at p is defined as a differentiable map $\mathrm{Exp}_p : U \subset T_p M \to M$ such that $\mathrm{Exp}_p = \gamma(1) \in M$.

For more mathematical details about Riemannian manifold, *see* Boothby (1986) and Lee (2006).

Now we can check that the hypersphere S^d, defined as

$$S^d = \{x \in \mathbb{R}^{d+1} : x^T x = 1\},$$

has the following facts as a Riemannian manifold:

- The geodesic distance between two points $p, q \in S^d$ is defined as the angle between p and q at the center of S^d.
- We can explicit exponential maps of S^d using the coordinates of $\mathbb{R}^{d+1}$.

  Without loss of generality, let $p = (0, \ldots, 0, 1) \in S^d$ be the north pole of S^d. The exponential map at p is expressed as
  $$\mathrm{Exp}_p(v) = \left(v_1 \frac{\sin \|v\|}{\|v\|}, \ldots, v_d \frac{\sin \|v\|}{\|v\|}, \cos \|v\| \right)$$
  with $v = (v_1, \ldots, v_d) \in T_p S^d$. The inverse exponential map at p is expressed as
  $$\mathrm{Exp}_p^{-1}(w) = \left(w_1 \frac{\theta}{\sin\theta}, \ldots, w_d \frac{\theta}{\sin\theta} \right),$$
  where $w = (w_1, \ldots, w_{d+1}) \in \mathbb{R}^{d+1} \backslash (0, \ldots, 0, -1)$ with $w^T w = 1$ and $\theta = \arccos(w_{d+1})$.

By exploiting the concepts of geodesic and geodesic distance, we can define an analogue of "mean" for dataset on the hypersphere S^d. One of the examples is *Fréchet mean* (Fréchet,1948; Bhattacharya and Patrangenaru,2003).

We first consider X to be a S^d-valued random variable, which is a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ into $(S^d, \mathcal{G})$, where $\mathcal{G}$ denotes the Borel σ-algebra of S^m. Let $X_1, \ldots, X_n$ be random variables generated identically and independently from the same law as X.

Fréchet mean minimizes the least expected squared geodesic distance between a candidate Fréchet mean and data point. Specifically, the population and sample Fréchet mean set is defined as
$$F = \left\{ p \in S^d : \mathbb{E}_X \rho^2(X, p) = \inf_{q \in S^d} \mathbb{E}_X \rho^2(X, q) \right\}$$

and

$$F_n = \left\{ p \in S^d : \mathbb{E}_X \rho^2(X, p) = \inf_{q \in S^d} \mathbb{E}_X \rho^2(X, q) \right\}.$$

Here the Fréchet mean sets F and F_n are guaranteed to be the set of points that reside on S^d. Compare them with a naive definition of mean on S^d: vector-sum all data points on S^d and scalar-divide with the number of data points, n — the result does not even locate on the hypersphere surface, S^d. We see that building a statistical descriptor using a geometrical information (such as geodesic distance) instrinsic to manifold can enhance the interpretablity of data set.

The notion of geodesic distance on S^d will play a key role in Chapter 3. Fréchet mean will be generalized to sphere descriptor of data points (called small/great sphere estimator in there), *see* Section 3.2 and 3.4.

Chapter 2

Partially-joint Structure Identification

2.1 Introduction

In various fields of science and technology, there is a growing interest in analyzing multi-source data in an integrative way. By the *multi-source data*, we mean data obtained for multiple groups of variables on the same set of subjects. Each group of variables is observed from a common source, and form a data block. A prominent example of multi-source data is modern multi-omics data that include gene expressions, RNA sequencing, mutations, epigenetic markers or metagenomic materials (Subramanian et al.,2020). The recent development of high-throughput technologies enables us to extract these sources of information comprehensively from a given preparation of cancer/normal tissue samples (Reuter et al.,2015; Norris et al.,2017).

One of the main challenges in analyzing multi-omics data is that data blocks come from distinct measurements of different sources. For example, our motivating data set consists of three data blocks, each from a drug response panel, genome-wide DNA methylation profiles and RNA sequencing profiles (Dietrich et al.,2018); *see* Section 2.6 for a detailed description of the data.

Separately analyzing each data block hinders the assessment of inter-relations among different data blocks. To capture the potentially joint association structures in these multi block data, the *linked component model* has been oftentimes used (Smilde et al.,2003; VanDeun et al.,2009; Lock et al.,2013).

Suppose that there are K data blocks, $X_k \in \mathbb{R}^{p_k \times n}$, for $k = 1, \ldots, K$, observed for n subjects. The subjects in the data are common and matched, i.e., the measurements for the ith subject appear in the ith column of each data matrix. Assuming zero mean, each data block is decomposed into

$$X_k = Z_k + E_k = U_k V_k^T + E_k, \quad k = 1, \ldots, K, \tag{2.1}$$

where the low-rank “signal” matrix Z_k is factored into a loading matrix U_k and a score matrix V_k, which is perturbed by “noise” matrix E_k. The linked component model further assumes that two or more data blocks can potentially share a common score component VanDeun et al. (2009). An extreme example is that all scores are common to all data blocks, that is, $V_1 = \cdots = V_k$, as in Smilde et al. (2003). The models considered in Lock et al. (2013) and Feng et al. (2018) allow some scores to be common to all data blocks, explaining the joint variation among all data blocks, and some scores to be specific to each data blocks (explaining the individual variations in a single data block).

Following Gaynanova and Li (2019) and Gao et al. (2020), we in addition allow *partially-joint* scores that are shared across multiple, but not necessarily all, data blocks. An illustrative example of such a model, for $K = 3$ blocks of data, is $(X_1^T, X_2^T, X_3^T)^T = UW^T + E$,

$$UW^T = \begin{pmatrix} U_{(1),1} & U_{(1),2} & 0 \\ U_{(2),1} & U_{(2),2} & 0 \\ U_{(3),1} & 0 & U_{(3),3} \end{pmatrix} \begin{pmatrix} W_{\{1,2,3\}} & W_{\{1,2\}} & W_{\{3\}} \end{pmatrix}^T, \tag{2.2}$$

where U is the loading matrix, $W_{\{1,2,3\}}$ is a matrix of scores that affect all data blocks, and $W_{\{1,2\}}$ is a partially-joint score matrix, affecting only the first two data blocks, but not the third. The scores in $W_{\{3\}}$ are specific to the third data block. Our goal is to delineate such an association structure from a multi-source data.

In this chapter, we develop a novel framework and estimation strategy for integrative decomposition of multi-source data, by identifying scores that are fully joint to all data sources, partially joint to some, or specific to a single data source. The framework utilizes the signal score subspace $[V_k]$, a rank r_k subspace of $\mathbb{R}^n$, spanned by the rows of the signal matrix Z_k in (2.1). The rationale for using $[V_k]$ is straightforward: If a common score of rank r is shared by the first two data blocks, then the intersection of $[V_1]$ and $[V_2]$ is non-empty, and is a dimension r subspace. To capture the partially-joint block-wise association structure from the score subspaces $[V_k]$, we define the collection of tuples

$$\mathfrak{S} := \mathfrak{S}(\{[V_k]\}_{k=1,\ldots,K}) = \{(S, r(S)) : S \subset \mathcal{K}\},$$

in which a non-empty subset S of $\mathcal{K} = \{1, \ldots, K\}$ denotes a specific pattern of block-wise association, W_S denotes the corresponding common scores (that are either fully-joint, partially-joint or specific to data sources) of rank $r(S)$. See Section 2.2 for a detailed construction of $\mathfrak{S}$, and Section 2.4 for conditions to guarantee the uniqueness of $\mathfrak{S}$.

To identify the partially-joint score structure $\mathfrak{S}$ from noisy observations, we propose to compute the flag mean of signal subspaces (Draper,2014), and to test whether the mean is indeed close to signal subspaces. If a mean w is "close" to $[V_1]$ and $[V_2]$, but not to other subspaces, then it is a basis of (estimated) $W_{\{1,2\}}$. We use a tunable parameter to determine the closeness between two subspaces. The algorithm, detailed in Section 2.3, is quick in decomposing multi-source data sets, and boasts a superior performance in identifying $\mathfrak{S}$ and in the estimation of the subspaces spanned by the loading matrices and by the common score matrices.

Recently, there has been a growing interest in integrative decomposition of multi-source data (Lock et al.,2013; Li and Jung,2017; Feng et al.,2018; Li and Gaynanova,2018; Gaynanova and Li,2019; Gao et al.,2020). Among these, Gaynanova and Li (2019) and Gao et al. (2020) also considered modeling partially-joint association structures. However, these authors focused on the loading matrix U, and exploited the source-wise sparse structure of the matrix U, as seen in (2.2). On the contrary, we explicitly utilize the signal score matrices V_k in (2.1) in identifying partially-joint scores of (2.2). The loading matrix is naturally obtained as a subsequent step in our proposal. Our approach extends the angle-based joint and individual variation identification of Feng et al. (2018), in which partially joint variations were not considered. In Section 2.5, we numerically confirm that our proposal finds the true association structure much more accurately than the aforementioned decomposition methods.

The rest of chapter is organized as follows. In Section 2.2, we formally present our integrative decomposition framework and define the partially joint structure $\mathfrak{S}$, followed by the proposed decomposition algorithm and tuning parameter selection procedure in Section 2.3. In Section 2.4, conditions to guarantee unique decompositions are discussed. Section 2.5 is devoted to numerical illustrations and comparisons to existing methods. In Section 2.6, we demonstrate the use of the proposed method in an analysis of a blood cancer multi-omics data set associated with drug responses, and reveal that the proposal detects a latent signal pattern, partially joint across two, but not all, data sources, which is not easily seen by a separate analysis of individual data blocks. Proofs, technical lemmas, examples and extended numerical results are given in the supplementary materials.

2.2 Statistical Framework

2.2.1 Partially-joint structures

Consider a set of row-centered data matrices $X_k \in \mathbb{R}^{p_k \times n}$ for $k = 1, \ldots, K$, where the ith column of each data matrix corresponds to the same (ith) subject. Let $p = \sum_k p_k$. We assume that each X_k is additively decomposed into the rank r_k true signal block Z_k and random errors E_k so that $X_k = Z_k + E_k$.

We consider a model that the signal $Z = (Z_1^T, \ldots, Z_K^T)^T$ is decomposed

$$Z = UW^T,$$

where each column of W has Euclidean norm 1, but not necessarily orthonormal. We assume that the number of columns of U and W is less than $\min(n, p)$.

Our goal is to find an association structure among Z_k's, where multiple, but not necessarily all, of Z_k's share a joint signal component. We describe this manner using a word "partially-joint".

We give partially-jointness to the model by assuming block-sparsity on U. For example, when $K = 3$, we can have a model

$$\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} U_{(1),1} & U_{(1),2} & 0 \\ U_{(2),1} & U_{(2),2} & 0 \\ U_{(3),1} & 0 & U_{(3),3} \end{pmatrix} \begin{pmatrix} W_1 & W_2 & W_3 \end{pmatrix}^T,$$

where the blocks $U_{(1),3}$, $U_{(2),3}$ and $U_{(3),2}$ are exactly zero. In this case, $(U^T_{(1),1}U^T_{(2),1}U^T_{(3),1})^T W_1^T$ is a signal component shared by all Z_1, Z_2 and Z_3 whereas $(U^T_{(1),2}U^T_{(2),2})W_2^T$ are shared "partially" by Z_1 and Z_2 and $U_{(3),2}W_3^T$ is the one specific to Z_3.

We associate each block-sparse pattern with with an *index-set*, a subset of $\{1, \ldots, K\}$. For each index-set S, we say the *rank* of S, or $r(S)$, the number of the columns of W associated with S. We summarize the block-sparsity as the *partially-joint structure*, the collection of pairs $(S, r(S))$. In our example, if W_1 and W_2 are composed of two and one columns, the partially-joint structure $\mathfrak{S}$ is given $((\{1, 2, 3\}, 2), (\{1, 2\}, 1), (\{3\}, 1))$.

For each index-set S, the columns of W associated with S are concatenate into a $n \times r(S)$ matrix, called *partially-joint score.* The columns of U associated with S are also concatenated into a $p \times r(S)$ matrix, which can be split into block matrices of $p_k \times r(S)$ corresponding to Z_k's. We call these block matrices *partially-joint loading.* From the block-sparsity, block matrices that corresponds to Z_k, $k \notin S$ is zero. In our example, W_1,W_2 and W_3 are partially-joint scores and $U_{(1),1}$, $U_{(2),1}$, $U_{(3),1}$, $U_{(2),1}$, $U_{(2),2}$ and $U_{(3),3}$ are partially-joint loadings.

For given Z_i's, our aim is to find partially-joint loadings U_i and scores W_i. Only with the block-sparsity constraint, there may be multiple candidates of both U and W. Now we give a plausible strategy of defining U and W — by finding the column space of each W_i.

Remark 2.1. *The decomposition of Z_k's into the partially-joint loading and score matrices is not unique. If instead of W_i (and its corresponding loading matrix $U_{(k),i}$), one chooses $W_i' = W_i R$, for an orthogonal matrix R, then the corresponding loading matrix becomes $U'_{(k),i} = U_{(k),i} R$. Nevertheless, if $U_{(k),i}$ is zero, then $U_{(k),i} R$ is zero for any orthogonal R, so the block-sparsity is invariant to* *the choices of basis.*

2.2.2 Partially-joint score subspace and index-set ordering

We now define the *partially-joint score subspace*, a linear subspace spanned by the columns of a partially-joint score, in a constructive manner. The definition depends mainly on the following two concepts, the signal score subspace of signal block Z_k and the ordering of index-sets.

Let the true signal matrix be further decomposed, $Z_k = U_k V_k$, as in the factor analysis, where $V_k \in \mathbb{R}^{n \times r_k}$ is the factor score matrix satisfying $V_k^T V_k = I_{r_k}$. We call the column space of V_k, denoted $[V_k]$, as the *signal score subspace* for Z_k. Note that $[V_k]$ is regarded as a rank-r_k subspace of $\mathbb{R}^n$, spanned by the columns of V_k.

Our geometrical implication is as follows: If, given a $S \subset \{1, \ldots, K\}$, $[V_k]$'s overlap, then the corresponding signal blocks Z_k, $k \in S$ share a common score, represented by the subspace $\cap_{k \in S} [V_k]$, which is at least partially-joint to signal blocks Z_k, $k \in S$. As we have assumed that signal blocks share a signal component through the corresponding partially-joint score, the intersections of signal score subspace plays an important role as a building block in constructing partially-joint score subspace.

However, one of the challenge of this approach is that index-sets, as subsets of $\{1, \ldots, K\}$, have a partially-ordered relation with respect to set-inclusion. As two index-sets can be either nested or intersected, the construction of a partially-joint score subspace on one index-set may affect the one on the other. Thus we suggest a plausible way of constructing partially-joint score subspaces by giving an ordering among index-sets.

We give an *ordering to index-sets*, sorted by the number of elements in descending order. Each index-sets is allocated with index numbers as S_i for $i = 1, \ldots, 2^K - 1$. We set $S_1 = \{1, \ldots, K\}$, the maximal set, and the next K index-sets are of size $K-1$, and then the next $K(K-1)/2$ index-sets are of size $k-2$, … *etc.* Note that index-set ordering is not unique for given K, since there can be permutations among index-sets of the same size.

Example 2.1. *For* $K = 3$*,* $S_1 = \{1, 2, 3\}$ *as appeared in the top row of Fig. 2.1, and the next three sets* (S_2, S_3, S_4) *are given by the next row of the figure. That is,* S_2*,* S_3*, and* S_4 *are* $\{1, 2\}$*,* $\{2, 3\}$*,* $\{1, 3\}$*, respectively. Likewise,* $S_5 = \{1\}$*,* $S_6 = \{2\}$*,* $S_7 = \{3\}$*. Each element* S_i *stands for the indices of data blocks that potentially possess a common score. Of course, such indexing is not unique, and one may set, e.g.,* $S_6 = \{3\}$*,* $S_7 = \{2\}$*.*

Conditions to guarantee invariance of the decomposition with respect to different choices of indexing will be discussed in Section 2.4.

Using the concepts of signal score subspace and index-set ordering, we now define partially-joint score subspace.

Definition 2.1. *Suppose matched data matrices* $X_k = Z_k + E_k \in \mathbb{R}^{p_k \times n}$ *for* $k = 1, \ldots, K$ *with true signal block* Z_k *are given. With an index-set ordering* $S_1, \ldots, S_{2^K - 1}$*, a set of corresponding partially-joint score subspace* $[W_i]$*'s for* $i = 1, \ldots, 2^K - 1$ *are constructed*

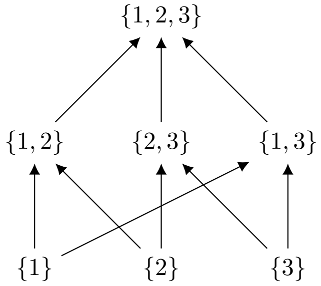


Figure 2.1: An example of index-set ordering is depicted as the indexed partially-ordered set. Arrow → stands for $\subset$, e.g. $\{1\} \to \{1,2\}$ means $\{1\} \subset \{1,2\}$.

sequentially

$$\begin{aligned}
[W_1] &:= \cap_{k\in S_1}[V_k],\\
[W_2] &:= P_1^{\perp}\left(\cap_{k\in S_i}[V_k]\right),\\
&\;\;\vdots\\
[W_i] &:= (\bigcirc_{\substack{\{j:j<i,\\ S_j\cap S_i\neq\phi\}}} P_j^{\perp})\left(\cap_{k\in S_i}[V_k]\right).
\end{aligned}$$

Here, $P_j^{\perp}$ is the projection transformation of $\mathbb{R}^n$ onto the orthogonal complement of $[W_j]$. The notation $\bigcirc_{j\in J}P_j^{\perp}([A])$, for an indexed set $J \subset \{1,\ldots,2^K-1\}$ and a subspace $[A]$, stands for the repeated applications of $P_j^{\perp}$ on $[A]$, where $P_j^{\perp}$s are applied one by one by the increasing order of $j \in J$.

Note that depending on $\{[V_k]\}$, a $[W_i]$ may be $\{0\}$.

Example 2.2. *Suppose $K = 3$ and index-set ordering is as in Example 2.1, with $r(S_1) = r(S_2) = r(S_4) = r(S_7) = 1$, and $r(S_i) = 0$ otherwise. Partially-joint score subspaces are obtained as*

$W_1 = \cap_{k \in S_1}[V_k]$, $W_2 = P_1^{\perp}(\cap_{k \in S_2}[V_l])$, $W_4 = P_2^{\perp} \circ P_1^{\perp}(\cap_{k \in S_4}[V_k])$ *and* $W_7 = P_4^{\perp} \circ P_1^{\perp}(\cap_{k \in S_7}[V_k])$.

Partially-joint score subspaces have the following basic properties.

(1) The partially-joint score subspaces involved with each Z_k can restore the signal score subspace, or

$$[V_k] = +_{i \in \{k \in S_i\}}[W_i], \tag{2.3}$$

as shown in Lemma A.3 in the supplementary material. Here the notation '+' means the sum of subspaces.

(2) For any $i \neq j$, $[W_i]$ and $[W_j]$ do not overlap, i.e., $[W_i] \cap [W_j] = \{0\}$, but they are not necessarily orthogonal. However, if both $[W_i]$ and $[W_j]$ are related to a common data source, then they are orthogonal:

Lemma 2.1. *For* $i, j \in \mathcal{I}_K$ *and* $S_i \cap S_j \neq \phi$, $[W_i] \perp [W_j]$.

Lemma 2.1 ensures that all partially-joint score subspaces relevant to the kth data block, $\{[W_{S_i}] : k \in S_i, i \in \mathcal{I}_K\}$, are orthogonal to each other. Thus, the subspaces in the right hand side of (2.3) are indeed orthogonal to each other.

2.2.3 Partially-joint score and loading

Given $\mathfrak{S}$, fix an orthogonal basis $W_i \in \mathbb{R}^{n \times r(S_i)}$ of $[W_i]$ for i such that $|S_i| > 0$. The potential partially-joint score is linked to the partially-joint loadings involved with Z_k by

$$U_{(k)} = Z_k W_{(k)},$$

where $U_{(k)}$ and $W_{(k)}$ are column-wise concatenations of U_i's and W_i's for $i \in \{i : k \in S_i\}$. Indeed we can easily check that $U_{(k)} W_{(k)}^T = Z_k W_{(k)} (W_{(k)}^T W_{(k)})^{-1} W_{(k)}^T = Z_k$, since $W_{(k)} (W_{(k)}^T W_{(k)})^{-1} W_{(k)}^T$ is the projection onto $[V_k]$ by the property (2.3) above. Then the partially-joint loadings $U_{(k),i}$ for $i \in \{i : k \in S_i\}$ are obtained by disjoining $U_{(k)}$. For $i \notin \{i : k \in S_i\}$, $U_{(k),i} = 0$.

2.3 Estimation

In practice, the signal Z_k and error E_k of row-centered data matrices $X_k \in \mathbb{R}^{p_k \times n}$ for $k = 1, \ldots, K$ are unknown. We assume that the ranks r_k of the signal matrices are pre-determined and extract the signal matrix using a low-rank approximation of X_k, and write the rank r_k approximation of X_k by $\widehat{Z}_k$. The basis of the signal score subspace $\widehat{V}_k$ is given either by the right singular vectors of Z_k from the SVD or by any off-the-shelf factor model estimates.

The overview of the estimation algorithm is as follows: (1) on the outer loop, we iterate through all index-sets S_i on a given ordering, and (2) on the inner loop, we obtain the *partially-joint score subspace estimate* $[\widehat{W_i}]$ from $\{[\widehat{V}_k]\}_{k \in S_i}$, identifying the basis of $[\widehat{W_i}]$ one-by-one.

The major challenge is that, we have the sample versions of signal blocks, $\widehat{Z}_k$, and their signal score subspaces, $[V_k]$, are contaminated by noise. Under noise, we may have $\cap_{k \in S_i} [\widehat{V}_k] = \{0\}$ even if $\cap_{k \in S_i} [V_k] \neq \{0\}$. Thus, there is a need to give a slack on identifying the "intersection" of $[V_k]$'s, accounting for random perturbations in $[\widehat{V}_k]$. We propose to use principal angles between subspaces for such identification, further developed in Section 2.3.1.

In Section 2.3.1, we propose a procedure that finds a basis matrix $\widehat{W}_i$ for the partially-joint score subspace estimate $[\widehat{W}_i]$ at each ith step. The rank of the corresponding S_i is set as $\hat{r}(S_i) = \mathrm{rank}(\widehat{W}_i)$. Finishing the iteration over all S_i's, an *estimated partially-joint structure* $\widehat{\mathfrak{S}} = \{(S_i, \hat{r}(S_i)\}$ and a set of corresponding $[\widehat{W}_i]$'s are obtained. In Section 2.3.2, we then discuss the estimation of corresponding partially-joint loading matrices from $\widehat{\mathfrak{S}}$ and $\widehat{w}$. In Section 2.3.3, a strategy for thresholding at the right principal angle is proposed.

2.3.1 Partially-joint Score Subspace Estimation

Our strategy on estimating partially-joint score subspace, $[\widehat{W}_i]$, is to estimate the partially-joint score subspace $[W_i]$ by collecting one-dimensional bases, that lie "close" to each and every score subspaces in $\{[\widehat{V}_k]\}_{k \in S_i}$.

For the notion of closeness, the principal angle between subspaces is used as a threshold parameter $\lambda \in [0, \pi/2)$. For now, λ is treated as a pre-determined tuning parameter, and a data-driven choice of λ will be discussed in Section 2.3.3.

For a given $\lambda \in [0, \pi/2)$, the estimation algorithm is applied sequentially for $i = 1, \ldots, 2^K - 1$. At the ith stage, we begin with $\mathcal{W}_i = \emptyset$, to which the identified one-dimensional bases of the partially-joint score subspace are added. Algorithm 1 summarizes the proposed partially-joint structure identification procedure, using steps (a)—(c).

Step (a) : Find a *mean direction* that lies closest to all score

```
Algorithm 1: Partially-joint Structure Identification
input: $\widehat{V}_1, \ldots, \widehat{V}_K, S_1, \ldots, S_{2^K-1}, \lambda$
for $i = 1, 2, \ldots, 2^K - 1$ do
    Set $\mathcal{W}_i = \phi$;
    while $\dim([\widehat{V}_k]) > 0$ for all $k \in S_i$ do
        (a) Compute the mean direction $\hat{w}$ of
         $\{[\widehat{V}_k]\}_{k \in S_i}$, see (2.4);
        if the condition (2.5) is satisfied then
            (b) Let $\mathcal{W}_i \leftarrow \mathcal{W}_i \cup \{\hat{w}\}$;
            (c) Update $\widehat{V}_k \leftarrow \widehat{V}_{k,trunc}$ for each $k \in S_i$
        else
            break;
        end
    end
    Let $\hat{r}(S_i) = |\mathcal{W}_i|$, write $\widehat{W}_i$ for the $n \times \hat{r}(S_i)$ matrix
     consisting of elements in $\mathcal{W}_i$ and record
     $(\widehat{S}_i, \hat{r}(S_i), \widehat{W}_i)$ in $\widehat{\mathfrak{S}}$;
end
Result: $\widehat{\mathfrak{S}}$
```

subspaces $\{[\widehat{V}_k]\}_{k \in S_i}$, as a candidate that may be included in $\widehat{W_i}$.

As a candidate for a basis vector of the partially joint score matrix, compute the mean direction $\hat{w}$ among $\{[\widehat{V}_k]\}_{k \in S_i}$, provided that $[\widehat{V}_k] \neq \{0\}$ for all $k \in S_i$. The mean direction minimizes the sum of the squares of the subspace distances between a candidate w and subspaces $[\widehat{V}_k]$ for $k \in S_i$, and is

$$\hat{w} = \underset{w^T w = 1}{\arg\min} \sum_{k \in S_i} d\left([w], [\widehat{V}_k]\right)^2, \tag{2.4}$$

where $d([w],[B]) = 1/\sqrt{2} \cdot \|ww^T - BB^T\|_F$ is the Frobenius-norm distance between subspaces $[w]$ and $[B]$ (Ye and Lim,2016). We chose the Frobenius norm, since for any choice of the basis $\widehat{V}_k$ for $[\widehat{V}_k]$,

$$\sum_{k \in S_i} d\left([w], [\widehat{V}_i]\right)^2 = |S_i| - w^T \left(\sum_{k \in S_i} \widehat{V}_k \widehat{V}_k^T \right) w = |S_i| - w^T (HH^T) w,$$

where H is the matrix given by the column-wise binding of $\widehat{V}_k$'s (Draper et al.,2014), and $\hat{w}$ is the first left singular vector of H. We mention in passing that $[\hat{w}]$ is the one-dimensional flag mean of subspaces $\{[\widehat{V}_k]\}_{k \in S_i}$ (Draper et al.,2014).

Step (b) : Check whether all of the signal score subspaces $[\widehat{V}_k]$ are not too dispersed from $\hat{w}$ in (2.4).

For this, given the prespecified λ, we check whether the *principal angle* between the mean direction and each of $[\widehat{V}_k]$ is at most λ, i.e.,

$$d([\hat{w}], [\widehat{V}_k]) < \sin(\lambda), \quad \text{for all } k \in S_i. \tag{2.5}$$

Here, the principal angle $\theta([w],[B]) := \arcsin(d([w],[B])) \in [0, \pi/2]$ is the acute angle formed by the vector w and the subspace $[B]$ (Björck and Golub,1973). If the condition (2.5) is not satisfied, then skip the following and move to the next stage for S_{i+1}. If (2.5) is satisfied, then $\mathcal{W}_i$ is updated to $\mathcal{W}_i \leftarrow \mathcal{W}_i \cup \{\hat{w}\}$.

Step (c) : Deflate each $[\widehat{V}_k]$ for $k \in S_i$, so that the next mean direction is orthogonal to previous ones.

We "peel" the basis $\hat{w}$, that has been added to $\mathcal{W}_i$, from each of $[\widehat{V}_k]$, $k \in S_i$. Since $\hat{w}$ is not exactly in $[\widehat{V}_k]$, the one-dimensional subspace closest to $\hat{w}$ is removed. Specifically, let $[\widehat{V}_{k,trunc}]$ be the orthogonal complement of $P_{\widehat{V}_k}\hat{w}$ within $[\widehat{V}_k]$. (Note that $\dim([\widehat{V}_{k,trunc}]) = \dim([\widehat{V}_k]) - 1$.) Write $[\widehat{V}_k]$ for $[\widehat{V}_{k,trunc}]$, for each $k \in S_i$. If any $[\widehat{V}_k]$ for $k \in S_i$ becomes $\{0\}$, then move to the next stage for S_{i+1}. Otherwise, move back to step (a).

We give an illustrative example for the three steps.

Example 2.3. *Suppose* $n = 3$ *and* $K = 3$. *Assume that we are at the stage* $S_i = \{1, 2\}$ *and* $[\widehat{V}_1]$ *has been deflated (in the previous stages) to be of dimension 1, while* $[\widehat{V}_2]$ *is of dimension 2, as in Fig. 2.2. There,* $[\widehat{V}_1]$ *is generated by* $(\cos\pi/6, 0, \sin\pi/6)^T$, *and* $[\widehat{V}_2]$ *by* $(1, 0, 0)^T$ *and* $(0, 1, 0)^T$. *In Step (a), the mean direction is* $\hat{w} = (\cos\pi/12, 0, \sin\pi/12)^T$. *In Step (b), for* $k = 1, 2$, $\theta([\hat{w}], [\widehat{V}_k]) = \pi/12$, *(2.5) is satisfied for any tuning parameter* $\lambda > \pi/12$. *Thus we say that* $[\widehat{V}_1]$ *and* $[\widehat{V}_2]$ *share a partially-joint score subspace, and* $\hat{w}$ *is included as a basis of* $[W_i]$. *In step (c),* $[\widehat{V}_2]$ *is updated to* $[\widehat{V}_2] \leftarrow [\widehat{V}_{2,trunc}] = span((0, 1, 0)^T)$, *and* $[\widehat{V}_1]$ *becomes* $\{0\}$. *Since there is no more scores left to exploit, we move onto the next stage for* $S_{i+1} = \{2, 3\}$.

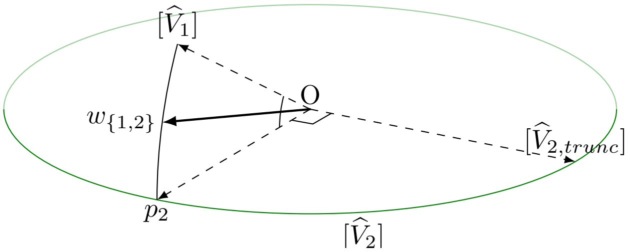


Figure 2.2: A figurative description for computing partially-joint score subspace $[\widehat{W}_i]$, for $S_i = \{1, 2\}$. The two-dimensional subspace $[\widehat{V}_2]$ is depicted as a disk. $w_{\{1,2\}}$ is the mean direction and p_2 stands for the projection of $w_{\{1,2\}}$ onto $[\widehat{V}_2]$.

For a singleton set $S_i = \{k\}$, the mean direction $\hat{w}$ is any unit vector in (the deflated) $[\widehat{V}_k]$, and the condition (2.5) is always satisfied. Thus, for this case, $\widehat{W}_i = \widehat{V}_k$, in place of steps (a)—(c) above.

Note that the estimated partially-joint score matrices $\widehat{W}_i$ consist of orthogonal columns, and $\widehat{W}_i^T \widehat{W}_i = I_{\hat{r}(S_i)}$.

2.3.2 Partially-joint Loading Matrix Estimation

Given the partially-joint structure estimate $\widehat{\mathfrak{S}} = \{(S_i, \hat{r}(S_i)) : i \in \mathcal{I}_K\}$ and the corresponding partially-joint score subspace estimates $[\widehat{W}_i]$'s, we obtain the estimated partially-joint loading matrix $\widehat{U}$.

Let $\widehat{Z} = (\widehat{Z}_1^T, \ldots, \widehat{Z}_K^T)^T \in \mathbb{R}^{p \times n}$, where $p = \sum_{k=1}^K p_k$. Denote the column-wise concatenation of $\widehat{W}_i$ as $\widehat{W} \in \mathbb{R}^{n \times \hat{r}}$, where $\hat{r} = \sum_{i \in \mathcal{I}_K} \hat{r}(S_i)$.

We estimate the partially-joint loading matrix from the optimization problem

$$\widetilde{U} = \mathop{\arg\min}_{U \in \mathbb{R}^{p \times \hat{r}}} \|\widehat{Z} - U \cdot \widehat{W}^T\|_F^2. \tag{2.6}$$

with a constraint that U in (2.6) has the block sparsity structure that corresponds to the estimated partially-joint structure $\widehat{\mathfrak{S}}$. For example, given $K = 2$ and $\widehat{\mathfrak{S}} = \{(\{1,2\},2),(\{1\},1)\}$, U has a block sparsity structure

$$\begin{array}{r} p_1\left\{\vphantom{\begin{array}{c}U\\U\end{array}}\right. \\ p_2\left\{\vphantom{\begin{array}{c}U\\U\end{array}}\right. \end{array} \left[\begin{array}{c:c} U_{(1),1} & U_{(1),2} \\ \hdashline U_{(2),1} & 0 \end{array}\right]$$

with the number of columns of $(U_{(1),1}^T, U_{(2),1}^T)^T$ and $U_{(1),2}$ is two and one.

With the block-wise sparse constraint imposed, the objective function (2.6) for U can be written separately for each data block, i.e.,

$$\|\widehat{Z} - U \cdot \widehat{W}^T\|_F^2 = \sum_{k=1}^{K} \|\widehat{Z}_k - U_{(k)}\widehat{W}_{(k)}^T\|_F^2 \tag{2.7}$$

Here $U_{(k)}$ and $\widehat{W}_{(k)}$ are the column-wise concatenation of each $U_{(k),i}$'s and $\widehat{W}_i$'s with $i \in \{i : k \in S_i \text{ and } \widehat{r}(S_i) > 0\}$. The minimizer of (2.7) is

$$\widehat{U}_{(k)} = \widehat{Z}_k \widehat{W}_{(k)} (\widehat{W}_{(k)}^T \widehat{W}_{(k)})^{-1},$$

and $\widehat{U}_{(k)i}$ for $i \in \mathcal{J}_{(k)}$ are obtained by disjoining $\widehat{U}_{(k)}$. By the constraint imposed, set $\widehat{U}_{(k)i} = 0$ if $k \notin S_i$. The estimated partially-joint loading matrix consists of $\widehat{U}_{(k)i}$, and is denoted by $\widehat{U} \in \mathbb{R}^{p \times \hat{r}}$.

2.3.3 Tuning Parameter Selection

The partially-joint structure identification, proposed in Section 3.1, depends heavily on the tuning parameter $\lambda \in [0, \pi/2)$. If λ is too small, then all scores are identified as individual scores, specific to each data blocks. If λ is too large, then individual and partially-joint scores may be falsely identified as fully-joint scores.

We use data splitting to select the value of tuning parameter $\lambda \in [0, \pi/2)$. For a single instance of data splitting, split n samples of $X = [X_1^T, \ldots, X_K^T]^T$ into two groups of equal proportions, the training set $X_{tr} = [X_{tr,1}^T, \ldots, X_{tr,K}^T]^T$ and the test set $X_{test} = [X_{test,1}^T, \ldots, X_{test,K}^T]^T$.

Given the signal rank r_k of each X_k, we then extract the training signal matrices $\widehat{Z}_{tr,k}$ for $k = 1, \ldots, K$ using the rank r_k approximation of $X_{tr,k}$. For each λ on the tuning parameter grid, we identify the partially-joint structure from $\widehat{Z}_{tr,k}$'s, and obtain the partially-joint score $\widehat{W}_{tr,\lambda}$ and the partially-joint loading matrix $\widehat{U}_{tr,\lambda}$, as discussed in Sections 3.1 and 3.2.

To assess the degrees to which the estimates are generalized to the test set, we first evaluate the score matrix for the test set, given by $\widehat{U}_{tr,\lambda}$. The test score matrix $\widehat{W}_{test,\lambda}$ is defined as the minimizer $\widehat{W}_{test,\lambda} \in \mathbb{R}^{n_{test} \times \hat{r}}$ of

$$\|X_{test} - \widehat{U}_{tr,\lambda} W^T\|_F^2$$

subject to $W^T W = W_{tr,\lambda}^T W_{tr,\lambda}$. The test score matrix is computed as $\widehat{W}_{test,\lambda} = P_\lambda Q_\lambda^T D_\lambda C_\lambda^T$, where P_λ and Q_λ are left and right singular vector matrices of $X_{test}^T \cdot \widehat{U}_{tr,\lambda} = P_\lambda \Sigma_\lambda Q_\lambda^T$, and where D_λ^2 and C_λ are diagonal and right singular vector matrices of

$W_{tr,\lambda}^T W_{tr,\lambda} = C_\lambda D_\lambda^2 C_\lambda^T$, using the Eckart-Young theorem (Eckart and Young,1936).

Let $\tilde{\lambda}$ be the value of λ for which the minimum of the empirical risk is attained. The empirical risk, defined for $\lambda \in [0, \pi/2]$, is

$$\text{Risk}(\lambda) = \sum_{k=1}^{K} \frac{\|X_{test,k} - \widehat{U}_{tr,\lambda,(k)} \widehat{W}_{test,\lambda}^T\|_F^2}{\|X_{test,k}\|_F^2}, \tag{2.8}$$

where $\widehat{U}_{tr,\lambda,(k)}$ is the kth row block of $\widehat{U}_{tr,\lambda}$. A similar form was used in Gaynanova and Li (2019). The corresponding partially-joint structure is $\widehat{\mathfrak{S}}_{tr}(\tilde{\lambda})$.

Finally, we generalize $\widehat{\mathfrak{S}}_{tr}(\tilde{\lambda})$ to the whole data. Let $\widehat{Z}_k$ be the rank r_k approximation of X_k. Again on the grid of λ's, we obtain the partially-joint structure $\mathfrak{S}(\lambda)$ from $\widehat{Z}_k$'s. Then we choose the best value $\widehat{\lambda}$ that minimize diff$(\widehat{\mathfrak{S}}_{tr}(\tilde{\lambda}), \mathfrak{S}(\lambda))$, where d is a measure of dissimilarity between two structure, which will be defined shortly. Indeed, d is a special version of matching distance between multi-sets (Bolt et al.,2022).

Measure of Dissimilarity

In generalizing $\widehat{\mathfrak{S}}_{tr}(\tilde{\lambda})$ to the whole data, we compute dissimilarity measure between $\widehat{\mathfrak{S}}_{tr}(\tilde{\lambda})$ and $\mathfrak{S}(\lambda)$ for each candidate λ. Here we define a measure of dissimilarity using K-row binary structure matrices, for example, when $K = 2$, $((\{1,2\},2),(\{1\},1))$ corresponds to

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Our strategy is that we consider the binary structure matrix as a multi-set of binary column vectors (of size K). We devise a measure that swiftly captures dissimilarity between two multi-sets, under a situation that we need to compute dissimilarity as many time as the number of λ's on grid for each instance of splitting.

Let $d_H(a, b)$ be the number of elements between column vectors $a, b \in \{0, 1\}^K$, i.e. Hamming distance, which we will use as a ground distance. The dissimilarity measure $d(\mathfrak{A}, \mathfrak{B})$ between two partially-joint structure $\mathfrak{A}$, $\mathfrak{B}$ is defined as follows

$$d(\mathfrak{A}, \mathfrak{B}) = \sum_{a \in A \setminus B} d_H(a, B \setminus A) + \sum_{b \in B \setminus A} d_H(b, A \setminus B),$$

where A and B are the binary structure matrices for $\mathfrak{A}$ and $\mathfrak{B}$. The set minus operation, denoted $\setminus$, is defined on a class of binary matrices of K rows, regarded as a multi-set of binary columns. See Section B.1 of the supplementary material for an exemplary description.

Remark 2.2. *It should be noted that* d *is not generally a distance: the conditions (1)* $d(\mathfrak{A}, \mathfrak{B})$ *if and only if* $\mathfrak{A} = \mathfrak{B}$ *and (2)* $d(\mathfrak{A}, \mathfrak{B}) = d(\mathfrak{B}, \mathfrak{A})$ *hold, but (3) the triangular inequality does not.*

2.4 Theory

Given an ordering of index-sets in $(2^{\mathcal{K}} \setminus \{\phi\}, \subset)$ and the signal blocks $Z = \{Z_k : k \in \mathcal{K}\}$, the partially-joint structure $\mathfrak{S}(Z) = \{(S_i, r(S_i), W_i) : i \in \mathcal{I}_K\}$ is uniquely determined. Unfortunately, for different orderings of $(2^{\mathcal{K}} \setminus \{\phi\}, \subset)$, the ranks $r(S_i)$ and the partially-joint score subspaces $[W_i]$ may be different. In this section, we introduce conditions on relations among $[V_k]$'s for $\mathfrak{S}(Z)$ to be uniquely determined regardless of the choice of the orders of $(2^{\mathcal{K}} \setminus \{\phi\}, \subset)$.

For each $l = 1, \ldots, K$, let $\mathcal{J}_l = \{i \in \mathcal{I}_K : |S_i| = l\}$ be the set of all indices with size l. For $l = 1, \ldots, K-1$, we set $[I_l] = +_{i \in \{i:|S_i|>l\}}(\cap_{k \in S_i}[V_k]) = +_{i \in \{i:|S_i|>l\}}[W_i]$, and the projection onto $[I_l]^\perp$ in $\mathbb{R}^n$ is denoted by $P_{I_l}^\perp$. Note that when evaluating the rank for $[W_i]$ for $i \in \mathcal{J}_l$, the definition of $[W_i]$ in (**??**) utilizes the deflated score subspaces orthogonal to $[I_l]$, using the given order of S_i's among $i \in \mathcal{J}_l$. We say $\{[V_k]\}_{k \in \mathcal{K}}$ to be *relatively independent* if, for every $l = 1, \ldots, K-1$ and $i \in \mathcal{J}_l$, $P_{I_l}^\perp(\cap_{k \in S_i}[V_k])$ is linearly independent to $[C_{l,-i}] = +_{j \in \mathcal{J}_l \setminus \{i\}}(P_{I_l}^\perp(\cap_{k \in S_i}[V_k]))$. In words, $\{[V_k]\}_{k \in \mathcal{K}}$ is relatively independent, if for each and every layer $\mathcal{J}_l$, each deflated subspace is lineary independent to $[C_{l,-i}]$, the sum of the other deflated subspaces in the layer $\mathcal{J}_l$. If $P_{I_l}^\perp(\cap_{k \in S_i}[V_k])$ is orthogonal to $[C_{l,-i}]$ for every $l = 1, \ldots, K-1$ and $i \in \mathcal{J}_l$, then $\{[V_k]\}_{k \in \mathcal{K}}$ is said to be *relatively orthogonal.* We immediately check that relative orthogonality implies relative independence.

Theorem 2.1. *Given matched data matrices* $X_k = Z_k + E_k \in \mathbb{R}^{p_k \times n}$ *for* $k = 1, \ldots, K$ *with true signal* Z_k *and error* E_k*, if* $\{[V_k]\}_{k \in \mathcal{K}}$*, the collection of* Z_k*'s signal score subspaces, is relatively independent, then, regardless of the ordering of index-sets in* $(2^{\mathcal{K}} \setminus \{\phi\}, \subset)$*, there exists a unique set of pairs* $\{(S_i, r(S_i))\}_{i \in \mathcal{I}_K}$.

Under only the relative independence condition, the determination of the partially-joint score subspaces $[W_i]$ corresponding to S_i, $i \in \mathcal{I}_K$, may not be unique, and depends on the ordering of index-sets (*see* Examples A.3 and A.4 in the supplementary materials). To ensure uniqueness of $[W_i]$'s, we require a rather strong assumption. We say that $\{[V_k]\}_{k \in \mathcal{K}}$ is *absolutely orthogonal,* if (1) $\{[V_k]\}_{k \in \mathcal{K}}$ satisfies relative orthogonality, and (2) for each $l = 1, \dots, K-1$ and for every $i \in \mathcal{J}_l$,

$$P_{I_l}^{\perp}(\cap_{k \in S_i}[V_k]) = P_{J_i}^{\perp}(\cap_{k \in S_i}[V_k]), \tag{2.9}$$

where $[J_i] = +_{j \in \mathcal{J}_{i,>l}}(\cap_{k \in S_j}[V_k]) = +_{j \in \mathcal{J}_{i,>l}}[W_j]$ and $\mathcal{J}_{i,>l} = \{j : |S_j| > l, S_i \cap S_k \neq \phi\}$. In Example 2.1, when $l = 1$ and $i = 6$, (2.9) holds if $(\bigcirc_{j \in \{1,2,3,4\}} P_j^{\perp})([V_1]) = (\bigcirc_{j \in \{1,2,3\}} P_j^{\perp})([V_2])$. Note that for $i = 6$, $S_i = \{2\}$ and the index-sets S_j's for $j \in \mathcal{J}_{i,>l}$ are $S_1 = \{1,2,3\}$, $S_2 = \{1,2\}$, $S_3 = \{2,3\}$, excluding $S_4 = \{1,3\}$.

Theorem 2.2. *Given matched data matrices* $X_k = Z_k + E_k \in \mathbb{R}^{p_k \times n}$ *for* $k = 1, \dots, K$ *with true signal* Z_k *and error* E_k, *if* $\{[V_k]\}_{k \in \mathcal{K}}$ *is absolutely orthogonal, then partially-joint score subspaces* $[W_i]$ *for* $i \in \mathcal{I}_K$ *are uniquely determined.*

The uniqueness of each partially-joint loading subspace $[U_{(k),i}]$ is deduced from the uniqueness of $[W_i]$'s.

Corollary 2.1. *Given matched data matrices* $X_k = Z_k + E_k \in \mathbb{R}^{p_k \times n}$ *for* $k = 1, \dots, K$ *with true signal* Z_k *and error* E_k, *if* $\{[V_k]\}_{k \in \mathcal{K}}$ *is absolutely orthogonal, then each partially-joint loading subspace* $[U_{(k),i}]$ *is uniquely determined for* $k = 1, \dots, K$ *and* $i \in \mathcal{I}_{(k)}$.

We provide proofs of theorems in this section and examples for relative independence and absolute orthogonality in Section A.2 of the supplementary materials.

2.5 Simulation Study

2.5.1 Example Dataset Generation

In the simulation study, we use the following data generation setting for numerically analyzing the performance of our proposal. Throughout, we use $K = 3$ blocks of data sets, in which the association structures are given by the ranks of index-sets.

First, we set a pre-determined rank $r(S_i)$ for each index-set S_i for $i = 1, \ldots, 2^K - 1$. The generic partially-joint score matrix $W_{comp,i} \in \mathbb{R}^{n \times r(S_i)}$ for each index-set is a column-wise concatenation of randomly generated vectors $w_{comp,i,j}$ for $j = 1, \ldots, r(S_i)$. Each $w_{comp,i,j}$ are generated element-wise and the entries of $w_{comp,i,j}$ follows $\mathcal{N}(0, \sigma^2_{i,j})$ independently; not only independent within a $w_{comp,i,j}$, but independent between $w_{comp,i,j}$'s as well. The magnitude of signal $\sigma^2_{i,j}$'s depends on the simulation settings, and summarized as $\sigma^2_M = \{(\sigma^2_{i,j}) : i \in \mathcal{I}_K, j = 1, \ldots, r(S_i), r(S_i) > 0\}$. The column-wise concatenation of $W_{comp,i}$ is denoted W_{comp} of size $n \times \sum_{i=1}^{2^K-1} r(S_i)$.

The generic loading matrices $U_{comp,k,i} \in \mathbb{R}^{p_k \times r(S_i)}$ for $i = 1, \ldots, 2^K - 1$ and $k = 1, \ldots, K$ are given as follows. The entries of $U_{comp,k,i}$ are generated independently from the uniform distribution, Unif$(0, 1)$, and all the columns of $U_{comp,k,i}$ are scaled. We give $U_{comp,k,i}$ orthonormality by the QR decomposition. We then derive the generic signal matrix Z_k for $k = 1 \ldots, K$ by $Z_k = \text{weight}_k \cdot \sum_{i=1}^{2^K-1} (U_{comp,k,i} \cdot W_{comp,i})$, where weight$_k$'s are weights for each dataset, which reflect the magnitude of signals. The concatenation of generic loading matrices is denoted U_{comp}, and is of size $\sum_{k=1}^{K} p_k \times \sum_{i=1}^{2^K-1} r(S_i)$.

The error matrix E_k is generated element-wisely, such that $e_{k,ij} \sim \mathcal{N}(0, \sigma^2)$ independently for $i = 1, \dots, p_k$ and $j = 1, \dots, n$. The magnitude of error σ^2 is set as the reciprocal of the signal-to-noise ratio (SNR), $\sigma^2 = 1/\text{SNR}$, where SNR is predetermined as a simulation setting.

We use the following six models. We set $n = 200$, $p_1 = p_2 = p_3 = 100$ and $\text{weight}_1 = \text{weight}_2 = \text{weight}_3 = 1$, for all six cases.

1. (Individuals) $\mathfrak{S} = \{(\{1\}, 2), (\{2\}, 2), (\{3\}, 2)\}$, $\sigma_M^2 = \{(1.4, 0.8), (1.3, 0.7), (1.2, 0.6)\}$
2. (Fully joint) $\mathfrak{S} = \{(\{1, 2, 3\}, 2)\}$, $\sigma_M^2 = \{(1.0, 0.9)\}$
3. (Circular, partially joint) $\mathfrak{S} = \{(\{1, 2\}, 2), (\{1, 3\}, 2), (\{2, 3\}, 2)\}$, $\sigma_M^2 = \{(1.4, 0.8), (1.3, 0.7), (1.2, 0.6)\}$
4. (Mix of fully joint and individuals) $\mathfrak{S} = \{(\{1, 2, 3\}, 2), (\{1\}, 2), (\{2\}, 2), (\{3\}, 2)\}$, $\sigma_M^2 = \{(1.5, 0.8), (1.4, 0.7), (1.3, 0.6), (1.2, 0.5)\}$
5. (Fully joint and partially joint) $\mathfrak{S} = (\{1, 2, 3\}, 2), (\{1, 2\}, 2), (\{1, 3\}, 2), (\{2, 3\}, 2)\}$, $\sigma_M^2 = \{(1.5, 0.8), (1.4, 0.7), (1.3, 0.6), (1.2, 0.5)\}$
6. (All possible combinations) $\mathfrak{S} = (\{1, 2, 3\}, 2), (\{1, 2\}, 2), (\{1, 3\}, 2), (\{2, 3\}, 2), (\{1\}, 2), (\{2\}, 2), (\{3\}, 2)\}$, $\sigma_M^2 = \{(1.8, 0.8), (1.7, 0.7), (1.6, 0.6), (1.5, 0.5), (1.4, 0.4), (1.3, 0.3), (1.2, 0.2)\}$

2.5.2 Results on Comparative Study

In this subsection, we numerically compare the performance of our proposal to other competitors, including SLIDE (Gaynanova and Li,2019), COBS (Gao et al.,2020), AJIVE (Feng et al.,2018), and JIVE (Lock et al.,2013). Note that from the estimates of each method, the partially-joint structure $\mathfrak{S}$ can be extracted, as well as the concatenated partially-joint score matrix $\widehat{W}$ and loading matrix $\widehat{U}$. See Section A.3.2 of the supplementary materials for a brief review of these methods. Our proposal will be called the method of partially-joint structure identification, or PSI for short.

To assess the efficacy of finding the true partially-joint structure $\widehat{\mathfrak{S}}$ and proper loading and score matrices, we use the following measures.

(1) *Partially-joint structure* $\widehat{\mathfrak{S}}$: The rate of finding the true partially-joint structure, $\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}}_0)$.

(2) *Partially-joint loading matrix* $\widehat{U}$: We find the difference between U_{comp} and $\widehat{U}$ as follows. We denote the principal angles between $U_{comp,k,i}$ and $\widehat{U}_{(k.)}$ by $\theta_{U,k,i,j}$ for $k = 1, \ldots, K$, $i = 1, \ldots, 2^K - 1$ and $j = 1, \ldots, r(S_i)$. We report the average of all the values of $\theta_{U,k,i,j}$'s as $\overline{\theta}(U, \widehat{U})$.

(3) *Partially-joint score matrix* $\widehat{W}$: We find the difference between W_{comp} and $\widehat{W}$ as follows. We denote the principal angle between $w_{comp,i,j}$ and $\widehat{W}$ by $\theta_{W,i,j}$ for $i = 1, \ldots, 2^K - 1$ and $j = 1, \ldots, r(S_i)$. We report the average of all the values of $\theta_{W,i,j}$'s as $\overline{\theta}(W, \widehat{W})$.

In the comparative study, the measures $\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$, $\overline{\theta}(U, \widehat{U})$ and $\overline{\theta}(W, \widehat{W})$ were computed to assess the performance of PSI and other four methods. The simulation was conducted on different values of SNR (10 and 5) for the example models 1 to 6. Given a fixed true partially-joint loading matrix U_{comp}, one hundred data

Table 2.1: Comparative Study on Model 1 to 3. The unit for $\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$ is percent. Numbers in parantheses are standard deviations.

Model	SNR	Measure	PSI	SLIDE bcv	COBS	AJIVE	JIVE
1	10	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	1	18	93	5
		$\bar{\theta}(U, \widehat{U})$	13.69 (0.48)	25.26 (14.08)	52.59 (9.43)	13.72 (0.52)	**13.62 (0.49)**
		$\bar{\theta}(W, \widehat{W})$	**18.47 (0.58)**	19.55 (0.79)	20.85 (1.39)	18.55 (0.95)	18.62 (0.64)
	5	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	3	48	93	6
		$\bar{\theta}(U, \widehat{U})$	20.01 (0.76)	36.26 (14.96)	67.24 (6.13)	20.14 (0.9)	**19.94 (0.77)**
		$\bar{\theta}(W, \widehat{W})$	**25.96 (0.71)**	28.14 (2.19)	34.48 (3.41)	26.33 (1.8)	26.1 (0.75)
2	10	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	**100**	**100**	**100**	0
		$\bar{\theta}(U, \widehat{U})$	**12.99 (0.59)**	13.03 (0.59)	19.27 (1.82)	**12.99 (0.59)**	13.33 (0.63)
		$\bar{\theta}(W, \widehat{W})$	10.86 (0.55)	11.47 (0.77)	11.17 (0.62)	10.86 (0.55)	**10.67 (0.53)**
	5	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	**100**	**100**	**100**	0
		$\bar{\theta}(U, \widehat{U})$	**18.43 (0.88)**	18.46 (0.88)	32.35 (2.8)	**18.43 (0.88)**	19.47 (1)
		$\bar{\theta}(W, \widehat{W})$	15.66 (0.85)	16 (0.83)	17.14 (1.12)	15.66 (0.85)	**15.38 (0.83)**
3	10	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	22	0	0	0
		$\bar{\theta}(U, \widehat{U})$	**13.27 (0.41)**	24.47 (8.16)	24.22 (2.82)	13.68 (0.57)	13.32 (0.44)
		$\bar{\theta}(W, \widehat{W})$	13.52 (0.39)	14.05 (0.48)	14.66 (0.54)	13.73 (0.53)	**13.12 (0.42)**
	5	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**89**	56	1	0	0
		$\bar{\theta}(U, \widehat{U})$	**19.13 (0.71)**	34.52 (7.09)	45.44 (5.54)	20.39 (0.84)	19.61 (0.75)
		$\bar{\theta}(W, \widehat{W})$	19.52 (0.62)	19.75 (0.66)	26.92 (2.77)	20.93 (1.15)	**18.84 (0.61)**

sets were generated for each SNR value. We tested all 5 methods over these datasets.

We give each method the initial ranks as follows : (1) PSI, AJIVE, JIVE : Truncate each X_k with its true rank. For example, in Model 3, each X_i was truncated with the true rank 4, (2) SLIDE : Use the whole X, the row-concatenation of X_i's. After standardization, X is then truncated with the sum of each index-set's true rank (for example, $6 = 2 + 2 + 2$ in Model 3), and (3) COBS : Use the whole X, the row-concatenation of X_i's. Set the number of components as the sum of each index-set's true rank.

Average and standard deviation for each measure are reported in Tables 2.1 and 2.2.

In the example model 1, in which case there are only individual

Table 2.2: Comparative Study on Model 4 to 6.

Model	SNR	Measure	PSI	SLIDE bcv	COBS	AJIVE	JIVE
4	10	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	2	24	**100**	0
		$\bar{\theta}(U, \widehat{U})$	13.31 (0.36)	20.03 (6.06)	34.19 (4.37)	**13.3 (0.36)**	13.45 (0.37)
		$\bar{\theta}(W, \widehat{W})$	17.28 (0.43)	18.07 (0.47)	19.69 (1.4)	**17.1 (0.42)**	17.36 (0.43)
	5	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	8	24	**100**	0
		$\bar{\theta}(U, \widehat{U})$	19.23 (0.57)	28.35 (8.82)	47.6 (2.85)	**19.19 (0.57)**	19.63 (0.61)
		$\bar{\theta}(W, \widehat{W})$	24.37 (0.64)	27.65 (3.66)	33.98 (2.65)	24.1 (0.63)	**24.04 (0.66)**
5	10	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	35	0	0	0
		$\bar{\theta}(U, \widehat{U})$	**13.28 (0.29)**	22.01 (6.24)	24.2 (2.2)	13.49 (0.38)	13.42 (0.33)
		$\bar{\theta}(W, \widehat{W})$	13.28 (0.32)	13.68 (0.41)	14.51 (0.58)	13.4 (0.43)	**12.9 (0.49)**
	5	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**69**	58	0	0	0
		$\bar{\theta}(U, \widehat{U})$	**19.19 (0.59)**	29.23 (4.95)	41.43 (3.36)	19.79 (0.53)	19.73 (0.53)
		$\bar{\theta}(W, \widehat{W})$	19.14 (0.6)	19.22 (0.62)	26.37 (1.89)	19.86 (0.72)	**18.33 (0.64)**
6	10	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	99	0	0	0	0
		$\bar{\theta}(U, \widehat{U})$	14.28 (0.33)	29.54 (9.7)	30.71 (2.27)	**14.12 (0.32)**	14.31 (0.34)
		$\bar{\theta}(W, \widehat{W})$	17.88 (0.45)	20.9 (2.29)	24.53 (1.21)	17.66 (0.45)	**17.54 (0.49)**
	5	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**55**	0	0	0	0
		$\bar{\theta}(U, \widehat{U})$	21.27 (0.69)	29.8 (9.71)	44.91 (2.77)	**21.24 (0.73)**	21.44 (0.67)
		$\bar{\theta}(W, \widehat{W})$	25.71 (0.84)	28.45 (1.68)	34.2 (1.09)	25.95 (1.07)	**24.52 (0.84)**

scores, PSI and AJIVE identified the true partially-joint structure for almost all instances. In the example model 2, in which case there are only joint score, all methods but JIVE find the true structure for almost all SNRs and instances. PSI, SLIDE and AJIVE have estimated identical score subspaces. In the example model 3, in which partially-joint scores are entangled in a cyclic structure, PSI boasts superior performances in identifying the true structures. Table 2.1 confirms these observations.

In the example model 4, in which case both joint and individual scores are composited, both PSI and AJIVE showed superior performance in identifying the true structure. Our method was as competent in estimating score subspaces as AJIVE. In the example model 5 and 6, the complicated cases with joint, partially-joint and individual scores mixed, PSI is prominent in estimating true structure, loading and score subspaces. See Table 2.2.

2.5.3 Results on Imbalanced Signal Strength between Joint and Individual Components

Next, we consider cases where the signal strengths of joint components and individual components are grossly imbalanced. Consider a new model, with $K = 3$ and $p_1 = p_2 = p_3 = 100$, whose index-set ordering is $S_1 = \{1, 2, 3\}$, $S_5 = \{1\}$, $S_6 = \{2\}$ and $S_7 = \{3\}$. We set inherent joint rank $r(S_1) = 10$ and individual ranks $r(S_5) = r(S_6) = r(S_7) = 10$. Other index-sets have zero ranks, $r(S_2) = r(S_3) = r(S_4) = 0$. Throughout, $n = 200$.

For the case in which there are larger variations in the joint component than the individual components, we set $\sigma^2_{1,j} \gg \sigma^2_{i,j}$ for $i = 5, 6, 7$ and all j's. In the opposite case, we give larger variations in the individual components than the joint component, $\sigma^2_{i,j} \gg \sigma^2_{1,j}$ for $i = 5, 6, 7$ and all j's. Details are given in Section A.3.5 of the supplementary materials.

We carried out comparative simulations on both cases at SNR levels ∞ and 5. Not only the success rate of finding true structure but also the numbers of estimated joint and individual components were evaluated. In Table 2.3, in which case the joint component has larger variations, PSI, AJIVE and JIVE showed superior performances to other methods. SLIDE only succeeded in finding joint components even when there is no noise. When the individual components have larger variations, in Table 2.4, our method still shows better performances in finding both joint and individual components when there is no noise, but failed to estimate any joint structure with noise. At SNR 5, AJIVE and JIVE was the most competent among all methods.

Table 2.3: The case where the joint component has larger variations than the individual components.

SNR	Measure	PSI	SLIDE bcv	COBS	AJIVE	JIVE
∞	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	0	0	0	**100**
	Joint	**10 (0)**	**10 (0)**	39.58 (0.59)	3.79 (1.95)	**10 (0)**
	Individual	**30 (0)**	0 (0)	0.02 (0.14)	48.63 (5.84)	**30 (0)**
5	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	0	0	1	0
	Joint	**10 (0)**	**10 (0)**	12.49 (1.76)	**10 (0)**	**10 (0)**
	Individual	**30 (0)**	0 (0)	18.54 (2.61)	33.19 (1.01)	43.29 (1.53)

Table 2.4: The case where the individual components have larger variations than the joint components.

SNR	Measure	PSI	SLIDE bcv	COBS	AJIVE	JIVE
∞	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	**100**	0	0	0	0
	Joint	**10 (0)**	9.84 (0.77)	38.33 (1.46)	3.67 (1.74)	15.05 (1.7)
	Individual	**30 (0)**	25.88 (1.29)	0.14 (0.35)	48.99 (5.21)	27.35 (1.2)
5	$\mathbb{E}\mathbb{1}(\mathfrak{S} = \widehat{\mathfrak{S}})$	0	0	0	0	**1**
	Joint	0 (0)	**10.18 (1.42)**	15.28 (3.04)	8.27 (0.58)	11.38 (2.94)
	Individual	59.4 (1.41)	21.48 (1.77)	12.33 (2.31)	36.41 (1.88)	**27.27 (1.14)**

2.6 Real Data Analysis

In this section, we apply the proposed PSI to a dataset called EGAS0000100174, a blood cancer multi-omics data set linked to a drug response panel (Dietrich et al.,2018). We have chosen to use 121 cases diagnosed with chronic lymphocytic leukemia (CLL). The drug response panel (X_{Drug}) records the *ex vivo* cell viabilities at a series of 5 concentrations, for each of 62 drugs that target onco-related pathways or are used widely in clinical practice. This multi-omics data set consists of the genome-wide DNA methylation profiles and the RNA sequencing profiles. The top 5,000 most variable CpG sites were selected (X_{Meth}) from the *450K il-*

lumina assay DNA methylation profile. As for the RNA sequencing profile, we selected the top 5,000 gene expressions with the largest stabilized variances (X_{Exp}) from the high-throughput sequencing (HTS) assay. In summary, we have three blocks of data sets $X_{\text{Drug}} \in \mathbb{R}^{121\times310}$, $X_{\text{Meth}} \in \mathbb{R}^{121\times5000}$ and $X_{\text{Exp}} \in \mathbb{R}^{121\times5000}$.

We repeat the estimation process over 100 repetitions (of data splitting) and select the mode structure, that is, the estimated partially-joint structure that appears the most out of 100 repetitions. The estimated mode structure is

$$\widehat{\mathfrak{S}} = \{(\{\text{Drug}, \text{Meth}\}, 1), (\{\text{Drug}\}, 4), (\{\text{Meth}\}, 41), (\{\text{Exp}\}, 3)\},$$

which means that PSI detected the index-set $\{\text{Drug}, \text{Meth}\}$ of rank 1, and no fully-joint score or other forms of partially-joint scores were detected. This structure appeared 53 times out of 100 repetitions. The index-set $\{\text{Drug}, \text{Meth}\}$ of rank 1 stands for the existence of one-dimensional latent score, partially-joint for Drug and Meth data sets (but not for Exp data set).

PSI showed better performance in computation time over other methods. In an Intel ® Xeon ® CPU E5-2640 v4 @ 2.40GHz system, it took 9.38 seconds per single data splitting instance (54.56 seconds for 100 instances with 40 cores). In comparison, the SLIDE with bcr method detected the same structure as our method, but it took almost 9 hours on the same machine. The COBS yielded an eccentric result, giving the fully-joint score of rank 50, taking about 14 minutes.

PSI was also robust over different choices of initial ranks. As a preprocessing, we set the ranks of the signal matrices using the principal component analysis which accounts for cumulative proportion of variances at 30%, 40%, 50% and 60%. Resulting signal ranks for Drug, Meth and Exp datasets are $(2, 6, 1)$, $(2, 15, 1)$, $(3, 28, 2)$ and $(5, 42, 3)$, respectively. For all cases, PSI detected the same index-set $\{\text{Drug}, \text{Meth}\}$ of rank 1.

The analysis by PSI for the patient sample multi-block data reveals a particular pattern in the latent partially-joint score, which cannot be identified by applying, *e.g.* principal component analysis to each data block. To support this claim, we plot the reconstructed matrices $\widehat{U}_{(k),i}\widehat{W}_i^T$ for the identified partially-joint and individual parts of the data in Fig. 2.3. Both the samples and the variables of X_{Drug} are ordered by a hierarchical bi-clustering applied to the partially-joint component $\widehat{U}_{(\text{Drug}),2}\widehat{W}_2^T$ of X_{Drug}, where $\widehat{W}_2 = \widehat{W}_{\{\text{Drug},\text{Meth}\}}$. The matrix $Z_{\text{Drug},S_2} := \widehat{U}_{(\text{Drug}),2}\widehat{W}_2^T$ is shown in the top left part of Fig. 2.3. The variables of X_{Meth} are similarly ordered.

Focusing on the partially-joint scores corresponding to $\{\text{Drug}, \text{Meth}\}$, the samples are clustered into two distinct subgroups. These subgroups are shown in Fig. 2.4(a), and are denoted by groups α and β. There, it can be seen that the variables in X_{Drug} and X_{Meth} show a contrasting pattern according to the two subgroups α and β. Comparing Z_{Drug,S_2} (Fig. 2.4(a)) with the whole X_{Drug} (Fig. 2.4(b)), we observe that the subgroups identified above are hidden in X_{Drug}. Moreover, the subgroups α and β are well-separated in the partially-joint score $\widehat{W}_{\{\text{Drug},\text{Meth}\}}$ of Z_{Drug,S_2} (Fig. 2.4(c)), while it is hard to find any subgroup in the principal component scores of whole X_{Drug} (Fig. 2.4(d)). The same conclusion can be made by inspecting the component of X_{Meth} corresponding to the partially-joint score $\widehat{W}_{\{\text{Drug},\text{Meth}\}}$, denoted Z_{Meth,S_2}, and the whole X_{meth} in Fig. 2.4(e) and (f). Thus we observed that PSI gives a more effective measure of finding inherent subgroups in a multi-omics data set than a separate application of the principal component analysis on each of the data matrix.

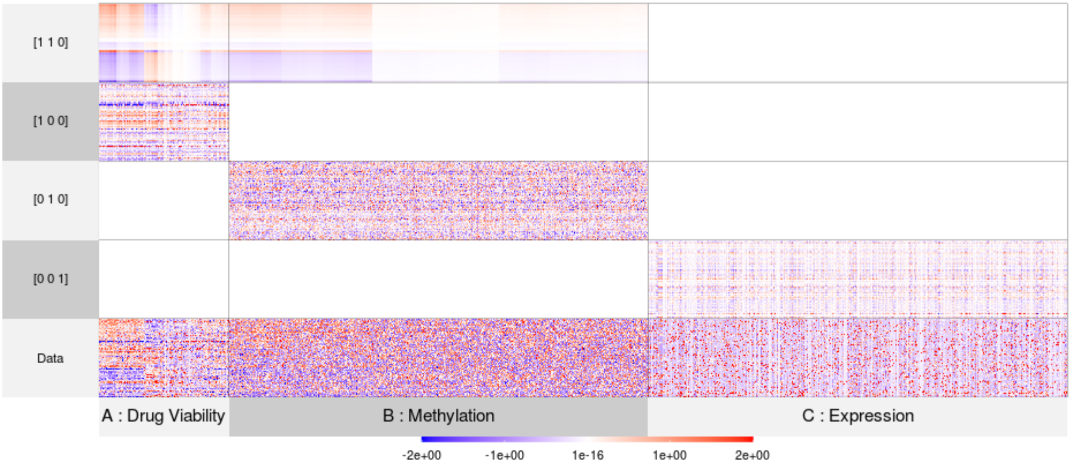
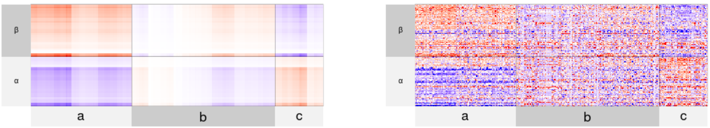


Figure 2.3: The reconstructed components according to the estimated partially-joint structure. The values are all normalized for each column, then truncated to line on $[-2, 2]$.

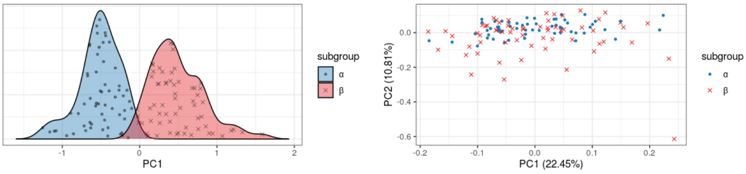
To find indicators that best explain the subgroups α and β, we conducted the Fisher exact test simultaneously on 59 gene mutations or chromosome defects of each patients, available as an ancillary information. The p-values from each Fisher exact test were adjusted by the Benjamini-Hochberg (BH) method (Benjamini and Hochberg,1995). The smaller p-values indicate stronger differences of each mutation or chromosome defect between the subgroup α and β. We found that the immunoglobulin heavy chain variable (IGHV) region mutation status has the most associated relation with the subgroups at BH-adjusted p-value 1.026×10^{-13}. We also present the 2×2 table of the IGHV status and the subgroups, *see* Tables A.4.1 and A.4.2 in the supplementary materials. We postulate IGHV mutation status gives a substantial explanation for the subgrouping of CLL patients, that is, wild type matches to the subgroup α and mutation type to the subgroup β. Survival analyses on overall survival rates was also conducted, and we found statistically significant differences in survival between the two subgroups α and β as shown in Fig. 2.5.

Again in Fig. 2.4(a), the variables in Z_{Drug,S_2} can be clustered into subgroups [a] and [b] showing a contrasting response pattern to the subgroups α and β (variables showing weak responses were excluded as subgroup [c]). The subgroup [a] shows higher viability for β (IGHV mutated) than α (IGHV wild type) and vice versa for [b]. Table A.4.3 (in the supplementary materials) presents the list of prominent drugs that have appeared in subgroups [a] and [b] at least 4 times out of 5 concentrations. For the subgroup [a], the table lists a number of inhibitor drugs that target the B cell receptor (BCR) components, such as Bruton's tyrosine kinase (BTK; spebrutinib, ibrutinib), phosphatidylinositol 3-kinase (PI3K; idelalisib, duvelisib) and spleen tyrosine kinase (STK; tamatinib, PRT062607 HCL). AKT inhibitor(MK-2206) or SRC inhibitors (dasatinib) targets signal transduction pathways that promotes survival and growth of B cell lymphocytes. Unexpected encounter with HSP90 inhibitor (AT13387 or Onalespib) may be related to the stability of lymphocyte-specific SRC family kinases (Mshaik et al.,2021). The appearance of CHK inhibitors (PF 477736, AZD7762, CCT241533) may be linked to repairing mechanisms of DNA damages at G2 phase, known to be associated with WEE1 kinase and the AKT/PKB pathway (Zhang and Hunter,2014). For the subgroup [b], the appearance of mTOR inhibitor (everolimus) may suggest that mTOR pathway and shows different drug sensitivities to the BCR component, despite the fact that it is on the downstream of AKR/PKB pathway. The role of IGHV in this implication requires further investigation. BCL2 inhibitor (navitoclax) and rotenone might be related to the role of mitochondria in apoptosis (Wang and Youle,2014).

(a) Z_{Drug,S_2} (b) the whole X_{Drug}



(c) Z_{Drug,S_2} (d) the whole X_{Drug}



(e) Z_{Meth,S_2} (f) the whole X_{Meth}

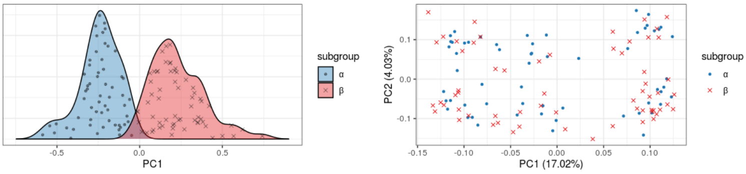


Figure 2.4: (a) The matrix $Z_{\text{Drug},S_2} := \widehat{U}_{(\text{Drug}),2}\widehat{W}_2^T$. (b) The whole X_{Drug} data matrix. (c) The density plot of the subgroups α and β along the first principal component (PC) score of Z_{Drug,S_2}; (d) The PC scores plot for the whole X_{Drug}; (e) The density plot of the subgroups α and β along the first principal component of Z_{Meth,S_2}; (f) The PC scores plot of the whole X_{Meth}.

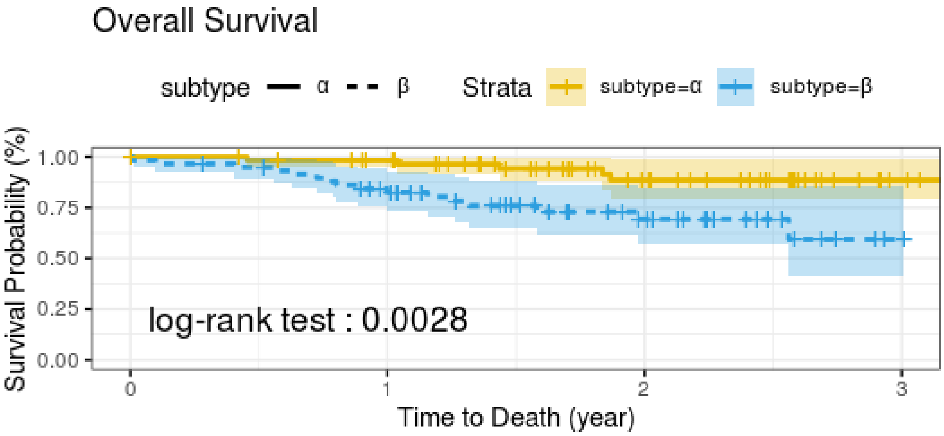


Figure 2.5: The difference in overall survival between the subgroups α and β is displayed on a Kaplan plot with p-value 0.0028 from the log-rank test.

Chapter 3

Penalized Principal Nested Spheres

3.1 Introduction

In this section, we review forward nestedness of PCA, the non-Euclidean generalization of PCA to PGA and backward nestedness of PNS. We then present the concept of overfitting phenomenon in PNS and give the motivations of PenPNS.

Nearly for a century, Principal Component Analysis (PCA) (Hotelling, 1933) has been a popular technique in multivariate data analysis. PCA is a dimension reduction method, in that it linearly transforms data points into a lower dimensional linear space with a new coordinate system. More specifically, PCA finds a coordinate, called the first principal component, on which the data points are projected so that the projected points have the largest variance. PCA next finds another coordinate, the second principal components, on which the data points are projected again and have the largest variance in the complementary space of the first principal component. The third principal components are found in the same manner and so on.

The mechanism described above on which PCA is operating is summarized as the concept *forward nestedness*. PCA is a stepwise procedure. At the current $m+1$th step, where m principal components are already identified, $m+1$th principal component is the best descriptor of the data points projected into the complement of the space that is spanned by the first to m-th principal component. By the nestedness, we describe that the subspace spanned by the principal components is a nested structure, that is,

$$\mathrm{span}(\mathrm{PC1}) \subset \mathrm{span}(\mathrm{PC1}, \mathrm{PC2}) \subset \mathrm{span}(\mathrm{PC1}, \mathrm{PC2}, \mathrm{PC3}) \subset \dots,$$

where PCm denotes the mth principal component. By the forward, we describe that the nested structure is built on from one dimension to larger step by step.

Our aim is to generalize PCA to a non-Euclidean spaces. There has been several attempts to apply PCA method on manifold surface (Gower,1975; Huckemann and Ziezold,2006; Huckemann et al.,2010; Mardia et al.,2021) and one of the elementary approach is Principal Geodesic Analysis (PGA) (Fletcher et al., 2004). PGA adopts forward nestedness, but with approximating the manifold surface by Euclidean space. The overall generalization framework is as follows.

(1) Given a base point μ on the manifold M, the data points $x_i \in M$ on the manifold surface are mapped to the points u_i's

of the tangent space $T_\mu M$ (approximation of the manifold surface by the tangent space, which is Euclidean space by definition)

(2) Then in the tangent space $T_\mu M$, find the descriptors v_k's that best explains the points u_i's in a manner of forward nestedness.

Specifically, PGA finds the base point μ as the intrinsic mean of x_i's, that is, a minimizer point of expected squared intrinsic distance (Kobayashi and Nomizu, 1969; Karcher, 1977). The data points $x_i \in M$ are mapped to u_i by an inverse exponential map, Exp_μ^{-1}. In the tangent space $T_\mu M$, PGA performs PCA on points u_i's. Each descriptor $v_k \subset T_\mu M$ is mapped back to a geodesic through μ on M, $\mathrm{Exp}_\mu(v_k)$, by an exponential map. Thus, we have geodesic descriptors $\mathrm{Exp}_\mu(v_k)$ of data points $x_i \in M$.

The major pitfall of this framework is that its result depends largely on the choice of the base point μ. The distribution of points u_i's in the tangent space $T_\mu M$ varies greatly on the choice of the base point and can lead to way different results in finding descriptos v_k. Also, PGA's choice of intrinsic mean is somewhat arbitrary. Moreover, there is a counter-example where the geodesic descriptors $\mathrm{Exp}_\mu(v_k)$ do not explain the data points x_i's well.

Example 3.1. *Let $M = S^2$, a sphere in $\mathbb{R}^3$. Suppose the data points x_i's are distributed uniformly on the equator. It can be easily shown that both north and south poles are the intrinsic mean μ of x_i's. Then the geodesic descriptors $Exp_\mu(v_k)$ are longitude lines passing through both north and south poles, but intuitively they are not good descriptors for data points on the equator.*

To cope with these difficulties, Analysis of Principal Nested Spheres (PNS) (Jung, 2012) takes a different strategy, *backward nestedness.* PNS is an analogue of PCA for data points on hypersphere, S^d. Instead of building up a nested structure from dimension one, PNS starts from S^d and reduces dimension one by one. The PNS framework, at the mth step, is

(1) Finds a best-fitting subsphere A_{d-m} of dimension $d-m$ in S^{d-m+1}.

(2) Project the points $x_i \in S^{d-m+1}$ onto A_{d-m}.

(3) As the subsphere A_{d-m} is homeomorphic to S^{d-m}, map the points of A_{d-m} to S^{d-m} using an isomorphism f_{d-m}.

As a result, PNS yields the following nested structure

$$S^d \supset \mathfrak{A}_{d-1} \cong S^{d-1} \supset \mathfrak{A}_{d-2} \cong S^{d-2} \supset \ldots \supset \mathfrak{A}_0 \cong S^0,$$

where $\mathfrak{A}_{d-1} = A_{d-1}$ and $\mathfrak{A}_{d-m} = f_{d-1}^{-1} \circ \cdots \circ f_{d_m}^{-1}(A_{d-m})$ for $m = 2, \ldots, d$. By adopting backward nestedness, there is no more need to choose a base point and mapping data points to the tangent space. Instead, we only need to estimate the parameters that describe the subsphere. A subsphere is defined by the axis $v \in S^m$ and radius $r \in [0, \pi/2]$ as

$$A_{m-1}(v, r) = \{x \in S^m : \arccos(v^T x) = r\}.$$

For more details about the framework of PNS, *see* Section 3.2.

One of the major characteristics of PNS is that the subsphere is not necessarily a great sphere. Here a great sphere means a subsphere with radius $r = \pi/2$ and it is a high-dimensional analogue

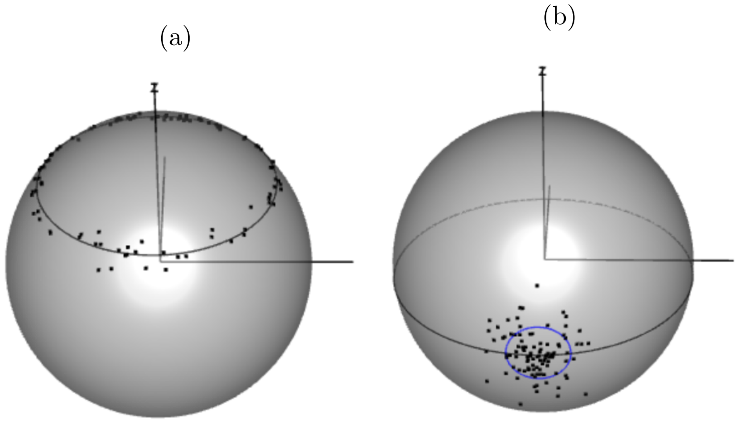


Figure 3.1: (a) Data points generated along a small sphere with $r = \pi/4$. (b) Data points generated along a great sphere (the equator), but inside a short interval of 10° (perturbed by von Mises-Fisher distribution with $\kappa = 50$). Blue circle is the fitted small sphere by PNS.

of a geodesic circle of S^2. Permitting a small sphere ($r < \pi/2$) gives PNS more flexibility in fitting a certain type of non-geodesic variation on a hypersphere. For example, if data points are distributed along a circle which is posited on latitude 45° of S^2, it is reasonable to fit the data points with a small sphere with parameters $v = (0, 0, 1)^T$ and $r = \pi/4$, *see* Figure 3.1a.

However, at the same time, permitting a small sphere can induce a somewhat pathological phenomenon, called *overfitting phenomenon*. Overfitting phenomenon in PNS indicates a situation where the fitted radius is less than $\pi/2$ even if data points are generated along a great subsphere. In this chapter, we discuss two types of overfitting phenomenon in PNS.

Type 1 Overfitting Phenomenon : When data points are generated along a great sphere, the angle between the true axis and data point would make a distribution around $\pi/2$. However, when the estimated radius $\widehat{r}$ is over $\pi/2$, the estimated axis $\widehat{v}$ is flipped to $-\widehat{v}$ and we take the estimated radius $\pi - \widehat{r}$, since radius as a parameter ranges from 0 to $\pi/2$ (For example, if the parameters are yielded as $((0,0,1)^T, \pi/2 + 0.01)$, then we take $(\widehat{v}, \widehat{r}) = ((0,0,1)^T, \pi/2 - 0.01)$ instead). Hence, the distribution of the estimated radius is a folded version of the distribution of the angle between the true axis and the data points. As a result, the expectation of estimated radius, $\mathbb{E}\widehat{r}$ is underestimated even if the true radius is $\pi/2$.

Type 2 Overfitting Phenomenon : Suppose data points are generated along a great subsphere but within a very short interval. This would make the point cloud have a disc shape. In this case, subsphere fitting would yield a small subsphere inside the point cloud, not the original great subsphere where data points were generated, *see* Figure 3.1b.

To mitigate overfitting phenomena, we propose Penalized Principal Nested Spheres (PenPNS). First, PenPNS alleviate type 1 overfitting phenomenon by regularizing r in parameter estimation. Well-known regularization, like penalizing with L_1, L_2 or MCP terms, are examined in simulation studies. Second, to deal with type 2 overfitting phenomenon, PenPNS imposes a penalizing term on the cross-validation error in choosing tuning parameter. This penalizing term is derived from a concept *Index of Dispersion* (IoD), an index which has a large value when a disc-shaped distribution is fitted with a small radius.

The rest of this chapter is organized as follows. In Section 3.2, we formally review the nested subsphere structure of PNS and present the PenPNS framework. The computational algorithm and validation method are explained in Section 3.3. In Section 3.4, the geometry of parameter space is described and conditions to guarantee strong consistency are discussed. In Section 3.5, we present a heuristic method to avoid type 2 overfitting phenomenon. The concepts of IoD and IoD criterion are defined therein. In Section 3.6, the results of numerical simulation are presented, mainly focused on the mitigation of overfitting phenomena. In Section 3.7, we compare the result of the proposed PenPNS with the original PNS and PNGS (Jung,2012) on the Cephalometric X-ray Image Data. Proofs, technical lemmas and extended numerical results are given in Chapter 3 of the supplementary materials.

3.2 Statistical Framework

3.2.1 Nested Subsphere Structure

In this section, we present the concept of the nested subsphere structure, proposed in Jung (2012). Given a natural number m and the unit sphere S^m in $\mathbb{R}^{m+1}$, a subsphere $A_{m-1}(v, r)$ of S^m is defined with a unit vector $v \in S^m$, called an *axis*, and a *radius* $r \in [0, \pi/2]$.

Definition 3.1. *When $m > 1$, a subsphere $A_{m-1}(v, r)$ of S^m with an axis $v \in S^m$ and a radius $r \in [0, \pi/2]$ is defined*

$$A_{m-1}(v, r) = \{x \in S^m : \rho_{m-1}(v, x) = r\},$$

where $\rho_{m-1}(x, y) = \arccos(x^T y)$ denotes the geodesic distance between $x, y \in S^m$. When $m = 1$, a subsphere A_0 is a point on a unit circle S^1.

Note that a subsphere $A_{m-1}(v, r)$ is the intersection of S^m with a m-dimensional affine subspace in $\mathbb{R}^{m+1}$, denoted $H_m(v, r) = \{x \in \mathbb{R}^{m+1} : v^T x - \cos r = 0\}$. The projection from S^m onto $A_{m-1}(v, r)$ is defined as

$$\pi_{m-1}(v, r)(x) = \underset{s \in A_{m-1}(v,r)}{\arg\min}\ \rho_{m-1}(x, s) = \frac{\sin r \cdot x + \sin(\rho_{m-1}(x, v) - r) v}{\sin(\rho_{m-1}(x, v))}$$

for $x \in S^m$ and $m \geq 2$.

We observe that there is an isomorphism between $A_{m-1}(v, r)$ and S^{m-1}, given as $f_{m-1}(v, r) : A_{m-1}(v, r) \to S^{m-1}$ such that

$$f_{m-1}(v, r)(x) = R^-(v) x / \sin r,$$

which has an inverse $f_{m-1}^{-1}(v, r) : S^{m-1} \to A_{m-1}(v, r)$ such that

$$f_{m-1}^{-1}(v, r)(y) = R^T(v) \begin{pmatrix} \sin r \cdot y^T & \cos r \end{pmatrix}^T.$$

Here, $R(v) \in \mathbb{R}^{(m+1)\times(m+1)}$ is a rotation mapping that moves v to the north pole and $R^-(v)$ consists of the first m rows of $R(v)$. By f_{m-1} and f_{m-1}^{-1}, $A_{m-1}(v, r)$ is identified with a unit sphere S^{m-1}.

Let d be a natural number and consider a family of subspheres $A_{m-1}(v_{m-1}, r_{m-1})$ in S^m for each $m = 1, \ldots, d$. A nested sphere $\mathfrak{A}_{m-1}$ corresponding to $A_{m-1}(v_{m-1}, r_{m-1})$ is defined in S^d as follows.

Definition 3.2. *Suppose d is a natural number and there is a family of subspheres $A_{m-1}(v_{m-1}, r_{m-1})$ with axis $v_{m-1} \in S^m$ and radius $r_{m-1} \in [0, \pi/2]$ for each $m = 1, \dots, d$. The projection from S^m onto $A_{m-1}(v_{m-1}, r_{m-1})$ is denoted π_{m-1}. A m-dimensional nested sphere $\mathfrak{A}_{m-1}$ of S^d is defined as*

(1) if $m = d$, $\mathfrak{A}_{d-1} = A_{d-1}(v_{d-1}, r_{d-1})$,

(2) if $1 \leq m < d$, $\mathfrak{A}_{m-1} = f_{d-1}^{-1} \circ \cdots \circ f_{m-1}^{-1}(A_{m-1}(v_{m-1}, r_{m-1}))$,

where $f_{m-1}^{-1} = f_{m-1}^{-1}(v_{m-1}, r_{m-1})$ for each $m = 1, \dots, d$. Then the nested subsphere structure $\mathfrak{A}$ is defined $\mathfrak{A} = \{\mathfrak{A}_0, \dots, \mathfrak{A}_{d-1}\}$.

3.2.2 Estimation of Subspheres

For $d \geq 2$, suppose n data points $x_1^{(1)}, \dots, x_n^{(1)} \in S^d$ are sampled. Our goal is to fit subspheres $\widehat{A}_{d-k}$ for $k = 1, \dots, d$ sequentially. At each step $k = 1, \dots, d-1$, the estimated subsphere $\widehat{A}_{d-k}$ is obtained by estimating its parameters v_{d-k} and r_{d-k}, that is, $\widehat{A}_{d-k} = A_{d-k}(\widehat{v}_{d-k}, \widehat{r}_{d-k})$. After fitting $\widehat{A}_{d-k}$, we update data points as $x_i^{(k+1)} = f_{d-k}(\widehat{v}, \widehat{r}) \circ \pi_{d-k}(\widehat{v}, \widehat{r})(x_i^{(k)})$ for all $i = 1, \dots, n$, so that the new data points reside on S^{d-k}.

For each step $1 \leq k < d$, we seek to minimize the objective function

$$F(v, r) = \frac{1}{2n} \sum_{i=1}^{n} L\left(x_i^{(k)}; v, r\right) + j_{\tau_k}(r), \tag{3.1}$$

over $(v, r) \in S^{d-k+1} \times [0, \pi/2]$. Here, $L(x; v, r)$ is a loss function that measures a discrepancy between a subsphere $A_{d-k}(v, r)$ and a data point $x \in S^{d-k+1}$. The function j_{τ_k} introduces the form of penalization, where the value of j_{τ_k} is larger for the smaller value of the radius r. The tuning parameter $\tau_k \in [0, \infty)$ controls the degree of penalization. The choice of loss function and penalty function is presented in the subsections below.

At the last step $k = d$, we compute $\widehat{A}_0$ as the Fréchet mean of $x_1^{(d)}, \ldots, x_n^{(d)} \in S^1$, that is,

$$\widehat{A}_0 = \arg\min_{s \in S^1} \sum_{i=1}^{n} \rho(x_i, s).$$

The loss function corresponding to the original PNS proposal (Jung,2012) is

$$L_I(x; v, r) = (\rho_{m-1}(x, v) - r)^2 = (\arccos(x^T v) - r)^2, \qquad (3.2)$$

which may be called the *intrinsic squared loss* function. For PenPNS, we follow the instrinsic squared loss as the original PNS.

We use the penalty function j_{τ_k} in order to penalize the departure from great sphere ($r = \pi/2$) by assigning large value for smaller r. For the intrinsic loss functions, (3.2), we set the penalty function in the forms of L_1 and L_2 norms,

$$j_\tau^{L_1}(r) = \tau \cdot |\pi/2 - r|, , \qquad (3.3)$$

$$j_\tau^{L_2}(r) = \tau \cdot (\pi/2 - r)^2, \qquad (3.4)$$

and also use the MCP (Zhang,2010)

$$j_\tau^{MCP}(r) = \begin{cases} \tau \cdot (\pi/2 - r) - (\pi/2 - r)^2/2\gamma, & \text{if } (\pi/2 - r) \leq \gamma\tau \\ \gamma\tau^2/2, & \text{otherwise} \end{cases} \qquad (3.5)$$

with a constant $\gamma > 1$.

3.3 Computational Algorithms

A general strategy of solving (1) is an alternating minimization method. Given a current iterate (v^0, r^0), we first optimize over v with the fixed r^0, obtaining v^1 satisfying $F(v^1, r^0) \leq F(v^0, r^0)$. Then we obtain r^1 from given v^1, satisfying $F(v^1, r^1) \leq F(v^1, r^0)$.

The algorithm for minimization of F with respect to v is discussed in Section 3.3.1. Given v, the problem of minimizing F in terms of r has a closed form solution with appropriate L_1, L_2 or MCP penalties, as discussed in Section 3.3.2. We propose to optimize $F(v, r)$ as follows. For the jth iterate (v^j, r^j),

1. Obtain v^{j+1} satisfying $F(v^j, r^{j+1}) \leq F(v^j, r^j)$ (Section 3.3.1).
2. Obtain r^{j+1} satisfying $F(v^{j+1}, r^{j+1}) \leq F(v^j, r^{j+1})$ (Section 3.3.2).
3. Stop if $F(v^j, r^j) - F(v^{j+1}, r^{j+1}) < \epsilon$ for a predetermined threshold $\epsilon > 0$. Otherwise, return to the $j + 1$th step.

3.3.1 Update of v

The current axis v^j is updated by solving the following nonlinear least-squares problem,

$$v^{j+1} = \underset{v \in S^{d-k+1}}{\arg\min} \sum_{i=1}^{n} \left(\rho_{d-k}(x_i^{(k)}, v^j) - r^j \right)^2,$$

numerically, *e.g.* Levenberg-Marquardt Algorithm (Levenberg,1944; Marquardt,1963).

3.3.2 Update of r

Suppose we have datasets $x_i^{(k)} \in S^{m-k+1}$ for $i = 1, \dots, n$. It is easy to check that the solution to

$$r^{j+1} = \arg\min_{r \in \mathbb{R}} \frac{1}{2n} \sum_{i=1}^{n} (\rho_{m-1}(x_i^{(k)}, v^{j+1}) - r)^2 + j_\tau(r),$$

is given as (1) for L_1 penalty, $r^{j+1} = \pi/2 - \text{sign}(\pi/2 - \overline{a})(|\pi/2 - \overline{a}| - \tau)_+$ and (2) for L_2 penalty, $r^{j+1} = (\overline{a} + \tau\pi/2)/(1+\tau)$ and (3) for the MCP,

$$r^{j+1} = \begin{cases} \overline{a}, & \text{if } |\pi/2 - \overline{a}| > \gamma\tau \\ \frac{\pi}{2} - \frac{\gamma}{\gamma-1}\text{sign}(\pi/2 - \overline{a})\left(|\pi/2 - \overline{a}| - \tau\right)_+, & \text{otherwise} \end{cases},$$

where $\overline{a} = \frac{1}{n}\sum_{i=1}^{n} \arccos(v^T x_i^{(k)})$. Here, for $a \in \mathbb{R}$, $a_+ = a$ if $a \geq 0$ and $a_+ = 0$ if $a < 0$.

Remark 3.1. *In particular, using the L_1 penalty and the MCP, i.e., $j_\tau^{L_1}(\pi/2 - r)$ and $j_\tau^{MCP}(\pi/2 - r)$, will ensure that the updated radius r^{j+1} is $\pi/2$ for large enough τ.*

Remark 3.2. *In the original PNS, $r^{j+1} = \overline{a}$.*

3.3.3 Tuning Parameter Selection

For every steps $k = 1, \dots, d$, We use the cross-validation method to select the value of tuning parameter $\tau_k \in [0, \infty)$. Let the number of cross-validation be B. For the bth instance of the cross-validation with $b = 1, \dots, B$, we split the n data points $X^{(k)} = \{x_1^{(k)}, \dots, x_n^{(k)}\}$ into two disjoint sets, the training set $X_{tr,b}^{(k)}$ of size $\lfloor (B-1)/B \cdot n \rfloor$ and the test set $X_{test,b}^{(k)} = \{x_{1,test,b}^{(k)}, \dots, x_{n',test,b}^{(k)}\}$ of size $n' = \lceil 1/B \cdot n \rceil$. Given $X_{tr,b}^{(k)}$ and a grid of τ_k's, we solve the optimization problem (3.1) for each τ_k and the estimated parameters are denoted $\widehat{v}_{tr,b}(\tau_k)$ and $\widehat{r}_{tr,b}(\tau_k)$.

To evaluate the accuracy of the prediction, we estimate the prediction error for each τ_k by the cross-validation error,

$$\widehat{\mathrm{Err}}_k(\tau_k) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{n'} \sum_{i=1}^{n'} L\left(x_{i,test,b}^{(k)}; \widehat{v}_{tr,b}(\tau_k), \widehat{r}_{tr,b}(\tau_k)\right),$$

where $L(x; v, t)$ is a loss function. The best tuning parameter $\widehat{\tau}_{best,k}$ is the one that minimizes $\widehat{\mathrm{Err}}_k$ over the interval $\tau_k \in [0, \infty)$. Finally, we refit the model on the full data $X^{(k)}$ with $\widehat{\tau}_{best,k}$ and the estimated parameters are denote $\widehat{v}_{k,best}$ and $\widehat{r}_{k,best}$.

3.4 Theory

We consider X to be a S^m-valued random variable, which is a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ into $(S^m, \mathcal{G})$, where $\mathcal{G}$ denotes the Borel σ-algebra of S^m. Let $X_1, \ldots, X_n$ be random variables generated identically and independently from the same law as X.

3.4.1 Geometry of Parameter Space

In this section, let a small subsphere $p(v, r)$ on S^m for $m \geq 2$ be parameterized with a pair of a center and a radius $(v, r) \in \Theta_0 = S^m \times [0, \pi/2]$, or explicitly

$$p(v, r) = \{x \in S^m : \rho(v, x) = r\},$$

where $\rho(x, y) = \arccos(x^T y)$ denotes the geodesic distance between $x, y \in S^m$.

However, when $r = \pi/2$, two small subspheres $p(v, \pi/2)$ and $p(-v, \pi/2)$ represent the same small sphere. To remedy this, we define an equivalence relation $\sim$ that identifies $p(v, r)$ with $p(-v, r)$ whenever $r = \pi/2$. The parameter space Θ is defined as $\Theta = \Theta_0/\sim$, that is, $\Theta = \{[v, r] : (v, r) \in \Theta_0\}$. Note that both Θ_0 and Θ are compact spaces.

A distance between $p_1 = [v_1, r_1]$ and $p_2 = [v_2, r_2] \in \Theta$ is defined as

$$d(p_1, p_2) = \min\left(d_1(p_1, p_2), d_2(p_1, p_2)\right)$$

where

$$\begin{aligned} d_1(p_1, p_2) &= \left(\arccos(v_1^T v_2)^2 + |r_1 - r_2|^2\right)^{1/2} \\ d_2(p_1, p_2) &= \left(\arccos(-v_1^T v_2)^2 + |\pi - r_1 - r_2|^2\right)^{1/2}. \end{aligned}$$

Lemma 3.1. *The distance d is a metric.*

For a pair of parameters $p = [v, r] \in \Theta$, the projection from S^m onto $p_0(v, r)$ is defined as

$$\pi(p)(x) = \underset{s \in p(v,r)}{\arg\min}\, \rho(x, s) = \frac{\sin r \cdot x + \sin(\rho(x, v) - r)v}{\sin(\rho(x, v))}$$

for $x \in S^m$. Abusing the notation, we denote $\rho(x, p) = \inf_{s \in p(v,r)} \rho(x, s)$ with $p \in \Theta$. Since ρ is continuous and S^m is compact, uniform continuity holds for ρ, that is, for every $x \in S^m$, $p \in \Theta$ and $\epsilon > 0$, there exists a $\delta > 0$ such that $|\rho(x, p') - \rho(x, p)| < \epsilon$ for all $p' \in \Theta$ with $d(p, p') < \delta$.

3.4.2 Strong Consistency

We discuss a consistency of the subsphere estimator $\widehat{p} = p(\widehat{v}, \widehat{r}) \in \Theta$ for $X_1, \ldots, X_n$ on S^m, where the model is equipped with a penalty. Our aim is to show the strong consistency with respect to the distance d, where $\widehat{p}$ is obtained by solving the minimization of a penalized likelihood function

$$\min_{p\in\Theta}\sum_{i=1}^{n}\rho^2(X_i,p)+j_\tau(r),$$

where $j_\tau:[0,\pi/2]\to[0,\infty)$ is a function with the following properties.

(P1) $j_\tau(x)$ is decreasing on $x\in[0,\pi/2]$

(P2) $j_\tau(\pi/2)=0$ and $j_\tau(x)$ is left continuous as $x\to\pi/2$.

(P3) $j_\tau(x)$ is continuous in $x\in[0,\pi/2)$

(P4) $j_\tau(x)\to 0$ monotonically as $j\to 0$ for all $x\in[0,\pi/2]$

Example 3.2. *Penalties in (3.3),(3.4) and (3.5), that is, L_1, L_2, and MCP penalties, satisfy (P1) - (P4).*

In Bhattacharya and Patrangenaru (2003) and Huckemann (2011), the population and sample Fréchet ρ-mean set for the unpenalized likelihood are suggested by as

$$E=\left\{p\in\Theta:\mathbb{E}_X\rho^2(X,p)=\inf_{q\in\Theta}\mathbb{E}_X\rho^2(X,q)\right\}$$

and

$$E_n=\left\{p\in\Theta:\sum_{i=1}^{n}\rho^2(X_i,p)=\inf_{q\in\Theta}\sum_{i=1}^{n}\rho^2(X_i,q)\right\},$$

For the penalized function with a penalty j_τ, we define the population solution set as

$$E^\tau=\left\{p=[v,r]\in\Theta:\mathbb{E}_X\rho^2(X,p)+j_\tau(r)=\inf_{q=[v',r']\in\Theta}\mathbb{E}_X\rho^2(X,q)+j_\tau(r')\right\}$$

and the sample solution set as

$$E_n^\tau = \left\{ p = [v, r] \in \Theta : \sum_{i=1}^{n} \rho^2(X_i, p) + j_\tau(r) = \inf_{q=[v',r']\in\Theta} \sum_{i=1}^{n} \rho^2(X_i, q) + j_\tau(r') \right\}.$$

Next, we introduce the following modes of convergence (Schotz,2022).

Definition 3.3. *Let (Θ, d) be a metric space. For a sequence of sets $E_n \subset \Theta$ for $n = 1, 2, \ldots$, the outer limit of E_n is*

$$\cap_{n=1}^{\infty} \overline{\cup_{k=n}^{\infty} E_k}.$$

The one-sided Hausdorff distance between two sets E_1 and E_2 in (Θ, d) is defined as

$$d_H(E_1, E_2) = \sup_{x_1 \in E_1} \inf_{x_2 \in E_2} d(x_1, x_2).$$

Strong consistency of random closed sets, on the ground of two modes of convergence above, were developed by Ziezold (1977) and Bhattacharya and Patrangenaru (2003) as follows.

Definition 3.4. *Let (Θ, d) a metric space. For a random closed set $E \subset \Theta$ and a sequence of random closed sets $E_n \subset \Theta$ for $n = 1, 2, \ldots$,*

(ZC) *E_n converges to E strongly in the sense of Ziezold (1977) if*

$$\cap_{n=1}^{\infty} \overline{\cup_{k=n}^{\infty} E_k} \subset E \quad a.s.$$

(BPC) *E_n converges to E strongly in the sense of Bhattacharya (2003) if for every $\epsilon > 0$, there is a sufficiently large $n > 0$ such that*

$$\cup_{k=n}^{\infty} E_k \subset \{p \in \Theta : \exists q \in E \ s.t. \ d(p, q) \le \epsilon\} \quad a.s.$$

(ZC) and (BPC) are indeed random closed set version of convergence in outer limit or one-sided Hausdorff distance respectively. Under the condition that Θ is compact, (ZC) and (BPC) imply each other.

Proposition 3.1. *Let* (Θ, d) *be a metric space with* Θ *compact. Consider a random closed set* E *and a sequence of random closed sets* E_n *for* $n = 1, 2, \ldots$. E_n *converges to* E *in the sense of (ZC) if and only if* E_n *converges to* E *in the sense of (BPC).*

We now prove the strong consistency of the sample solution sets E_n^τ, extending the discussions in Huckemann (2011). First, we assert the strong consistency (ZC) for a given tuning parameter $\tau \geq 0$.

Theorem 3.1. *For a given tuning parameter* $\tau \geq 0$, E_n^τ *converges to* E^τ *in the sense of (ZC) if the penalty function* j_τ *satisfies (P1) to (P3).*

The sets E^τ are deterministic and for any decreasing sequences $\tau_k \to 0$, E^{τ_k} converges to E in outer limit.

Theorem 3.2. *Let* $\tau_k \to 0$ *be any decreasing sequences. Then* E^{τ_k} *converges to* E *in outer limit if the penalty function* j_τ *satisfies (P4).*

Or we can show the consistency of E_n^τ by setting $\tau_n \to 0$ as n goes to infinity.

Theorem 3.3. *Let* τ_n *be a decreasing sequence as* $n \to 0$. *Then* E_n^τ *converges to* E^τ *in the sense of (ZC) with respect to* d *in* Θ *if the penalty function* p_τ *satisfies (P1) to (P4).*

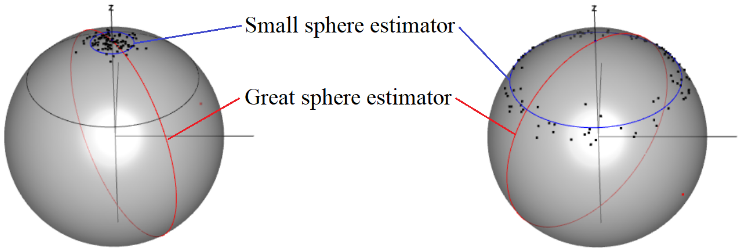


Figure 3.2: A small sphere estimator and a great sphere estimator on a disc-shaped distribution (left) and a doughnut-shaped distribution (right).

We provide proofs of theorems in this section and examples for relative independence and absolute orthogonality in Section B.1 of the supplementary materials.

3.5 IoD Criterion to Avoid Overfitting

Type 2 overfitting phenomenon usually occurs when data points are generated along a great subsphere but within a very short interval. To emulate this case, we assume that the data points follow a highly concentrated unimodal distribution, what we call a disc-shaped distribution. This disc-shaped distribution is a counterpart of a doughnut-shaped distribution, where the distribution is highly concentrating along a small sphere mode, like a doughnut shape. Small sphere estimator resides inside the disc-shaped cloud of data points and its radius is unnecessarily small but with much lower variance, *see* Figure 3.2.

For more practical interpretability, our aim is to estimate this type of data with a great circle estimator. For that, we propose a concept of *Index of Dispersion* (IoD), an index which reflects how much data points are congested in a small area. If the data point cloud has a disc shape, IoD will have a high value. In contrast, if the data point cloud has a doughnut shape, IoD will have a low value. We utlize IoD in tuning parameter selection, where the radius of small sphere estimator varies as the tuning parameter τ does. By penalizing cross-validation errors with IoD value, we avoid estimate a disc-shaped distribution with a small subsphere with a small radius.

3.5.1 Index of Dispersion (IoD) and IoD Criterion

We first define the index of dispersion (IoD) with respect to a direction v for a distribution on the sphere S^m.

Definition 3.5. *Suppose a random variable X has a distribution ν on S^m. Given a vector $v \in S^m$, the angle between v and X is denoted as $\theta_v = \arccos(v^T X)$. The first and second moments of θ_v are given as*

$$\mu_1 = \int_{S^m} \theta_v \, d\nu(X), \qquad \mu_2 = \int_{S^m} \theta_v^2 \, d\nu(X).$$

Then the index of dispersion (IoD) with respect to a direction v is defined as

$$IoD(v) = \left(\frac{\mu_2 - \mu_1^2}{\mu_1} \right)^2 .$$

Remark 3.3. *Our concept of IoD is inspired from the original version of IoD from Cox (1966), which has a form σ^2/μ, thus it is indeed correct to call our version as squared IoD. Bur for* *convenience in naming, we will allow a slight misnomer for this concept in our thesis.*

For a highly concentrated unimodal (disc-shaped) distribution around $v \in S^m$, $\mathrm{IoD}(v)$ tends to be large since density gathers in a small area on the sphere, and therefore the mean value of θ_v is small. In contrast, when density is highly concentrated along a small or great circle (doughnut-shaped) and v coincides with the center of the circle, the mean value of θ_v grows larger and the value of $\mathrm{IoD}(v)$ decreases as much.

The sample version of the IoD is defined in view of the penalized PNS estimates.

Definition 3.6. *Given a small circle estimator* $A(\widehat{v}(\tau), \widehat{r}(\tau))$ *from the penalized PNS, we compute the azimuthal angles* $\{\theta_{v,i}\}_{i=1}^n$ *from the dataset* $\{x_i\}_{i=1}^n$ *such that* $\theta_{v,i} = \arccos(\widehat{v}(\tau)^T x_i)$*. Then the sample version of IoD with respect to* $\widehat{v}(\tau)$ *is defined as*

$$\widehat{IoD}(\widehat{v}(\tau)) = \left(\frac{\sum_{i=1}^n \theta_{v,i}^2/n - (\sum_{i=1}^n \theta_{v,i}/n)^2}{\sum_{i=1}^n \theta_{v,i}/n} \right)^2$$

From now on we denote $\widehat{IoD}(\widehat{v}(\tau))$ *as* $\widehat{IoD}(\tau)$*.*

When τ goes down, say τ_s, the small circle estimator has a smaller radius and $\widehat{v}$ tends to approach to the cloud of data points. Thus, in this case, the small circle estimator should have a larger sample IoD, $\widehat{\mathrm{IoD}}(\tau_s)$. In contrast, when τ becomes bigger, say τ_l, the small circle tends to have a larger radius and the axis gets further from the cloud of data points. In this case, the small circle estimator has a smaller sample IoD, $\widehat{\mathrm{IoD}}(\tau_l)$, that is, $\widehat{\mathrm{IoD}}(\tau_l) < \widehat{\mathrm{IoD}}(\tau_s)$.

Our strategy is, in order to compensate the overfitting phenomenon in PNS, to add an compensating IoD to the validation error.

For that, we first suggest a concept of a criterion by adding an IoD term to the *empirical risk*,

$$\frac{1}{n}\sum_{i=1}^{n} L\left(x_i; \widehat{v}(\tau), \widehat{r}(\tau)\right),$$

given each value of tuning parameter τ_k. Given data points, we define a criterion for the small sphere estimator by penalizing the empirical risk as follows.

Definition 3.7. *Given a dataset* $x_1, \ldots, x_n$ *on* S^m *and a small sphere estimator* $A(\widehat{v}(\tau), \widehat{r}(\tau))$ *from the penalized PNS, we define the IoD criterion* $\widehat{IC}_C$ *with a constant* $C > 0$

$$\widehat{IC}_C(\tau) = \frac{1}{n}\sum_{i=1}^{n} L\left(x_i; \widehat{v}(\tau), \widehat{r}(\tau)\right) + C \cdot \widehat{IoD}(\tau)$$

In tuning parameter selection, we derive the IoD criterion $\widehat{\mathrm{IC}}_{C,(b)}(\tau)$ for each test set $X_{test,b}$ with $b = 1, \ldots, B$

$$\widehat{\mathrm{IC}}_{C,(b)}(\tau) = \frac{1}{n'}\sum_{i=1}^{n'} L(x_{i,test,b}; \widehat{v}_{tr,b}(\tau), \widehat{r}_{tr,b}(\tau)) + C \cdot \widehat{\mathrm{IoD}}_{(b)}(\tau),$$

where $\widehat{\mathrm{IoD}}_{(b)}(\tau)$ is obtained from the test set $X_{test,b}$. The IoD criteria are aggregated to $\widehat{\mathrm{IC}}_{test,C}(\tau) = 1/B \cdot \sum_{b=1}^{B} \widehat{\mathrm{IC}}_{C,(b)}(\tau)$ and we select the best tuning parameter $\tau_{best,C}$ that minimizes $\widehat{\mathrm{IC}}_{test,C}$.

3.5.2 Choice of the Constant C

The major issue here is the choice of the constant C. We assume a situation in the population setting, where we estimate a data point cloud with two representing cases, a small sphere estimator and a great sphere estimator. We determine the value of the constant C such that the IoD term compensates the difference of risks between small and great sphere estimators.

To put it formally, we first define IoD criterion in the population setting, by adding an IoD term to the *risk*, $\mathbb{E}L\left(X;v,r\right)$, as follows.

Definition 3.8. *Suppose we have a random variable $X \in S^m$. Given a small circle $A(v,r)$, which is independent from X, we define a IoD criterion for X and $A(v,r)$*

$$IC_C(v,r) = \mathbb{E}L\left(X;v,r\right) + C \cdot IoD(v)$$

with some constant $C > 0$.

We adopt a heuristic method to choose the value of C, whose brief overview is as follows.

We consider a disc-shaped distribution scenario. For each cases, we suppose generic situations where the distribution is estimated with small or great sphere estimator. Then the variance and IoD of two estimators are computed and we find the range of C where the overfitting phenomenon is mitigated.

We denote the parameters for a small sphere estimator as (v_{small}, r_{small}) and the ones for a great sphere estimator as (v_{great}, r_{great}). Then, our aim is to find a constant C such that type 2 overfitting phenomenon is mitigated for disc-shaped distribution, that is, IoD criterion is bigger for small sphere estimator than great sphere estimator. Our aim is summarized as follows.

$$\mathrm{IC}_C(v_{small}, r_{small}) > \mathrm{IC}_C(v_{great}, r_{great}). \tag{3.6}$$

For notations, suppose $X \in S^m$ follows a disc or doughnut-shaped distribution with a pole μ. Let v_s and v_g be the center of the small and great sphere estimator for X. Let $\theta_g = \arccos(v_g^T X)$, $\theta_s = \arccos(v_s^T X)$ and $\theta_\mu = \arccos(\mu^T X)$.

Disc-shaped distribution

We call a high concentrated unimodal distribution with a mode μ as a *disc-shaped distribution* on the sphere S^m with a pole μ. We consider a disc-shaped distribution invariant to the rotation around μ and whose support is a convex subset of S^m that includes μ, which is not a point density on μ nor uniform density on the whole surface of S^m.

For a disc-shaped distribution, we set a small sphere estimator concentric to X, that is, $\theta_s = \theta_\mu$ and its radius coincides with $\mathbb{E}\theta_s$. In contrast, we set a *great sphere estimator* to pass across the density through μ. Then the center of the great sphere v makes a perpendicular angle with the pole μ, $\mathbb{E}\theta_g = \pi/2$, *see* Figure 3.2 left. The small sphere estimation represents the overfitting phenomenon.

Note that the risk of the small sphere estimator, $\mathbb{E}L(X; v_s, r_s)$ is the same as the variance of θ_s. It is the same for the great sphere estimator, that is, $\mathbb{E}L(X; v_g, r_g) = \mathrm{Var}(\theta_g)$.

To satisfy (3.6), we require C should be large than the division of the variance difference by the IoD difference, that is,

$$C > \left(\mathrm{Var}(\theta_g) - \mathrm{Var}(\theta_s)\right) / \left(\mathrm{IoD}(v_s) - \mathrm{IoD}(v_g)\right), \tag{3.7}$$

so that the IoD criterion (IC_C) of the small sphere estimator is larger than that of the great sphere estimator.

In practice, we set C as two times the RHS value of (3.7). The determination of C depends on the distribution of data points.

In the following examples, we choose the value of C for disc-shape distributions. For convenience, we use $p = m + 1$ to follow the notational convention $S^{p-1} = S^m$.

Example 3.3. *We first consider the spherical τ-ball uniform distribution, $\mathcal{S}_\tau$, for some small $\tau > 0$. The sectional function of $\mathcal{S}_\tau$ is*

$$g_{S_\tau}(\theta_1) = c(p, \tau)\mathbf{1}_{(\theta_1 < \tau)},$$

where $c(p, \tau)$ is the normalizing term. Using the following approximation formula

$$\int_0^\tau x^q \sin^p x \, dx = \frac{1}{p+q+1}\tau^{p+q+1} - \frac{1}{6}\frac{p}{p+q+3}\tau^{p+q+3} + \mathcal{O}(\tau^{p+q+5}),$$

we can obtain an approximating expression for the normalizing term $c(p, \tau)$,

$$c(p, \tau) = \frac{\Gamma\left(\frac{p-1}{2}\right)}{2\pi^{\frac{p-1}{2}}}\left[\frac{p-1}{\tau^{p-1}} + \frac{1}{6}\frac{(p-1)^2(p-2)}{p+1}\frac{1}{\tau^{p-3}}\right] + \mathcal{O}(\tau^{5-p}).$$

The variance and IoD for both small and great circle estimates are approximated as

$$\begin{aligned}
Var_{small} &= \frac{p-1}{p^2(p+1)}\tau^2 + \mathcal{O}(\tau^4) \\
Var_{great} &= \frac{1}{(p+1)}\tau^2 + \mathcal{O}(\tau^4) \\
IoD_{small} &= \frac{1}{p^2(p+1)^2}\tau^2 + \mathcal{O}(\tau^4) \\
IoD_{great} &= \frac{4}{\pi^2}\frac{1}{(p+1)^2}\tau^4 + \mathcal{O}(\tau^6)
\end{aligned}$$

From these, the compensating factor C must satisfy

$$C \geq (p+1)(p^2 - p + 1).$$

We observe that C diverges at a polynomial rate of order p^3 with fixed τ.

Example 3.4. *In the case of positive kurtosis, we propose the following probability density on S^{p-1}*

$$S_{2,\tau}(x; \mu, \tau) = c(p,\tau)\mathbf{1}_{(\arccos \mu^T x < \tau)} \left(1 - \frac{1}{\tau^2}\arccos^2(\mu^T x)\right), \qquad x \in S^{p-1}$$

with the mode $\mu \in S^{p-1}$ and sufficiently small $\tau > 0$. The normalizing term in an approximated polynomial form is

$$c(p,\tau) \approx \frac{\Gamma\left(\frac{p-1}{2}\right)}{2\pi^{\frac{p-1}{2}}} \left[\frac{(p-1)(p+1)}{2}\frac{1}{\tau^{p-1}} + \frac{1}{12}\frac{(p-1)^2(p-2)(p+1)}{p+3}\frac{1}{\tau^{p-3}}\right]$$

The sectional function is

$$g_{S_{2,\tau}}(\theta_1) = \mathbf{1}_{(\theta_1 < \tau)}\left(1 - \frac{1}{\tau^2}\theta_1^2\right).$$

The variance and IoD for both small and great circle estimates are approximated as

$$\begin{aligned}
Var_{small} &= \frac{(p-1)(2p^2+4p+3)}{p^2(p+2)^2(p+3)}\tau^2 + \mathcal{O}(\tau^4) \\
Var_{great} &= \frac{1}{p+3}\tau^2 + \mathcal{O}(\tau^4) \\
IoD_{small} &= \frac{(2p^2+4p+3)^2}{p^2(p+1)^2(p+2)^2(p+3)^2}\tau^2 + \mathcal{O}(\tau^4) \\
IoD_{great} &= \frac{4}{\pi^2}\frac{1}{(p+3)^2}\tau^4 + \mathcal{O}(\tau^6)
\end{aligned}$$

From these, the compensating factor C must satisfy

$$C > \frac{p^2(p+1)^2(p+2)^2(p+3)}{(2p^2+4p+3)^2}.$$

Here C diverges at a polynomial rate of order p^3 with fixed τ too.

Example 3.5. *In the case of von Mises-Fisher distribution (Fisher,1953; Mardia and Jupp,1999), it is computationally intractable to derive an analytic form for moments of the azimuthal angle θ_1 directly. Instead, we approximate highly-concentrated von Mises-Fisher distribution as a normal distribution on a q-dimensional tangent plane ($q = p - 1$). On the tangent plane, a radius r from the tangent point is equivalent to the azimuthal angle θ_1 from the pole μ on the sphere S^{p-1}. Here we set a spherical coordinate on the tangent plane as $(r, \theta'_1, \ldots, \theta'_{q-2}, \phi')$. The marginal density of r is given as*

$$g_{normal}(r) = \frac{1}{(2\pi\sigma^2)^{q/2}} \frac{2\pi^{q/2}}{\Gamma(q/2)} r^{q-1} e^{-r^2/2\sigma^2}$$

where σ^2 is the variance of the normal distribution on the tangent plane and very close to 0. The moments of r are then

$$\begin{aligned}\mu_1 &= \sqrt{2}\sigma \frac{\Gamma((q+1)/2)}{\Gamma(q/2)} \\ &= \sqrt{2}\sigma\sqrt{\frac{q}{2}}\left(1 - \frac{1}{4q} + \mathcal{O}(q^{-2})\right) \\ \mu_2 &= \sigma^2 q.\end{aligned}$$

For μ_1, we used the following asymptotic expansion (Qi,2010)

$$\frac{\Gamma(x+1/2)}{\Gamma(x)} = x^{1/2} - \frac{1}{8}x^{-1/2} + \mathcal{O}(x^{-3/2})$$

as x goes to infinity. From these, the variance and IoD for small circle estimate are

$$
\begin{aligned}
Var_{small} &= \frac{\sigma^2}{2} + \mathcal{O}(q^{-1}) \\
IoD_{small} &= \frac{\sigma^2}{2(2q-1)} + \mathcal{O}(q^{-2})
\end{aligned}
$$

The variance of θ_r for a great circle estimate, or the second moment of the signed residual

$$\frac{\pi}{2} - \theta_v \approx r \sin\theta'_1 \ldots \sin\theta'_{q-2} \sin\phi'$$

is given by

$$
\begin{aligned}
Var_{great} = &\frac{1}{(2\pi\sigma^2)^{q/2}} \int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} e^{-r^2/2\sigma^2} r^{q+1} \sin^{q+1}\theta'_1 \cdots \sin^3\theta'_{q-2} \sin^2\phi' \, d\theta'_1 \cdots d\phi \\
&= \sigma^2.
\end{aligned}
$$

and the IoD for a great circle estimate is

$$IoD_{great} = \sigma^4/(\pi/2)^2.$$

From these, the compensating factor C must satisfy

$$C > 2p - 1.$$

Here C diverges at a polynomial rate of order 1 with respect to p.

In Simulation Study (Section 3.6) and Real Data Analysis (Section 3.7), we use the result of Example 3.5. Heuristically, to emphasize the contrast, we use the constant, two times $2p - 1$.

Notice that the determination of C only depends on the dimension p (or equivalently, m).

3.5.3 Geometric Interpretation of Model Bias

Suppose the data points on S^m are concentrated near a subsphere A_{m-1} of radius r_m, and we wish to recover A_{m-1}. Generally, it turns out that the subsphere estimate does not work well, but rather it is biased in radius. The bias occurs from the curvature of A_{m-1}, in that there are more chances the noise mass falls outward A_{m-1} than inward, and so the estimate tends to have a larger radius than A_{m-1}. (Briefly, we use a term "outward" for a point $x \in S^m$ if $\arccos(v^T x) > r_m$ and "inward" if $\arccos(v^T x) < r_m$.)

We give a illustrative example. Let A_{m-1} be a subsphere on S^m which has an axis v at the north pole and a radius $0 < r_m \leq \pi/2$. The points x_{0i} for $i = 1, \ldots, n$ are distributed on A_{m-1}. Each data point x_i is generated from a rotationally invariant distribution with a mean direction x_{0i}.

It is easy to notice that the noise mass outward A_{m-1} is bigger than that of inward, so that x_i has more chance to posited outward A_{m-1} than inward, *see* Figure 3.3. Thus the estimated radius $\widehat{r}_m$ is bigger than the true radius r_m generally. A similar example on the plane was discussed in Hastie (1984).

We quantify the bias in the case where the noise follows the von Mises-Fisher distribution on S^m (Fisher,1953; Mardia and Jupp,1999). Consider a vector μ such that $\arccos(\mu^T v) = r$ and the von Mises-Fisher distribution with the mean direction μ and the concentration parameter κ whose density is given as

$$f_{\text{vMF}}(\mu^T x; \kappa) = C_{m+1}(\kappa) \exp(\kappa \cdot \mu^T x),$$

where the normalizing constant is $C_{m+1}(\kappa) = \kappa^{(m-1)/2} / \left((2\pi)^{(m+1)/2} I_{(m+1)/2}(\kappa)\right)$ with $I_k(x)$ the modified Bessel function of order k for $x \geq 0$. We

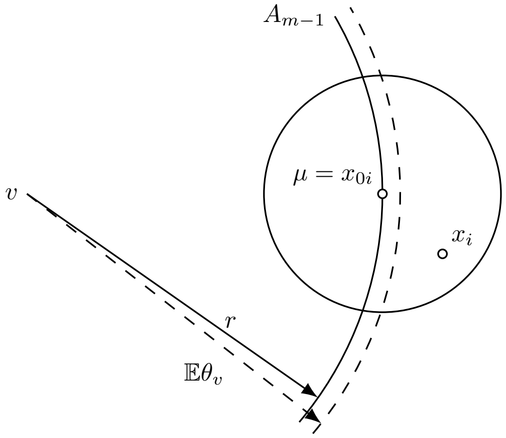


Figure 3.3: A bird's eye view from the direction of v. Each data point x_i is generated from a rotationally invariant distribution (circle) with a mean direction $\mu = x_{0i}$.

suggest the following proposition about computing the expectation of $\theta_v = \arccos(v^T X)$, the angle between X and an arbitrary point $v \in S^m$ on the sphere. Then the bias is defined as $\mathbb{E}\theta_v - r$.

Proposition 3.2. *Let $v \in S^m$. Suppose a random vector $X \in S^m$ follows the von Mises-Fisher distribution with the mean direction $\mu \in S^m$ and the concentration parameter κ. Then the expectation of $\theta_v = \arccos(v^T X)$ is given as*

$$\mathbb{E}\theta_v = \int_{S^m} \arccos(v^T x) C(\kappa) exp(\kappa \cdot \mu^T x)\, d\sigma(x) \tag{3.8}$$

$$= \sum_{l=0}^{\infty} A(l) Y_l^0(v^T \mu) \frac{\omega_m}{C_l^{(m-1)/2}(1)} \int_{-1}^{1} \arccos(t) C_l^{(m-1)/2}(t)(1-t^2)^{(m-2)/2}\, dt, \tag{3.9}$$

where $\lambda = (m-1)/2$, $\omega_{m+1} = 2\pi^{(m+1)/2}/\Gamma((m+1)/2)$ *the surface area of* S^{m+1}, $Y_l^0(\cos\theta)$ *the spherical harmonic function of degree* l *and order* 0 *for* $0 \leq \theta \leq \pi$, $C_l^k(x)$ *the Gegenbauer polynomial of degree* l *and order* k *for* $x \in [-1,1]$, $\Gamma(x)$ *the gamma function for* $x > 0$,

$$c_l = \frac{\sqrt{\pi}}{(m-1)_l 2^{m-2}} \frac{\Gamma(m+l-1)}{\Gamma(m/2)\Gamma((m-1)/2)} \frac{I_{(m-1)/2+l}(\kappa)}{I_{(m+1)/2-1}(\kappa)},$$

and

$$A(l) = \sum_{l=0}^{\infty} \frac{1}{\omega_{m+1}} \frac{l+\lambda}{\lambda} c_l \sqrt{\frac{(m-1)_l \frac{m-1}{2}}{l!(l+\frac{m-1}{2})}}.$$

For the numerical computation of $\mathbb{E}\theta_v$, *see* Remark B.3 of Section B.1.6 in the supplementary materials. In a brief numerical study, where the angle $\arccos(v^T\mu)$ ranges from 0 to $\pi/2$, we observe that the bias always shows positive values and decreases monotonically, *see* Figure 3.4. The bias vanishes only when v and μ are perpendicular. Also the bias becomes larger as the dimension m increases. This is due to the curse of dimensionality, in that the more the dimension m grows, the more the angle between separate points on S^m become perpendicular. However, the bias becomes smaller as the concentration parameter κ increases, that is, the noise mass is more concentrated around μ.

3.6 Simulation Study

3.6.1 Example Dataset Generation

In this section, we shall evaluate the performance of our method. We consider the following random point generation procedures on S^d with *von Mises-Fisher error along a small arc.*

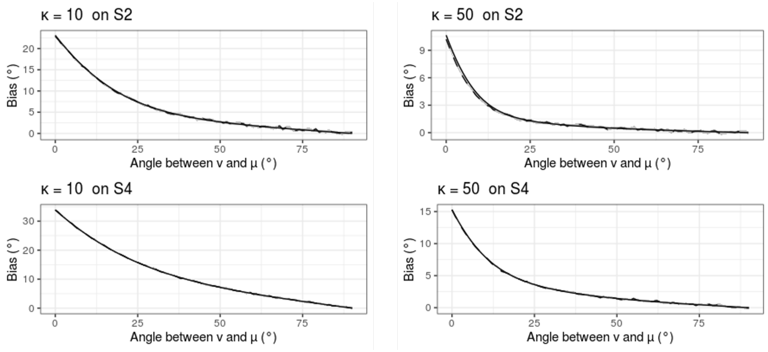


Figure 3.4: The biases calculated from (3.8) are drawn in solid lines. The biases from Monte Carlo simulations are in dashed lines. The confidence intervals for Monte Carlo simulations are in gray-shaded areas.

Fix $v = (0, \ldots, 0, 1)^T \in S^d$ and $r \in [0, \pi/2]$. Given a fixed vector $(t_1, \ldots, t_{d-1})^T \in [0, 2\pi]^{d-1}$, we generate spherical angles $(\psi_1, \ldots, \psi_{d-1})^T$ with each component $\psi_i \sim \text{Unif}([-0.5 \cdot t_i, 0.5 \cdot t_i])$. Then the signal point $x' \in S^d$ is obtained by the coordinate change,

$$x' = ((\prod_{i=1}^{d-1} \sin\psi_i) \sin r, \cos\psi_1 (\prod_{i=2}^{d-1} \sin\psi_i) \sin r, \ldots, \cos\psi_{d-1} \sin r, \cos r)^T.$$

Let the point x follow von Mises-Fisher distribution with a pole x' and a concentration parameter κ. We generate n signal points $x_1, \ldots, x_n$ identically and independently as the same law as x. For figurative examples, *see* Figure 3.5.

Note that $\theta_v = \arccos(v^T x)$ is not independent from $\theta_i = \arccos(e_i^T x)$, where e_i is the ith coordinate vector with $i = 1, \ldots, d$.

We generate 100 sets of n data points for given parameters. We then compute the average and standard deviation of estimated ra-

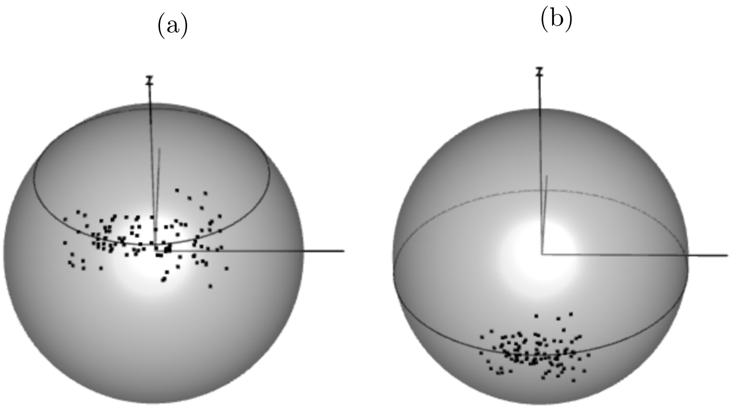


Figure 3.5: Examples of generic data generation ($n = 100$) (a) $r = 45°$, $\kappa = 100$, $t_1 = 30°$, (b) $r = 90°$, $\kappa = 100$, $t_1 = 10°$.

dius for L_1, L_2 penalties and MCP. In validation, we used 5-cross validation on a grid of the tuning parameter τ, $(0, 10^{-3}, 10^{-2.9}, \ldots, 10^{-0.1}, 10^{0})$.

For the case where the true radius $r = 90°$, we additionally record the number of data sets that mitigate type 1 overfitting phenomenon, that is, when the estimated radius $\widehat{r} = 90°$. For every dataset, we compute the mean and median of the angles between the true axis $v = (0, \ldots, 0, 1)^T$ and data points x_i's, what we call oracle mean and median of radius.

3.6.2 Results

We first consider the case where $d = 2$ and $n = 100$. We assess the performance at the combination of settings, $r = 90°, 45°$, $\kappa = 100, 30$ and $t_1 = 90°, 60°$.

We start with the results when the true radius r is 90°. Here we want to see if the penalization of r works well in mitigating type 1 overfitting phenomenon.

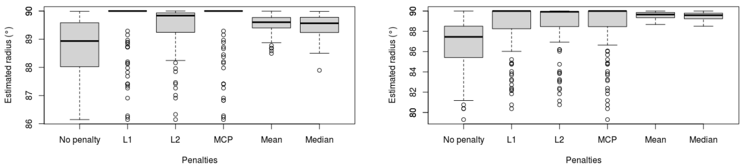
We set $\kappa = 100$. When $t_1 = 90^\circ$, we observe that the estimated radius has median at 90° for L_1 penalty and MCP, *see* Figure 3.6a. In 76 of L_1 penalty and 78 of MCP penalty out of 100 data sets, type 1 overfitting was mitigated, *see* Table 3.1. No other penalties (original PNS and L_2 penalty) could not mitigate type 1 overfitting phenomenon. When $t_1 = 60^\circ$, 74 of L_1 penalty and 74 of MCP out of 100 data sets achieved $\widehat{r} = 90^\circ$, *see* Figure 3.6b and Table 3.1.

The pattern of estimated radius is quite different between $t_1 = 90^\circ$ and 60° cases. When $t_1 = 60^\circ$, as data points spans only 1/3 of the true arc (equator of S^2), the estimation of axis, $\widehat{v}$, suffers larger variance compared to $t_1 = 90^\circ$. Indeed, we can observe the variances of estimated radius, $\widehat{r}$, is much larger when $t_i = 60^\circ$, *see* Figure 3.7. However, even though larger variance in estimating v and r, the rate of overfitting mitigation is quite similar in both $t_1 = 90^\circ$ and 60° cases, *see* Table 3.1.

Next, for $\kappa = 30$, where von Mises-Fisher distribution has a larger variance. When $t_1 = 90^\circ$, 71 of L_1 penalty and 73 of MCP penalty out of 100 data sets mitigated type 1 overfitting, *see* Figure 3.6c and Table 3.1. When $t_1 = 60^\circ$, 75 of L_1 penalty and 76 of MCP penalty out of 100 data sets mitigated type 1 overfitting, *see* Figure 3.6b and Table 3.1. We observe larger variances in radius estimations than $\kappa = 100$ cases. Even though the instability in axis estimation, the rate of overfitting mitigation showed no significant difference from the case $\kappa = 100$.

For a larger data set size, $n = 1000$ with the same setting, $r = 90^\circ$, $\kappa = 100, 30$ and $t_1 = 90^\circ, 60^\circ$. The rate of overfitting

(a) $r = 90°$, $t_1 = 90°$, $\kappa = 100$ (b) $r = 90°$, $t_1 = 60°$, $\kappa = 100$



(c) $r = 90°$, $t_1 = 90°$, $\kappa = 30$ (d) $r = 90°$, $t_1 = 60°$, $\kappa = 30$

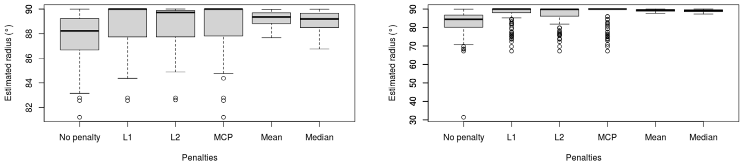


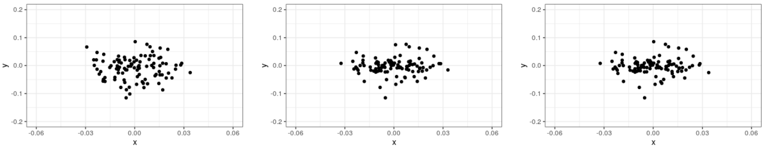
Figure 3.6: Results of radius estimation when $d = 2$, $r = 90°$ and $n = 100$.

mitigation improved compared to $n = 100$ cases, *see* Figure B.2.1 and Table 3.1. The mean and standard deviation of estimated radius are improved significantly.

When $r = 45°$, there is no danger of type 1 overgfitting. L_1 and L_2 penalties shows significantly large bias in estimating radius, *see* Figure 3.8 and Table B.2.1, whereas MCP shows the identical results as original PNS. L_1 and L_2 penalties also shows larger variance than original PNS and MCP due to the instability in axis estimation.

To present the results on type 2 overfitting phenomenon, we consider two situtation, (1) $r = 5°$ with $t_1 = 90°$, $\kappa = 100$, $n = 100$ and (2) $r = 90°$ with $t_1 = 10°$, $\kappa = 100$, $n = 100$. The first case mimics a disc-shaped distribution (high concentrated unimodal

(a) $t_1 = 90°$, PNS (b) $t_1 = 90°$, L_1 (c) $t_1 = 90°$, MCP



(d) $t_1 = 60°$, PNS (e) $t_1 = 60°$, L_1 (f) $t_1 = 60°$, MCP

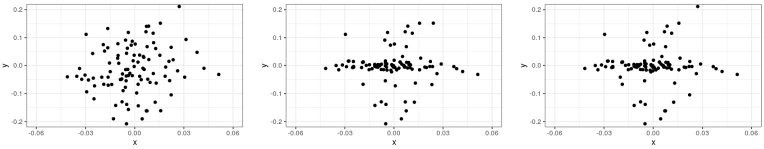
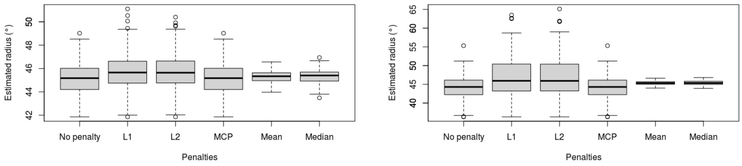


Figure 3.7: Results of axis estimation when $d = 2$, $r = 90°$ and $n = 100$. The estimated axis $\widehat{v}$ are mapped to the tangent plane at $(0, 0, 1)^T$ by the exponential map $\text{Exp}_{(0,0,1)^T}$.

(a) $r = 45°$, $t_1 = 90°$, $\kappa = 100$, $n = 100$ (b) $r = 45°$, $t_1 = 60°$, $\kappa = 100$, $n = 100$



(c) $r = 45°$, $t_1 = 90°$, $\kappa = 30$, $n = 100$ (d) $r = 45°$, $t_1 = 90°$, $\kappa = 100$, $n = 1000$

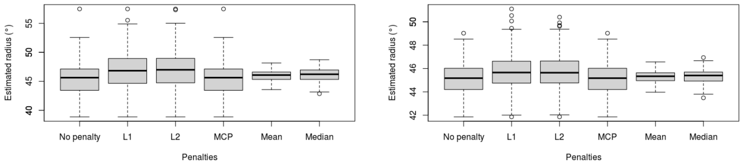


Figure 3.8: Results of radius estimation when $d = 2$, $r = 45°$.

		Original PNS	L_1	L_2	MCP	Mean (Oracle)	Median (Oracle)
$n = 100$							
$r = 90^\circ$, $t_1 = 90^\circ$,	mean (sd)	88.72 (1.04)	89.42 (1.00)	89.42 (0.87)	89.51 (1.03)	89.97 (0.58)	89.93 (0.66)
$\kappa = 100$	$\widehat{r} = \pi/2$	0	**76**	0	**78**	0	0
$r = 90^\circ$, $t_1 = 60^\circ$,	mean (sd)	88.67 (2.49)	88.62 (2.62)	88.65 (2.39)	88.57 (2.76)	90.05 (0.52)	90.10 (0.59)
$\kappa = 100$	$\widehat{r} = \pi/2$	0	**74**	0	**74**	0	0
$r = 90^\circ$, $t_1 = 90^\circ$,	mean (sd)	87.75 (1.86)	88.97 (1.82)	88.83 (1.65)	88.95 (1.96)	90.05 (0.96)	90.13 (1.22)
$\kappa = 30$	$\widehat{r} = \pi/2$	0	**71**	0	**73**	0	0
$r = 90^\circ$, $t_1 = 60^\circ$,	mean (sd)	82.31 (7.46)	87.12 (5.51)	86.95 (5.29)	87.07 (5.80)	90.09 (0.94)	90.18 (1.10)
$\kappa = 30$	$\widehat{r} = \pi/2$	0	**75**	0	**76**	0	0
$n = 1000$							
$r = 90^\circ$, $t_1 = 90^\circ$,	mean (sd)	89.96 (0.24)	89.92 (0.20)	89.88 (0.18)	89.92 (0.22)	90.01 (0.17)	90.00 (0.22)
$\kappa = 100$	$\widehat{r} = \pi/2$	0	**82**	0	**84**	0	0
$r = 90^\circ$, $t_1 = 60^\circ$,	mean (sd)	88.91 (0.80)	89.67 (0.76)	89.63 (0.67)	89.68 (0.79)	90.00 (0.16)	90.00 (0.22)
$\kappa = 100$	$\widehat{r} = \pi/2$	0	**81**	0	**83**	0	0
$r = 90^\circ$, $t_1 = 90^\circ$,	mean (sd)	89.27 (0.48)	89.78 (0.47)	89.72 (0.40)	89.77 (0.52)	90.02 (0.32)	90.01 (0.41)
$\kappa = 30$	$\widehat{r} = \pi/2$	0	**79**	0	**80**	0	0
$r = 90^\circ$, $t_1 = 60^\circ$,	mean (sd)	87.06 (1.45)	89.15 (1.68)	89.08 (1.63)	89.06 (1.93)	90.00 (0.31)	89.99 (0.40)
$\kappa = 30$	$\widehat{r} = \pi/2$	0	**75**	0	**79**	0	0

Table 3.1: Results of radius estimation when $d = 2$ and $r = 90^\circ$.

distribution) (Figure 3.2 left) and the second case represents a situation where the data points are generated along a great sphere but on a short interval (Figure 3.1b).

In both cases, IoD criterion successfully mitigated type 2 overfitting phenomenon. In the first case ($r = 5^\circ$), only L_1 penalty with IoD and MCP with IoD estimated radius as 90° with rates 91 and 92 out of 100 data sets. In the second case ($r = 90^\circ$), L_1 penalty and MCP mitigated overfitting only 5 and 4 times respectivelym but L_1 penalty with IoD and MCP with IoD achieved 94 and 99 times out of 100 data sets. The original PNS only yielded overfitted results.

3.7 Real Data Analysis

In this section, we apply the proposed PenPNS to a dataset called the Cephalometric X-ray Image Data, a landmark dataset on the sagittal-view X-ray images of patients' skulls (Wang et al.,2016). We choose to use 400 adolescent cases with 19 landmarks on 2-D dental X-ray images, that is, the dataset has dimension $19 \times 2 \times 400$. The dataset was preprocessed by the Generalized Procrustes Analysis method (GPA) (Mardia and Jupp,1999; Chapter 3 of Dryden and Mardia,2016), which results in 400 points on S^{36}.

We applied PNS, PenPNS (MCP), Principal Geodesic Nested Spheres (PNGS) (Jung,2012) and PenPNS (MCP) with IoD criterion (for short, PenPNS+IoD)on the Cephalometric X-ray Image Data. Here, PNGS is a variant of PNS, such that radius r is fixed as $\pi/2$ while estimating parameters of each subsphere A_m, thus the data points are forced to be estimated by great sphere at every step.

We want to compare the results of the methods with respect to variance explained (Figure B.3.1 and B.3.2) and estimated radii (Figure 3.10). For overall results, *see* Table B.3.1 - B.3.4 in the supplementary materials.

Variance explained is obtained as follows. We use the residual $\Xi(m)$ of $\mathfrak{A}_m$ after fitting $\mathfrak{A}_m$ inside $\mathfrak{A}_{m+1}$ (they are indeed analogues of the $(m+1)$th PC of classical PCA in Euclidean space). The variance s_m^2 of the mth component is the squared sum of $\Xi(m)$. The percent of variance s_m^2 is the proportion of s_m^2 over the sum of all variances. For more details, *see* Section 2.4 of (Jung,2012).

PenPNS yielded 66.53% in the first component of variance proportions. This figure is slightly lower than the one of PNS, 68.49%, *see* Figure B.3.1 and Table B.3.1 and B.3.2 in the supplementary materials.

However, two methods show a different implication in the estimated radii. The estimated radii of PNS vary from 15.47° to 90° (Table B.3.1). In contrast, PenPNS yielded mostly 90° except 8.86° of A_3 and 22.23° of A_1 (Table B.3.2), which means PenPNS mitigated type 1 overfitting phenomenon in nearly every dimension. It is notable that type 1 overfitting mitigations in high dimensions show little influence on variance proportions in lower dimensions.

In constrast, PNGS and PenPNS+IoD featured 39.41% and 40.79% in the first component of variance proportions, *see* Figure B.3.2 and Table B.3.3 and B.3.4 in the supplementary materials. It is notable that every estimated radii of PenPNS+IoD is 90°

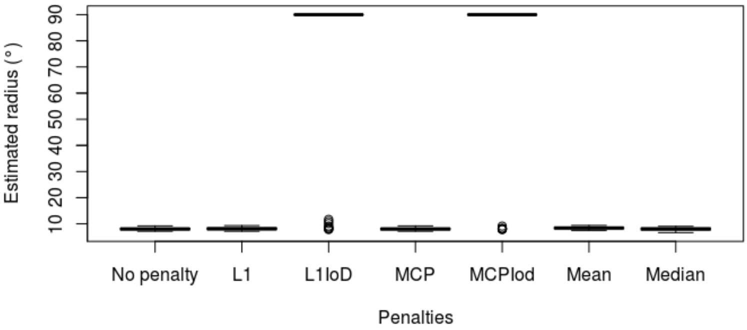
(Figure 3.10) and PenPNS+IoD yielded nearly identical results as PNGS (Table B.3.3 and B.3.4).

One of the key observations here is that PenPNS+IoD mitigated type 2 overfitting at A_3 (8.86° to 90°), in addition to type 1 overfitting mitigation by PenPNS. It may seem a small correction in radius estimation, but resulted in large differences in variance proportions between two methods.

We interpret the results as follows. By mitigating type 2 overfitting at A_3, PenPNS+IoD corrected overfitting tendency shown in the original PNS and PenPNS (66.53 and 68.59% to 40.79% in the first component of variance proportions). It also corrected slight underfitting in PNGS (39.41% to 40.79%).

We visualize the results by drawing the principal mode of variation (call PC1 and PC2 from $\mathfrak{A}_1$ and $\mathfrak{A}_2$ respectively) on the data points, *see* Figure 3.11 and 3.12. We observe that PC1 of the original PNS and PenPNS make small circles in data point clouds, which implies type 2 overfitting phenomenon. In contrast, PC1 of PenPNS+IoD and PNGS draw gentle arcs through data point clouds, which implies the mitigation of type 2 overfitting.

(a) $r = 5°$, $t_1 = 90°$, $\kappa = 100$, $n = 100$



(b) $r = 90°$, $t_1 = 10°$, $\kappa = 100$, $n = 100$

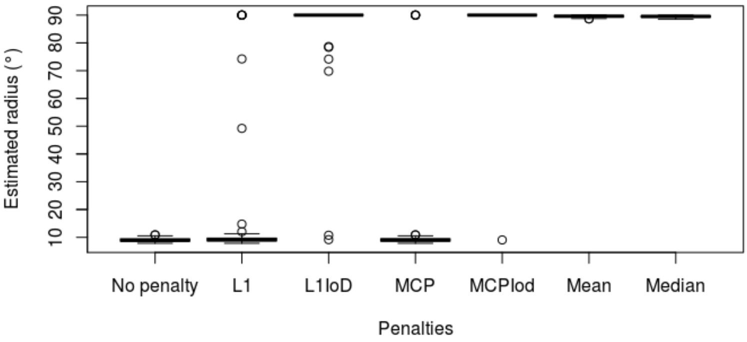


Figure 3.9: Results on mitigating type 2 overfitting phenomenon.

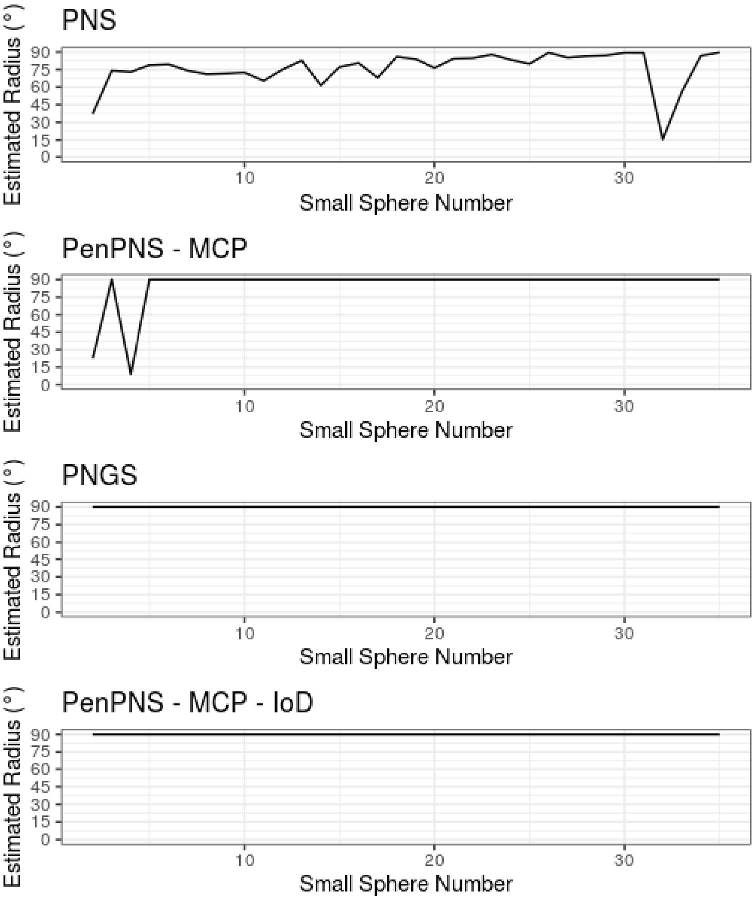
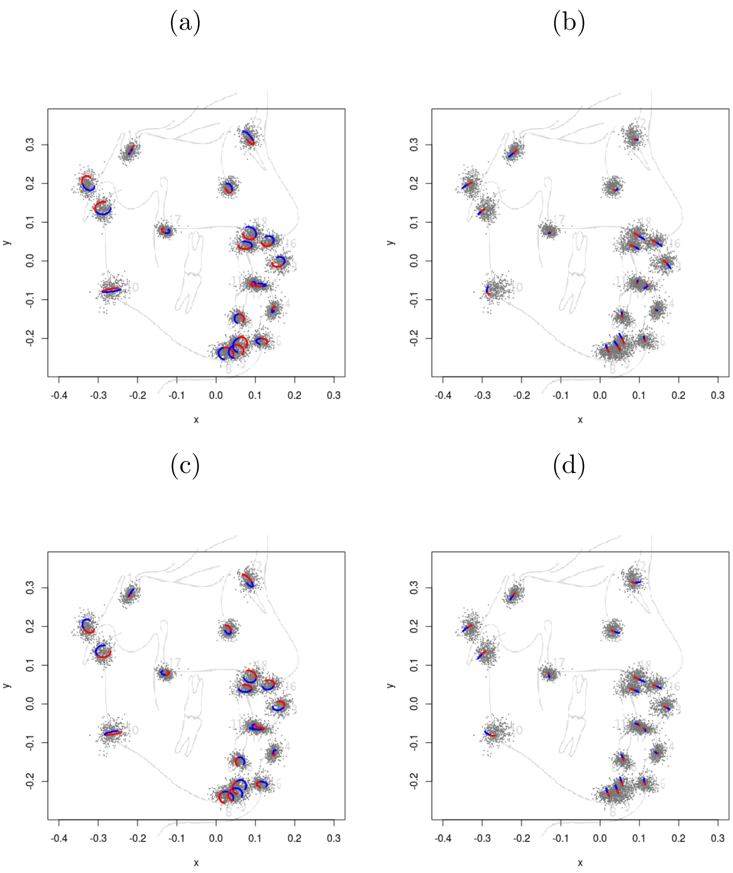
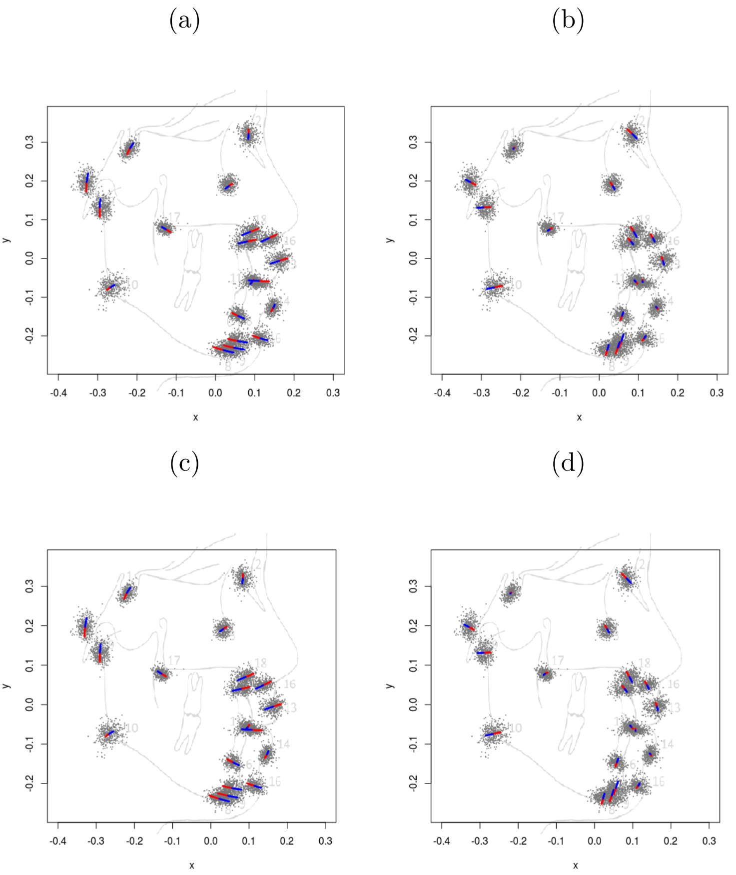


Figure 3.10: Cephalometric X-ray Image Data : Estimated radii $\widehat{r}_m$ of small sphere estimators $A_m(\widehat{v}_k, \widehat{r}_m)$ for PNS, PenPNS (MCP), PNGS and PenPNS (MCP) + IoD Criterion.


(a) (b)

(c) (d)

Figure 3.11: Cephalometric X-ray Image Data : The principal mode of variation by $\mathfrak{A}_1$ (left column) and $\mathfrak{A}_2$ (right column). (a)(b) PNS, (c)(d) PenPNS (MCP).


(a) (b)

(c) (d)

Figure 3.12: Cephalometric X-ray Image Data : The principal mode of variation by $\mathfrak{A}_1$ (left column) and $\mathfrak{A}_2$ (right column). (a)(b) PNGS, (c)(d) PenPNS (MCP) + IoD criterion.

Appendix A

Supplementary Materials for Chapter 2

A.1 Mathematical Backgrounds of Section 2.3.1

A.1.1 Principal Angles

There had been suggested several definitions of the distance between two linear subspaces $[A]$ and $[B]$ in $\mathbb{R}^n$, see Table 2 and Theorem 12 of Ye and Lim (2016). Indeed, these distances can be expressed by principal angles between two linear subspaces.

In this section, we present the concept of *principal angles* (Bj orck and Golub,1973). Principal angles between two linear subspaces are defined as follows: Given $[A] \in \mathrm{Gr}(n, p_1)$ and $[B] \in \mathrm{Gr}(n, p_2)$, let $p = \min(p_1, p_2)$. The ith principal vectors (a_i, b_i), $i = 1, \ldots, p$

are defined as a pair of unit vectors that satisfies

$$a_i^T b_i = \max_{v \in [A]} \max_{u \in [B]} v^T u$$

with the condition $a_i \perp a_1, \ldots, a_{i-1}$ and $b_i \perp b_1, \ldots, b_{i-1}$. Principal angles are then a set of angles $\theta = (\theta_1, \ldots, \theta_p)^T$ such that

$$\cos \theta_i = a_i^T b_i$$

for $i = 1, \ldots, p$. We call θ_i as ith principal angle between $[A]$ and $[B]$. Note that $0 \leq \theta_1 \leq \ldots, \leq \theta_p \leq \pi/2$.

The computation of principal angles can be performed using SVD. Suppose $[A]$ and $[B]$ are represented by $A \in \mathrm{V}(n, p_1)$ and $B \in \mathrm{V}(n, p_2)$ respectively. By the untruncated SVD, we decompose $A^T B$ into

$$A^T B = U \Sigma V^T,$$

where $U \in \mathbb{R}^{p_1 \times p_1}$ and $V \in \mathbb{R}^{p_2 \times p_2}$ are orthonormal and $\Sigma \in \mathbb{R}^{p_1 \times p_2}$ is a diagonal matrix of a rectangular shape,

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}$$

with a diagonal $\Sigma_1 \in \mathbb{R}^{p \times p}$. The diagonal elements of Σ_1 are written $\sigma_1 \geq \ldots \geq \sigma_p \geq 0$ in order and the principal angles between $[A]$ and $[B]$ can be easily calculated as

$$\sigma_i = \cos \theta_i$$

for $i = 1, \ldots, p$.

We show an example of subspace distance expressed using principal angles: the Frobenius norm distance, appeared in Section 1.1

and 2.3.1, can be expressed as

$$
\begin{aligned}
d([A],[B]) &= \frac{1}{\sqrt{2}}\|AA^T - BB^T\|^2 \\
&= \sqrt{2}(n - \mathrm{Tr}(A^T AB^T B)) \\
&= \sqrt{2}(n - (\cos^2\theta_1 + \ldots + \cos^2\theta_n)) \\
&= \sqrt{2}\sum_{i=1}^{p} \sin^2\theta_i \\
&= \sqrt{2}\|\sin\theta\|_2^2,
\end{aligned}
$$

where A and B are orthonormal matrices which represent $[A] \in \mathrm{Gr}(n, p_1)$ and $[B] \in \mathrm{Gr}(n, p_2)$ and $p = \min(p_1, p_2)$.

A.1.2 Calculation of Mean Direction

We wish to explain the concept of mean direction in perspective of the flag mean. The mean direction among a set of linear subspaces $[V_1], \ldots, [V_K]$ is defined as a one-dimensional subspace $v_{\{[V_i]\}_{i=1}^K}$ that minimize the sum of the squares of the subspace distances between $v_{\{[V_i]\}_{i=1}^K}$ and each linear subspace $[V_i]$. That is,

$$
v_{\{[V_i]\}_{i=1}^K} = \underset{[v]\in \mathrm{Gr}(n,1)}{\arg\min} \sum_{i=1}^{K} \mathrm{d}\left([v],[V_i]\right)^2, \tag{A.1}
$$

where $\mathrm{d}([A],[B])$ is a subspace distance between subspaces $[A]$ and $[B]$.

In this definition, we adopt the Frobenius norm distance as a measure of distance between two linear subspaces. The optimization problem (A.1) can be rewritten as follows (we quoted the derivations directly from Draper et al. (2014):

$$
\begin{aligned}
v_{\{[V_i]\}_{i=1}^K} &= \underset{[v]\in \mathrm{Gr}(n,1)}{\arg\min} \sum_{i=1}^{K} \mathrm{d}_{pF}\left([v],[V_i]\right)^2 \\
&= \underset{[v]\in \mathrm{Gr}(n,1)}{\arg\min} \sum_{i=1}^{K} \|\sin\theta_i\|^2 \\
&= \underset{[v]\in \mathrm{Gr}(n,1)}{\arg\max} \sum_{i=1}^{K} \|\cos\theta_i\|^2 \\
&= \underset{[v]\in \mathrm{Gr}(n,1)}{\arg\max} \sum_{i=1}^{K} \cos^2\theta_i \\
&= \underset{[v]\in \mathrm{Gr}(n,1)}{\arg\max} \sum_{i=1}^{K} v^T V_i V_i^T v \\
&= \underset{[v]\in \mathrm{Gr}(n,1)}{\arg\max}\; v^T \left(\sum_{i=1}^{K} V_i V_i^T\right) v.
\end{aligned}
$$

There is a merit of using Frobenius norm distance we can see in the fifth equality, in that there is a closed form converting an expression of the principal angle θ into of v and V_is, or, $\cos^2\theta_i = v^T V_i V_i^T v$. Other distances, such as the Procrustes distance $2\|\sin^2\frac{1}{2}\theta\|$ or the geodesic distance $\|\theta\|_2$, are not suitable for making this kind of linear expression using v and V_is.

Another merit comes from the last equality that this optimization problem is equivalent to the variational characterization of SVD (Draper et al.,2014). The solution is indeed the first singular vector of the symmetric matrix

$$
\left(\sum_{i=1}^{K} V_i V_i^T\right).
$$

This sum of matrices term can be rewritten with respect to a concatenated matrix. Let V be a matrix that is a column concatenation of V_is,

$$V = [V_1; V_2; \ldots; V_K] \in \mathbb{R}^{n \times \sum r_i}.$$

It is easy to show that $VV^T = \sum_{i=1}^K V_i V_i^T$. Therefore, the principal direction among $V_1, \ldots, V_K$ can be conveniently computed using SVD. The SVD of V is expressed as

$$V = Q \Sigma W^T$$

and subsequently

$$VV^T = Q \Sigma \Sigma^T Q^T.$$

Note that the rank of VV^T is the same as that of V. The first column of Q is the principal direction $v_{\{V_i\}_{i=1}^K}$ we are looking for. In practice, it can be computed efficiently using truncated SVD numerical solutions.

A.2 Proofs

A.2.1 Some Basic Facts for Proofs

As it is the most elementary fact, we begin by recapitulating Lemma 2.2 and show its proof.

Lemma 2.2. *For* $i, j \in \mathcal{I}_K$ *and* $S_i \cap S_j \neq \phi$, $[W_i] \perp [W_j]$.

Proof. From the definition of the partially-joint score subspace, $[W_i]$, we immediately find that

$$[W_i] \perp [W_j], \quad i, j \in \mathcal{I}_K \text{ and } S_j \cap S_i \neq \phi$$

by the range-kernel complementarity property of the vector space projection transformation. □

We denote $\mathcal{I}_{<i} = \{j : j < i, S_j \cap S_i \neq \phi\}$. Hereafter $\mathcal{N}(T)$ is the null space of a linear transformation T of $\mathbb{R}^n$ and $\mathcal{R}(T)$ its range space.

Lemma A.1. *For* $i \in \mathcal{I}_K$, *we have* $\mathcal{N}(\bigcirc_{k\in\mathcal{I}_{<i}} P_k^{\perp}) = \oplus_{k\in\mathcal{I}_{<i}}[W_k]$.

Proof. Let $v \in \oplus_{k\in\mathcal{I}_{<i}}[W_k]$. Then there exists a unique $\{v_k\}_{k\in\mathcal{I}_{<i}}$ with $v_k \in [W_k]$ such that the sum of all v_k is v. For each $j_1 \in \mathcal{I}_{<i}$ and $v_{j_1} \in [W_{j_1}]$, it can be easily checked that $(\bigcirc_{k\in\mathcal{I}_{<j_1}} P_k^{\perp})(v_{j_1}) = \{0\}$ and then $(\bigcirc_{k\in\mathcal{I}_{<i}} P_k^{\perp})(v_{j_1}) = \{0\}$ follows. Thus $(\bigcirc_{k\in\mathcal{I}_{<i}} P_k^{\perp})(v) = \{0\}$ and $\mathcal{N}(\bigcirc_{k\in\mathcal{I}_{<i}} P_k^{\perp}) \supset \oplus_{k\in\mathcal{I}_{<i}}[W_k]$.

Conversely, let $v' \notin \oplus_{k\in\mathcal{I}_{<i}}[W_k]$. Then there exists a unique $\{v_k\}_{k\in\mathcal{I}_{<i}}$ with $v_k \in [W_k]$ and non-zero $a \in (\oplus_{k\in\mathcal{I}_{<i}}[W_k])^{\perp}$ such that v' is the sum of all v_k and a. For each $j_1 \in \mathcal{I}_{<i}$ and $v_{j_1} \in [W_{j_1}]$, we have $(\bigcirc_{k\in\mathcal{I}_{<i}} P_k^{\perp})(v_{j_1}) = \{0\}$ as before. Since $a \perp [W_{k''}]$ for

any $k'' \in \mathcal{I}_{<i}$, we also have $(\bigcirc_{k\in\mathcal{I}_{<i}} P_k^\perp)(a) \neq 0$. Thus we have $(\bigcirc_{k\in\mathcal{I}_{<i}} P_k^\perp)(v') \neq \{0\}$ and $\mathcal{N}(\bigcirc_{k\in\mathcal{I}_{<i}} P_k^\perp) \subset \oplus_{k\in\mathcal{I}_{<i}}[W_k]$. ☐

Lemma A.2. *For $i \in \mathcal{I}_K$, we have $\oplus_{k\in\mathcal{I}_{<i}}[W_k] = +_{k\in\mathcal{I}_{<i}}(\cap_{k'\in S_k}[V_{k'}])$.*

Proof. We give a proof by induction on k. If $k = 1$, there is nothing to prove. If $k = 2$, $[W_1] = \cap_{k'\in S_1}[V_{k'}]$, so the statement is true for $k = 1, 2$.

For any $k \geq 3$, suppose the statement holds, that is,

$$\oplus_{j\in\mathcal{I}_{<m}}[W_j] = +_{j\in\mathcal{I}_{<m}}(\cap_{k'\in S_j}[V_{k'}])$$

for all $1 \leq m \leq k$. Let $\tilde{k}$ be the largest element in $\mathcal{I}_{<k+1}$. We denote $P = \bigcirc_{k'\in\mathcal{I}_{<\tilde{k}}} P_{k'}^\perp$ and $P^\perp$, the projection onto $\mathcal{N}(P)$ of $\mathbb{R}^n$. Then, we have

$$\begin{aligned} +_{j\in\mathcal{I}_{<k+1}}(\cap_{k'\in S_j}[V_{k'}]) &= \oplus_{j\in\mathcal{I}_{<\tilde{k}}}[W_j] + \cap_{k'\in S_{\tilde{k}}}[V_{k'}] \\ &= \oplus_{j\in\mathcal{I}_{<\tilde{k}}}[W_j] + P(\cap_{k'\in S_{\tilde{k}}}[V_{k'}]) + P^\perp(\cap_{k'\in S_{\tilde{k}}}[V_{k'}]) \end{aligned}$$

Indeed, $P^\perp(\cap_{k'\in S_{\tilde{k}}}[V_{k'}]) \subset \mathcal{N}(P)$ and $\mathcal{N}(P) = \oplus_{k'\in\mathcal{I}_{<\tilde{k}}}[W_{k'}]$ by the previous lemma. Thus

$$\begin{aligned} +_{j\in\mathcal{I}_{<k+1}}(\cap_{k'\in S_j}[V_{k'}]) &= \oplus_{j\in\mathcal{I}_{<\tilde{k}}}[W_j] + P(\cap_{k'\in S_{\tilde{k}}}[V_{k'}]) \\ &= \oplus_{j\in\mathcal{I}_{<\tilde{k}}}[W_j] \oplus [W_{\tilde{k}}] \\ &= \oplus_{j\in\mathcal{I}_{<k+1}}[W_j]. \end{aligned}$$

Therefore the statement holds for any $k \geq 3$ and the proof is completed. ☐

Lemma A.3. *For $k = 1, \ldots, K$, $+_{i\in\{k\in S_i\}}[W_{S_i}] = [V_k]$.*

Proof. In Lemma A.2, we set i such that $S_i = \{k\}$. Then $\mathcal{I}_{<i} = \{j : k \in S_j, |S_j| > 1\}$ and $\oplus_{j\in\mathcal{I}_{<i}}[W_j] = +_{j\in\mathcal{I}_{<i}}(\cap_{k'\in S_j}[V_{k'}])$. The proof goes similarly as in Lemma A.2. Let $\tilde{i}$ be the largest element in $\mathcal{I}_{<i}$. We denote $P = \bigcirc_{k'\in\mathcal{I}_{<\tilde{i}}} P_{k'}^{\perp}$ and $P^{\perp}$, the projection onto $\mathcal{N}(P)$ of $\mathbb{R}^n$.

$$\begin{aligned}
+_{j\in\mathcal{I}_{<i}\cup S_i}(\cap_{k'\in S_j}[V_{k'}]) &= \oplus_{j\in\mathcal{I}_{<i}}[W_j] + \cap_{k'\in S_i}[V_{k'}] \\
&= \oplus_{j\in\mathcal{I}_{<i}}[W_j] + P(\cap_{k'\in S_i}[V_{k'}]) + P^{\perp}(\cap_{k'\in S_i}[V_{k'}]) \\
&= \oplus_{j\in\mathcal{I}_{<i}}[W_j] + P(\cap_{k'\in S_i}[V_{k'}]) \\
&= \oplus_{j\in\mathcal{I}_{<i}}[W_j] \oplus [W_i] \\
&= \oplus_{j\in\mathcal{I}_{<i}\cup S_i}[W_j].
\end{aligned}$$

By the set inclusion-exclusion principle, as $\mathcal{I}_{<i} \cup S_i$ involves every index-set that contains k, it is immediate that $+_{j\in\mathcal{I}_{<i}\cup S_i}(\cap_{k'\in S_j}[V_{k'}]) = [V_k]$. □

A.2.2 Proof of Theorem 2.1

Proof. We claim that rank($[W_i]$) is uniquely determined. By Sylvester's law of nullity, we have

$$\text{rank}([W_i]) = \text{rank}(\cap_{k\in S_i}[V_k]) - \text{rank}(\mathcal{N}(\bigcirc_{j\in\mathcal{I}_{<i}} P_j^{\perp}) \cap (\cap_{k\in S_i}[V_k])),$$

where $\mathcal{I}_{<i} = \{j : j < i, S_j \cap S_i \neq \phi\}$ as before. By lemma A.1 and A.2, Sylvester's law of nullity for our theorem is restated as

$$\text{rank}([W_i]) = \text{rank}(\cap_{k\in S_i}[V_k]) - \text{rank}\left((+_{k\in\mathcal{I}_{<i}}(\cap_{k'\in S_k}[V_{k'}])) \cap (\cap_{k\in S_i}[V_k])\right).$$

We want to make this expression in a more explicit form. For that, we suggest the following assertions.

Let $l = |S_i|$. We set $[I_l] = +_{k \in \{k : |S_k| > l\}}(\cap_{k' \in S_k}[V_{k'}])$. The projection onto $[I_l]$ in $\mathbb{R}^n$ is denoted by P_{I_l} and the projection onto $[I_l]^\perp$ is denoted by $P_{I_l}^\perp$. Moreover, we consider an index set $\mathcal{J}_{i,>l} = \mathcal{I}_{<i} \cap \{k : |S_k| > l\}$ and let $[J_i] = +_{k \in \mathcal{J}_{i,>l}}(\cap_{k' \in S_k}[V_{k'}])$. The projection onto $[J_i]$ in $\mathbb{R}^n$ is denoted by P_{J_i} and the projection onto $[J_i]^\perp$ by $P_{J_i}^\perp$. Finally, we define $\mathcal{J}_{i,l} = \mathcal{I}_{<i} \cap \{k : |S_k| = l\}$.

Lemma A.4. $P_{I_l}^\perp = P_{I_l}^\perp \circ P_{J_i}^\perp$.

Proof. Trivial from the fact $\mathcal{N}(P_{J_i}^\perp) \subset \mathcal{N}(P_{I_l}^\perp)$. □

Lemma A.5. *If* $v_j \in P_{J_i}^\perp(\cap_{k' \in S_j}[V_{k'}])$ *with* $j \in \mathcal{J}_{i,l} \setminus \mathcal{J}'_{i,l}$, *then* $v_j \in (\cap_{k' \in S_j}[V_{k'}]) \cap (\cap_{k' \in S_m}[V_{k'}])$ *for some* $m < i$, $m \neq j$ *such that* $S_m \cap S_i = \{0\}$ *and not for any* $m < i$, $m \neq j$ *such that* $S_m \cap S_i \neq \{0\}$.

Proof. Consider the cases

(1) $v_j \in (\cap_{k' \in S_j}[V_{k'}]) \cap (\cap_{k' \in S_m}[V_{k'}])$ for some $m < i$, $m \neq j$ such that $S_m \cap S_i \neq \{0\}$ but not for any $S_m \cap S_i = \{0\}$,

(2) $(\cap_{k' \in S_j}[V_{k'}]) \cap (\cap_{k' \in S_m}[V_{k'}]) = \{0\}$ for only $m \in \mathcal{J}_{i,l}$

In case (1), v_j becomes automatically an element of $\cap_{k' \in S_t}[V_{k'}]$ such that $S_j \subset S_t$ and $S_m \subset S_t$. Then $v_j \in [I_i]$ since $S_i \cap S_t \neq \phi$ and $|S_t| > |S_i|$, and this is a contradiction. In case (2), $v_j \notin \mathcal{R}(P_{I_l})$ and this is contradict to the assumption. □

Proposition A.1. *If* $[V_k]_{k \in \mathcal{K}}$ *is relatively independent, then* $+_{k \in \mathcal{J}_{i,l}} P_{J_i}^\perp(\cap_{k' \in S_k}[V_{k'}])$ *and* $P_{J_i}^\perp(\cap_{k \in S_i}[V_k])$ *are independent.*

Proof. We want to show that the relative independence of $[V_k]_{k \in \mathcal{K}}$ is violated if $+_{k \in \mathcal{J}_{i,l}} P_{J_i}^\perp(\cap_{k' \in S_k}[V_{k'}])$ and $P_{J_i}^\perp(\cap_{k \in S_i}[V_k])$ are linearly dependent. Suppose that there is nonzero $v \in +_{k\in\mathcal{J}_{i,l}} P_{J_i}^{\perp}(\cap_{k'\in S_k}[V_{k'}]) \cap P_{J_i}^{\perp}(\cap_{k\in S_i}[V_k])$ with $l = |S_i|$. We are going to show that $+_{k\in\mathcal{J}_{i,l}} P_{J_i}^{\perp}(\cap_{k'\in S_k}[V_{k'}]) \cap P_{J_i}^{\perp}(\cap_{k\in S_i}[V_k])$ is not $\{0\}$ in the following cases.

First we exclude the case where $P_{I_l}^{\perp}(\cap_{k'\in S_i}[V_{k'}]) = \{0\}$. If so we have $\cap_{k'\in S_i}[V_{k'}] = \oplus_{k''\in\{k'':S_i\subset S_{k''}\}}[W_k''] = +_{k''\in\{k'':S_i\subset S_{k''}\}}[V_k'']$, then $\cap_{k'\in S_i}[V_{k'}]$ itself is in $[I_i] = \mathcal{N}(P_{J_i}^{\perp}))$.

Next, under the assumption that $P_{I_l}^{\perp}(\cap_{k'\in S_i}[V_{k'}]) \neq \{0\}$, we run through the following situations. Now on $\mathcal{J}_{i,l}' = \{k : |S_k| = l, k \in \mathcal{I}_{<i}, P_{I_i}^{\perp}(S_k) \neq \{0\}\}$.

(i) $|\mathcal{J}_{i,l}| = 0$: There is nothing to prove.

(ii) $|\mathcal{J}_{i,l}| = 1$ and $|\mathcal{J}_{i,l}'| = 0$:

We will show that this case is vacuous. Suppose there exist nonzero $v \in P_{J_i}^{\perp}(\cap_{k'\in S_i}[V_{k'}]) \cap P_{J_i}^{\perp}(\cap_{k'\in S_k}[V_{k'}])$ with $k \in \mathcal{J}_{i,l}$. As $P_{I_l}^{\perp}(\cap_{k'\in S_k}[V_{k'}]) = \{0\}$, we can deduce that $P_{I_l}^{\perp}(v) = 0$.

We pick a vector u in $\cap_{k'\in S_j}[V_{k'}]$ such that $S_i \subset S_j$ and $S_j \subset S_j$. As $\cap_{k'\in S_j}[V_{k'}] \subset \cap_{k'\in S_i}[V_{k'}]$ and $\cap_{k'\in S_j}[V_{k'}] \subset \cap_{k'\in S_k}[V_{k'}]$, we have $u \in (\cap_{k'\in S_i}[V_{k'}]) \cap (\cap_{k'\in S_k}[V_{k'}])$. As $S_i \cap S_j \neq \phi$, we observe that $\cap_{k'\in S_j}[V_{k'}] \subset [I_i] = \mathcal{N}(P_{J_i}^{\perp})$.

$$\begin{aligned} u &\in P_{J_i}((\cap_{k'\in S_i}[V_{k'}]) \cap (\cap_{k'\in S_k}[V_{k'}])) \\ &\subset P_{J_i}(\cap_{k'\in S_i}[V_{k'}]) \cap P_{J_i}(\cap_{k'\in S_k}[V_{k'}]). \end{aligned}$$

Let $w = u+v$. Since $u \in P_{J_i}(\cap_{k'\in S_k}[V_{k'}])$ and $v \in P_{J_i}^{\perp}(\cap_{k'\in S_k}[V_{k'}])$ with respect to S_k and the same for S_i, we find that $w \in (\cap_{k'\in S_i}[V_{k'}]) \cap (\cap_{k'\in S_k}[V_{k'}])$, and then, $w \in [I_i] = \mathcal{N}(P_{j_i}^{\perp})$. But this forces $P_{J_i}^{\perp}(w) = v$ and v to be zero and leads to vacuity.

(iii) $|\mathcal{J}_{i,l}| = 1$ and $|\mathcal{J}'_{i,l}| = 1$:

Suppose there is nonzero $v \in P^{\perp}_{J_i}(\cap_{k' \in S_k}[V_{k'}])$ with $k \in \mathcal{J}_{i,l}$. From Lemma A.4, we have $P^{\perp}_{I_l}(v) \in P^{\perp}_{I_l}(\cap_{k' \in S_k}[V_{k'}])$. By the same argument, $P^{\perp}_{I_l}(v) \in P^{\perp}_{I_l}(\cap_{k' \in S_i}[V_{k'}])$. Then there is a nonzero $P^{\perp}_{I_l}(v) \in P^{\perp}_{I_l}(\cap_{k' \in S_i}[V_{k'}]) \cap P^{\perp}_{I_l}(\cap_{k' \in S_k}[V_{k'}])$.

(iv) $|\mathcal{J}_{i,l}| \geq 2$, $|\mathcal{J}'_{i,l}| = 0$ and $v \in P^{\perp}_{J_i}(\cap_{k' \in S_k}[V_{k'}])$ with some $k \in \mathcal{J}_{i,l}$: the same as case (ii).

(v) $|\mathcal{J}_{i,l}| \geq 2$, $|\mathcal{J}'_{i,l}| = 0$ and $v \notin P^{\perp}_{J_i}(\cap_{k' \in S_k}[V_{k'}])$ with any of $k \notin \mathcal{J}_{i,l}$:

We will show that this case is a generalization of case (ii) and also vacuous. As $v \in +_{k \in \mathcal{J}_{i,l}} P^{\perp}_{J_i}(\cap_{k' \in S_k}[V_{k'}])$, we express

$$v = \sum_{j \in \mathcal{J}_{i,l}} a_j v_j$$

for each $a_j \in \mathbb{R}$ (at least two of them are nonzero) and $v_j \in P^{\perp}_{J_i}(\cap_{k' \in S_j}[V_{k'}])$.

By Lemma A.5, for each v_j for $j \in \mathcal{J}_{i,l}$, we can find $S_{m,j}$ such that $v_j \in S_{m,j}$ and $m < i$, $m \neq j$, $S_m \cap S_i = \{0\}$. And this also implies that there exists certain $S_{t,j}$ and $v_j \in S_{t,j}$ such that $S_j \subset S_{t,j}$ and $S_{m,j} \subset S_{t,j}$. Here we point out that as $S_{t,j} \cap S_i \neq \phi$, since $S_j \cap S_i \neq \phi$. Then all $\cap_{k' \in S_{t,j}}[V_{k'}]$ for $j \in \mathcal{J}_{i,l}$ and their linear combinations are subsets of $[I_i]$. This leads to the conclusion $v \in [I_i]$ and shows the vacuity of this case.

(vi) $|\mathcal{J}_{i,l}| \geq 2$, $|\mathcal{J}'_{i,l}| \geq 1$ and $v \in P^{\perp}_{J_i}(\cap_{k' \in S_k}[V_{k'}])$ with some $k \in \mathcal{J}_{i,l}$:

If $k \in \mathcal{J}'_{i,l}$, then the same as the case (iii). If $k \notin \mathcal{J}'_{i,l}$, then the same as the case (ii).

(vii) $|\mathcal{J}_{i,l}| \geq 2$, $|\mathcal{J}'_{i,l}| \geq 1$ and $v \notin P^{\perp}_{J_i}(\cap_{k' \in S_k}[V_{k'}])$ with any of $k \in \mathcal{J}_{i,l}$:

As $v \in +_{k \in \mathcal{J}_{i,l}} P^{\perp}_{J_i}(\cap_{k' \in S_k}[V_{k'}])$, we express

$$v = \sum_{j \in \mathcal{J}_{i,l}} a_j v_j$$

for each $a_j \in \mathbb{R}$ (at least two of them are nonzero) and $v_j \in P^{\perp}_{J_i}(\cap_{k' \in S_j}[V_{k'}])$. Note that if $a_{j'} = 0$ for all $j' \in \mathcal{J}'_{i,l}$, then this case is essentially the same as case (v), so we only consider the situation at least one $a_{j'} \neq 0$ for $j' \in \mathcal{J}'_{i,l}$.

By Lemma A.5 and its consequences in case (v), for each $j \in \mathcal{J}_{i,l} \setminus \mathcal{J}'_{i,l}$, there exists $S_{t,j}$ and $v_j \in S_{t,j}$ such that $S_j \subset S_{t,j}$ and $S_{m,j} \subset S_{t,j}$. As previously discussed, $\cap_{k' \in S_{t,j}}[V_{k'}]$ for $j \in \mathcal{J}_{i,l} \setminus \mathcal{J}'_{i,l}$ is a subset of $[I_i]$. So we rule out the terms involving $j \in \mathcal{J}_{i,l} \setminus \mathcal{J}'_{i,l}$ and then

$$P^{\perp}_{I_l}(v) = \sum_{j' \in \mathcal{J}'_{i,l}} a_{j'} P^{\perp}_{I_l}(v_{j'}).$$

Since each $\cap_{k' \in S_{j'}}[V_{k'}] \neq \{0\}$ for $j \in \mathcal{J}'_{i,l}$ is a subset of $[I_i]$ and at least one $a_{j'} \neq 0$, we deduce that $P^{\perp}_{I_l}(v)$ is non-zero. Therefore, the relative independence is violated.

□

Now the last term in the RHS of our law of nullity is re-expressed as

$$\begin{aligned}
&(+_{k\in\mathcal{I}_{<i}}(\cap_{k'\in S_k}[V_{k'}]))\cap(\cap_{k\in S_i}[V_k]) \qquad\qquad (A.2)\\
=&((+_{k\in\mathcal{J}_{i,<l}}(\cap_{k'\in S_k}[V_{k'}]))\oplus(+_{k\in\mathcal{J}_{i,l}}P^{\perp}_{J_i}(\cap_{k'\in S_k}[V_{k'}])))\cap(\cap_{k\in S_i}[V_k])\\
=&((+_{k\in\mathcal{J}_{i,<l}}(\cap_{k'\in S_k}[V_{k'}]))\oplus(+_{k\in\mathcal{J}_{i,l}}P^{\perp}_{J_i}(\cap_{k'\in S_k}[V_{k'}])))\\
&\quad\cap(P_{J_i}(\cap_{k\in S_i}[V_k])\oplus P^{\perp}_{J_i}(\cap_{k\in S_i}[V_k]))
\end{aligned}$$

In Proposition A.1, we have observed that $+_{k\in\mathcal{J}_{i,l}}P^{\perp}_{J_i}(\cap_{k'\in S_k}[V_{k'}])$ and $P^{\perp}_{J_i}(\cap_{k\in S_i}[V_k])$ are independent. As $P_{J_i}(\cap_{k\in S_i}[V_k])\subset +_{k\in\mathcal{J}_{i,<l}}(\cap_{k'\in S_k}[V_{k'}]$, the term (A.2) becomes

$$(+_{k\in\mathcal{I}_{<i}}(\cap_{k'\in S_k}[V_{k'}]))\cap(\cap_{k\in S_i}[V_k])=(+_{k\in\mathcal{J}_{i,<l}}(\cap_{k'\in S_k}[V_{k'}])\cap(P_{J_i}(\cap_{k\in S_i}[V_k])).$$

Finally, we have demonstrated

$$\begin{aligned}
\mathrm{rank}([W_i]) &= \mathrm{rank}(\cap_{k\in S_i}[V_k])-\mathrm{rank}\left((+_{k\in\mathcal{I}'_{<i}}(\cap_{k'\in S_k}[V_{k'}]))\cap(\cap_{k\in S_i}[V_k])\right)\\
&= \mathrm{rank}(\cap_{k\in S_i}[V_k])\\
&\quad -\mathrm{rank}\left((+_{k\in\mathcal{J}_{i,<l}}(\cap_{k'\in S_k}[V_{k'}])\cap(P_{J_i}(\cap_{k\in S_i}[V_k]))\cap(\cap_{k\in S_i}[V_k])\right)\\
&= \mathrm{rank}(\cap_{k\in S_i}[V_k])-\mathrm{rank}\left(P_{J_i}(\cap_{k\in S_i}[V_k])\right).
\end{aligned}$$

It is notable that the determination of $\mathrm{rank}([W_i])$ depends only on $\mathcal{J}_{i,>l}$, that is, the set of indices j such that $|S_j|>|S_i|$ and $S_j\cap S_i\neq\phi$. In other words, it does not depend on any index-sets of the same size as S_i and their orderings.

As partially-joint score subspaces $[W_i]$ are constructed recursively and the determination of each $[W_i]$'s rank only depends on $\mathcal{J}_{i,>l}$, not on the ordering index-sets of size $l=|S_i|$, we conclude that the set of pairs $\{(S_i,r(S_i)):i\in\mathcal{I}_K\}$ is unique.

□

A.2.3 Proof of Theorem 2.2

Proof. We give a proof recursively on $l \in \mathcal{K}$. If $l = K$, there only exists $[W_1] = \cap_{k' \in S_1}[V_{k'}]$ for $S_1 = \mathcal{K}$, so there is nothing to prove. If $l = K - 1$, by absolute orthogonality, all $P_1^{\perp}(\cap_{k' \in S_i}[V_{k'}])$ are orthogonal each other for $i = \{2, \ldots, K+1\}$. Therefore, in determining each $[W_i]$ for $i = \{2, \ldots, K+1\}$, other $[W_t]$ for $t \in \mathcal{J}_{i,2} = \mathcal{I}_{<i} \cap \{i' : |S_{i'}| = 2\}$ does not affect on the construction of $[W_i]$.

For any $l \leq K - 2$, suppose the statement holds, that is, a partially-joint score subspace $[W_j]$ such that $|S_j| = l' > l$ is uniquely determined only by $[W_{j'}]$ for $S_{j'} > S_j$. For each index $i \in \mathcal{J}_l = \{i' : |S_i| = l\}$ (regardless of ordering), we have $P_{I_l}(\cap_{k' \in S_i}[V_{k'}]) = P_{J_i}(\cap_{k' \in S_i}[V_{k'}])$ by absolute orthogonality. Then among all indices $i' \in \mathcal{J}_l$, $P_{J_{i'}}^{\perp}(\cap_{k' \in S_{i'}}[V_{k'}])$ are orthogonal each other.

Suppose an ordering on the set of all index-sets of size l is given, denoted by $(S_{i_1}, \ldots, S_{i_h})$ with $h = {}_K C_l$. For i_1, $[W_{i_1}]$ is just determined as $P_{J_{i_1}}^{\perp}(\cap_{k' \in S_{i_1}}[V_{k'}])$. Next for i_2, as $[W_{i_1}]$ and $P_{J_{i_2}}^{\perp}(\cap_{k' \in S_{i_2}}[V_{k'}])$ are orthogonal, we check that $P_{i_1}^{\perp} \circ P_{J_{i_2}}^{\perp}(\cap_{k' \in S_{i_2}}[V_{k'}]) = P_{J_{i_2}}^{\perp}(\cap_{k' \in S_{i_2}}[V_{k'}])$. Thus $[W_{i_2}]$ is determined regardless of $[W_{i_1}]$. In recursive manner, for $i \in \{i_3, \ldots, i_h\}$, $[W_i]$ is determined regardless of all $[W_{i'}]$ for $i' \in \mathcal{J}_{i,l}$, or in other word, is uniquely determined as $P_{J_i}^{\perp}(\cap_{k' \in S_i}[V_{k'}])$ depending only on $[W_{j'}]$s for $S_{j'} > l$.

$\square$

A.2.4 Proof of Corollary 2.1

Proof. From the discussions of Section 2, we know that $[W_{(k.)}]$ is indeed the score subspace of Z_k for each $k = 1, \ldots, K$. Thus, given unique $[W_i]$'s for $i \in \mathcal{I}_{(k)}$ from Theorem 2, the subspace $[U_{(k),i}]$ generated by $U_{(k),i}$ is unique. $\square$

A.2.5 Examples

The following two examples presents the cases where relative independence is satisfied and not:

Example A.1. *Let* $K = 3$*,* $n = 4$ *and*

$$V_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad V_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \\ 0 & 0 \end{pmatrix}, \quad V_3 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{pmatrix}$$

then $[I_1] = [(1,0,0,0)^T]$ *and* $P_{I_1}^{\perp}[V_1] = [(0,1,0,0)^T]$*,* $P_{I_1}^{\perp}[V_2] = [(0,1/\sqrt{2},1/\sqrt{2},0)^T]$ *and* $P_{I_1}^{\perp}[V_3] = [(0,0,1/\sqrt{2},1/\sqrt{2})^T]$ *are linearly independent. Thus* $\{[V_1],[V_2],[V_3]\}$ *is relatively independent.*

Example A.2. *Let* $K = 3$*,* $n = 4$ *and*

$$V_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad V_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \\ 0 & 0 \end{pmatrix}, \quad V_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1/\sqrt{2} \\ 0 & 0 & 1/\sqrt{2} \end{pmatrix}$$

As $P_{I_1}^{\perp}[V_3] \cap (P_{I_1}^{\perp}[V_1] + P_{I_1}^{\perp}[V_2]) = [(0,0,1,0)^T]$ *is not empty, thus* $\{[V_1],[V_2],[V_3]\}$ *is not relatively independent.*

We present the following examples that support Theorem 1, that relative independence indeed guarantee the uniqueness of the brief version of *partially-joint structure* $\mathfrak{S}_0 = \{(S_i, r(S_i)) : i \in \mathcal{I}_K, r(S_i) > 0\}$.

Example A.3 (cont'd from Example A.1.)**.** *Under the ordering* $S_1 = \{1,2,3\}, S_5 = \{1\}, S_6 = \{2\}, S_7 = \{3\}$*, we obtain* $[W_1] = [(1,0,0,0)^T]$*,* $[W_5] = [(0,1,0,0)^T]$*,* $[W_6] = [(0,0,1,0)^T]$ *and* $[W_7] = [(0,0,0,1)^T]$*. Then we have*

$$\mathfrak{S}_0 = \{(\{1,2,3\},1), (\{1\},1), (\{2\},1), (\{3\},1)\}.$$

On the other hand, in the case $S_1 = \{1,2,3\}, S_5 = \{2\}, S_6 = \{1\}, S_7 = \{3\}$*, we obtain* $[W_1] = [(1,0,0,0)^T]$*,* $[W_5] = [(0,1/\sqrt{2},1/\sqrt{2},0)^T]$*,* $[W_6] = [(0,-1/\sqrt{2},1/\sqrt{2},0)^T]$ *and* $[W_7] = [(0,0,0,1)^T]$*. The partially-joint structure is still the same as above.*

Example A.4 (cont'd from Example A.2.)**.** *Under the ordering* $S_1 = \{1,2,3\}, S_5 = \{1\}, S_6 = \{2\}, S_7 = \{3\}$*, we obtain* $[W_1] = [(1,0,0,0)^T]$*,* $[W_5] = [(0,1,0,0)^T]$*,* $[W_6] = [(0,0,1,0)^T]$ *and* $[W_7] = [(0,0,0,1)^T]$*. The partially-joint structure is*

$$\mathfrak{S}_0 = \{(\{1,2,3\},1), (\{1\},1), (\{2\},1), (\{3\},1)\}.$$

On the other hand, in the case $S_1 = \{1,2,3\}, S_5 = \{3\}, S_6 = \{2\}, S_7 = \{1\}$*, we obtain* $[W_1] = [(1,0,0,0)^T]$*,* $[W_5] = [(0,1,0,0)^T, (0,0,1/\sqrt{2},1/\sqrt{2})^T]$ $[W_6] = [(0,0,1/\sqrt{2},-1/\sqrt{2})^T]$ *and* $[W_7] = \{0\}$*. This time, the partially-joint structure is*

$$\mathfrak{S}_0 = \{(\{1,2,3\},1), (\{1\},0), (\{2\},1), (\{3\},2)\}.$$

We also present the following examples that support Theorem 2, that absolute orthogonality guarantee the uniqueness of the partially-joint score subspaces.

Example A.5. *Let* $K = 4$*,* $n = 7$ *and*

$$
\begin{aligned}
V_1 &= \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}^T \\
V_2 &= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}^T \\
V_3 &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}^T \\
V_4 &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}^T .
\end{aligned}
$$

This example satisfies absolute orthogonality. Between under two orderings $(S_{11} = \{3,4\}, S_{12} = \{1\}, S_{13} = \{2\})$ *and* $(S_{11} = \{3,4\}, S_{12} = \{2\}, S_{13} = \{1\})$*, the determinations of* $[W_{12}]$ *and* $[W_{13}]$ *are the same.*

Example A.6. *Let* $K = 4$*,* $n = 7$ *and*

$$
\begin{aligned}
V_1 &= \begin{pmatrix} 1/2\sqrt{2} & \sqrt{3}/2\sqrt{2} & 1/\sqrt{2} & 0 & 0 & 0 \end{pmatrix}^T \\
V_2 &= \begin{pmatrix} \sqrt{3}/2\sqrt{2} & 1/2\sqrt{2} & 0 & 1/\sqrt{2} & 0 & 0 \end{pmatrix}^T \\
V_3 &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}^T \\
V_4 &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}^T .
\end{aligned}
$$

This example satisfies relative orthogonality, but not absolute orthogonality. Between under two orderings $(S_{11} = \{3,4\}, S_{12} =$

$\{1\}, S_{13} = \{2\})$ *and* $(S_{11} = \{3,4\}, S_{12} = \{2\}, S_{13} = \{1\})$*, the determinations of* $[W_{12}]$ *and* $[W_{13}]$ *are not the same because* $[V_1]$ *and* $[V_2]$ *are not orthogonal and* $[W_{11}] = (1/\sqrt{2}, 1/\sqrt{2}, 0, 0, 0, 0)^T$ *does not have an effect on the determination of* $[W_{12}]$ *and* $[W_{13}]$ *by definition.*

A.3 Additional Information on Simulation Study

A.3.1 Examples of the Measure of Dissimilarity between Two Partially-Joint Structures

For the comparison between two partially-joint structure, we devised a following concept of the measure of dissimilarity between partially-joint structures.

First, we introduce a *partially-joint structure matrix* $\widehat{T}$ for a (brief version of) partially-joint structure $\widehat{\mathfrak{S}}_0$, a matrix each of whose columns indicates an identified index-set among datasets and each of whose elements show whether the corresponding dataset belongs to that index-set. For example, when $K = 3$, if the estimated partially-joint structure is

$$\widehat{\mathfrak{S}}_0 = \{(\{1,2,3\},2),(\{1,2\},1),(\{1,3\},1),(\{2,3\},1),(\{3\},1)\},$$

then

$$\widehat{T} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Next, consider two partially-joint structure matrices $\widehat{T}_1 \in \{0,1\}^{n\times m_1}$ and $\widehat{T}_2 \in \{0,1\}^{n\times m_2}$. Discarding all the identical columns between $\widehat{T}_1$ and $\widehat{T}_2$, we denote the remaining columns $\tilde{T}_1$ and $\tilde{T}_2$. For example, from

$$\widehat{T}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \widehat{T}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

we obtain

$$\tilde{T}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \qquad \tilde{T}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

For each remaining column of $\widehat{T}_1$ (or $\widehat{T}_2$), find the closest column of $\widehat{T}_2$ (or $\widehat{T}_1$) in the Frobenius norm sense. The measure of dissimilarity between $\widehat{T}_1$ and $\widehat{T}_2$ is the sum of the squares of all these Frobenius norms between the remaining columns between $\tilde{T}_1$ and $\tilde{T}_2$. In the example above, for $(1,1,0)^T$ of $\tilde{T}_1$, the closest column of $\tilde{T}_2$ is $(1,1,1)^T$ and the difference is 1. For $(1,1,1)^T$ of $\tilde{T}_2$, the difference between $(1,1,0)^T$ is 1 and for $(0,1,1)^T$ of $\tilde{T}_2$, it is 2. The overall difference between $\widehat{T}_1$ and $\widehat{T}_2$ is then $1^2 + 1^2 + 2^2 = 6$. The measure of dissimilarity between two partially-joint structure matrix $\widehat{T}_1$ and $\widehat{T}_2$ is denoted $\text{diff}(\widehat{T}_1, \widehat{T}_2)$.

Finally, if the partially-joint structure $\widehat{\mathfrak{S}}_{0,1}$ (or $\widehat{\mathfrak{S}}_{0,2}$) has partially-joint structure matrices $\widehat{T}_1$ (or $\widehat{T}_2$), then the measure of dissimilarity between $\widehat{\mathfrak{S}}_{0,1}$ and $\widehat{\mathfrak{S}}_{0,2}$ is $\text{diff}(\widehat{\mathfrak{S}}_{0,1}, \widehat{\mathfrak{S}}_{0,2}) = \text{diff}(\widehat{T}_1, \widehat{T}_2)$.

A.3.2 Review on Methodology of Other Methods

We briefly review the methodology of AJIVE (Feng et al.,2018), SLIDE (Gaynanova and Li,2019), COBS (Gao et al.,2020) and JIVE (Lock et al.,2013).

AJIVE In AJIVE, each signal matrix $Z_k \in \mathbb{R}^{p_k \times n}$ is regarded as a sum of joint structure J_k and individual structure I_k for $k = 1, \dots, K$. A joint structure J_k is viewed as the score subspace $[V_M] \in \mathbb{R}^n$ shared by all Z_i's.

AJIVE extracts each estimated signal matrix $\widehat{Z}_k$ from dataset X_k and $\widehat{Z}_k$ is of initial rank estimate $\widetilde{r}_k$. The estimated shared joint component $[\widehat{V}_M]$ is obtained as a flag mean among score subspaces of $\widehat{Z}_k$'s, or $[\widehat{V}_1], \ldots, [\widehat{V}_k]$, in a sense of the projection Frobenius norm distance as our method. The rank r_J of $[\widehat{V}_M]$ (called the joint rank) is estimated using the simulated distribution of the largest singular value of the concatenated matrix of random directions and that of Wedin bounds:

(1) If the largest squared singular value of the column concatenation matrix $\widehat{V}$ of $\widehat{V}_k$'s is larger than the 5th percentile of the simulated distribution of the largest squared singular value of the concatenation matrix of random orthogonal matrices of the same size as $\widehat{V}_k$'s (or random direction bound), then $[\widehat{V}_M]$ is not generated by noise in 95 percent of probability.

(2) If there are $\widehat{r}_J$ squared singular values of $\widehat{V}$ are larger than the 95th percentile of the simulated distribution of Wedin bounds, then the first $\widehat{r}_J$ right singular vectors are used as the basis for the estimated joint score subspace $[\widehat{V}_M]$.

The estimated joint structure $\widehat{J}_k$ is a projection of the dataset X_k onto the estimated joint score subspace $[\widehat{V}_M]$, that is, $\widehat{J}_k = X_k \widehat{V}_M \widehat{V}_M^T$. Each estimated individual structure $\widehat{I}_k$ is obtained as $X_k \cdot (I - \widehat{V}_M \widehat{V}_M^T)$. The row spaces of each estimated individual structure $\widehat{I}_k$ is orthogonal to $[\widehat{V}_M]$. There is no guarantee that individual structures are mutually orthogonal. The joint score matrix is just defined as $\widehat{V}_M$ and the corresponding joint loading matrix for kth data source is a regression of $\widehat{J}_k$ on $\widehat{V}_M^T$, computed as $\widehat{J}_k \cdot \widehat{V}_M$. The individual loading and score matrices are obtained from SVDs of $\widehat{I}_k$'s.

SLIDE SLIDE identifies the partially-joint structure with the penalized matrix factorization, that is,

$$(\widetilde{U}, \widetilde{V}) = \underset{U,V}{\arg\min} \sum_{k=1}^{K} \frac{1}{2} \|X_k - U_k V^T\|_F^2 + \lambda \sum_{j=1}^{r} \|U_{kj}\|_2 \quad \text{s.t. } V^T V = I,$$

where U_{kj} is the jth column of the loading matrix U_k of X_k and r is the number of all possible sparsity patterns. After computing $\widetilde{U}$ and $\widetilde{V}$ with an iterative algorithm, the corresponding structure matrix $\widehat{T}$ is obtained from the sparse structure of $\widetilde{U}$. Note that the concept of the structure matrix $\widehat{T}$ here is identical to the partially-joint structure matrix of ours in Section B.1. Even though this optimization problem is nonconvex and there is no guarantee about convergence to the global optimum, authors reported that a local solution can be obtained heuristically by initializaing V with the left singular matrix of concatenated X that works well in the simulation.

Then SLIDE estimate the loading and score matrices, $\widehat{U}$ and $\widehat{V}$, for the structure $\widehat{T}$ by solving the following optimization problem with an iterative algorithm,

$$(\widehat{U}, \widehat{V}) = \underset{U,V}{\arg\min} \|X - UV^T\|_F^2 \quad \text{s.t. } V^T V = I,$$

with the constraint that the loading U has the same sparsity structure as $\widehat{T}$.

In model validation, SLIDE adapt the block cross validation (BCV) procedure to select the best structure $\widehat{T}_{\text{best}}$. BCV splits rows and columns of each dataset X_k into submatrices $X_k^{11}, X_k^{12}, \ldots, X_k^{21}, \ldots$.

Then it holds out a set of submatrices $X^{ij} = [X_1^{ij}, X_2^{ij}, \dots, X_K^{ij}]$ of the same sub-block position in each dataset X_k and evaluate the prediction error on X^{ij}'s. Given a set of structure candidates $\widehat{T}_1, \widehat{T}_2, \dots$, we select one that minimizes the error across all folds.

JIVE Like in AJIVE, JIVE decompose each signal matrix X_i as a sum of joint structure J_i and individual structure I_i, for $i = 1, \dots, K$. After defining R to be a row concatenation of $R_i = X_i - J_i - I_i$, JIVE estimate both joint and individual structures by minimizing $\|R\|_F^2$ under the given ranks. An alternating iterative algorithm is implemented for the estimation finding individual component with given joint component at one step and vice versa at another step. The estimated joint structure is identical to the first r_J terms in the SVD of X with individual components removed and the estimated individual structures to the first r_i terms in the SVD of X_i with the joint component removed. The selection of r_J and r_is are validated using the permutation test.

COBS COBS iteratively estimates a sequence of loading vectors, u_i for $i = 1, \dots, r$ for given r, while updating the data matrix $X = [X_1, \dots, X_k] \in \mathbb{R}^{n \times \sum p_k}$. The algorithm starts with $X^{[0]} = X$. At the ith step, with the current data matrix $X^{[i-1]}$, the ith loading vector u_i is estimated solving the following maximization problem, that is,

$$\widetilde{u}_i = \max_u \|(X^{[i-1]})^T u\|_2^2 \quad \text{s.t. } u^T u = 1.$$

As each u_i is equipped with block structure $u_i = \left(u_{(1)}^T \dots u_{(K)}^T\right)^T$, we can give sparsity at two levels of thresholding, one for block-wise sparsity and the other for overall sparsity in estimating $\widetilde{u}_i$. The tuning parameters $\alpha_v \in [0, 1]$ and $\lambda_v \geq 0$ control the two levels of sparsity respectively, and $\widehat{u}$ is thresholded as a normalized solution of

$$\min_{x} \frac{1}{2}\|x - \widetilde{u}_i\|^2 + \gamma_1\|x\|_1 + \gamma_2\|x\|_{2,1},$$

where $\gamma_1 = \alpha_v \lambda_v$ and $\gamma_2 = (1 - \alpha_v)\lambda_v$. Then the score vector v_i is estimated as the empirical BLUP and dataset $X^{[i-1]}$ is updated as $X^{[i]} = X^{[i-1]} - \widehat{u}_i \widehat{v}^T$.

A.3.3 Results on Tuning Parameter Selection

We report the performances of our tuning parameter selection procedure of Section 2.3.3 using the six models given in Section 2.5. Since the tuning parameter λ represents a threshold for principal angles, the candidates for λ are given by $\lambda = 0°, 1°, \ldots, 90°$. For each value of λ, we evaluated the empirical risk. As discussed in Section 2.3.3, we take the parameter $\tilde{\lambda}$ that gives the smallest empirical risk and also compare $\mathrm{diff}(\widehat{\mathfrak{S}}_0(\tilde{\lambda}_0; \widehat{Z}_{tr}), \widehat{\mathfrak{S}}_0(\lambda; \widehat{Z}))$ as a function of λ. We also present $\mathrm{diff}(\widehat{\mathfrak{S}}_0(\lambda, \widehat{Z}), \mathfrak{S}_0)$, which reflects how much the estimated structure differs from the true structure on each value of angle threshold, under the situation where the true structure ('oracle') is known.

When SNR = 10, `getnfrac` function estimated true signal ranks correctly for models 1 to 5. The empirical risk is minimized at an interval of λ's, and for any λ in the interval, the corresponding structure $\widehat{\mathfrak{S}}_0$ matches the true $\mathfrak{S}_0$; the valley bottoms of empirical risk (solid line) are posited inside those of the measure of dissimilarity (dashed line) with value zero, as seen in Fig. A.3.1. For each model in the figure, solid line (empirical risk) shows a similar shape as dashed line (the measure of dissimilarity), which implies that empirical risk well reflects the difference between the estimated structure and the true structure. In Model 6, the true signal rank is 8 for each dataset, for instance, from $\{(\{1,2,3\},2), (\{1,2\},2), (\{1,3\},2), (\{1\},2)\}$ in the case X_1. However, `getnfrac` function only estimated $r_1 = 8$, $r_2 = 8$ and $r_3 = 7$. The estimated partially-joint structure lacks $(\{3\},1)$ from the true structure.

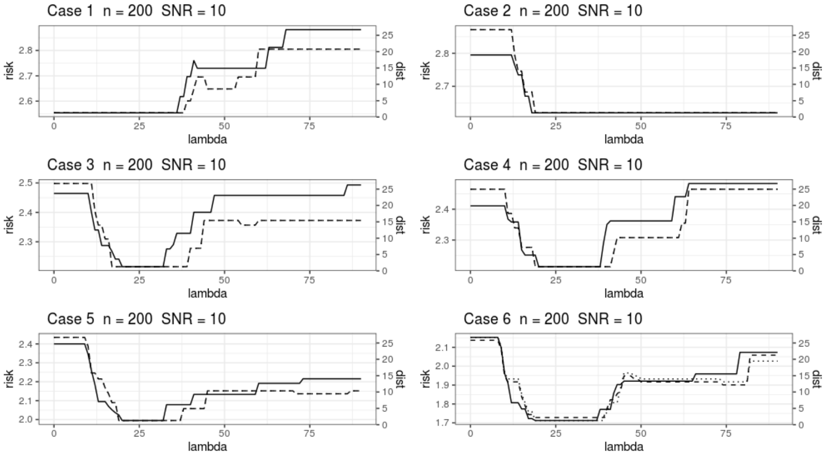


Figure A.3.1: The values of empirical risk (solid) and diff$(\widehat{\mathfrak{S}}_0(\tilde{\lambda}_0;\widehat{Z}_{tr}),\widehat{\mathfrak{S}}_0(\lambda;\widehat{Z}))$ (dotted), diff$(\widehat{\mathfrak{S}}_0(\lambda,\widehat{Z}),\mathfrak{S}_0)$, (dashed).

When the signal-to-noise ratio is small, SNR $= 2$, `getnfrac` function estimated signal ranks as zero for all six cases, so we give the true inherent signal ranks instead. Unfortunately, the empirical risk is minimized at smaller values of λ than desired; *see* Fig. A.3.2. Unlike Fig. A.3.1, solid line (empirical risk) shows a far different shape than dashed line (the measure of dissimilarity), except for Model 1, which implies that empirical risks fail to detect the true structure. This is due to the lower value of SNR, with which the magnitude of noise overwhelms that of signal. As the score vectors of each dataset have almost random directions in low SNRs, there is a tendency that signals from a partially-joint (and fully-joint) scores are counted separately as if they belong to individual data blocks.

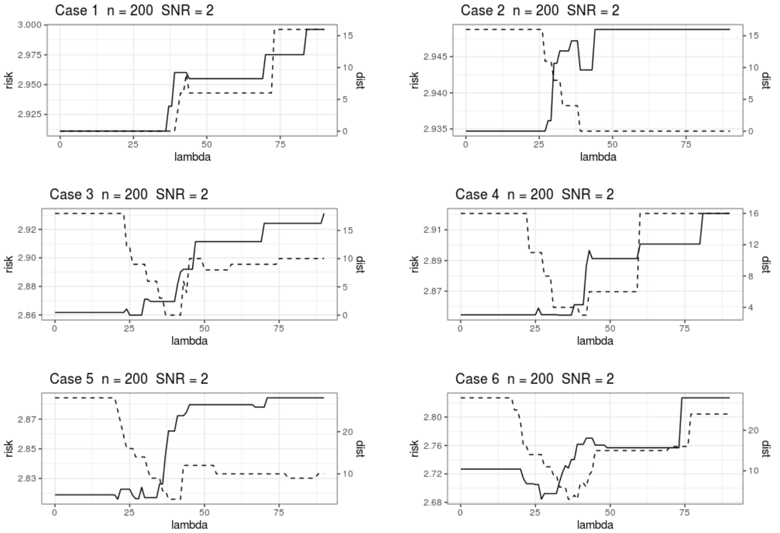


Figure A.3.2: The values of penalized empirical risk for λ when $n = 200$ and SNR $= 2$. The values of empirical risk (solid line) and the measure of dissimilarity, diff$(\widehat{\mathfrak{S}}_0(\lambda, \widehat{Z}), \mathfrak{S}_0)$, (dashed line) over varying λ are shown.

A.3.4 Simulation Settings for Section 2.5.3

In the unbalance of signal strength between joint and individual component settings, we set $n = 200$ and $K = 3$.

In the first case, we set the elements of σ_M^2 as

(1) Joint (S_1) : $(15, 14.5, \ldots, 5.5)$,

(2) Individual 1 (S_2) : $(0.150, 0.141, 0.132, \ldots, 0.069)$,

(3) Individual 2 (S_3) : $(0.147, 0.138, 0.129, \ldots, 0.066)$,

(4) Individual 3 (S_4) : $(0.144, 0.135, 0.126, \ldots, 0.063)$,

so the strength of joint signals are about 100 times stronger than those of individual signals.

In the second case case, we set the elements of σ_M^2 as

(1) Joint (S_1) : $(0.15, 0.145, \ldots, 0.055)$,

(2) Individual 1 (S_2) : $(15, 14.1, 13.2, \ldots, 6.9)$,

(3) Individual 2 (S_3) : $(14.7, 13.8, 12.9, \ldots, 6.6)$,

(4) Individual 3 (S_4) : $(14.4, 13.5, 12.6, \ldots, 6.3)$,

so the strength of individual signals are about 100 times stronger that those of joint signals.

A.4 Additional Information on Real Data Analysis

Table A.4.1: Fisher's exact tests between gene mutations/chromosome defects and the CLL subgroups, α and β with top 5 adjested p-values (Benjamini-Hochberg method).

Gene mutation/Chromosome defect	adjusted p-value
IGHV	$\mathbf{1.036 \times 10^{-13}}$
MED12	0.173
del17p13	0.174
del13q14	0.178
TP53	0.184

Table A.4.2: The association between IGHV mutation status and the CLL subgroups with adjusted p-value (9 missing values excluded).

Mutation \ Subgroup	α	β
IGHV Wild type	49	8
IGHV Mutated	7	48

Table A.4.3: The most appeared drugs in the subgroups [a] and [b] with appearances at least four times out of 5 concentrations.

The most appeared drugs in subgroup [a]		
Drug name	Target pathway	Appearances
spebrutinib	BTK	5
idelalisib	PI3K delta	5
duvelisib	PI3K gamma, PI3K delta	5
tamatinib	SYK	5
dasatinib	ABL1, KIT, LYN, PDGFRA, PDGFRB, SRC	5
PF 477736	CHK1, CHK2	5
MK-2206	AKT1/2 (PKB)	5
ibrutinib	BTK	4
selumetinib	MEK1/2	4
PRT062607 HCL	SYK	4
AZD7762	CHK1/2	4
CCT241533	CHK2	4
TAE684	ALK	4
MK-1775	WEE1	4
AT13387	HSP90	4

The most appeared drugs in the subgroup [b]		
Drug name	Target pathway	Appearances
everolimus	mTOR	5
thapsigargin	SERCA	5
orlistat	LPL	5
rotenone	Electron transport chain in mitochondria	5
afatinib	EGFR, ERBB2	4
fludarabine	Purine analogue	4
navitoclax	BCL2, BCL-XL, BCL-W	4

Appendix B

Supplementary Materials for Chapter 3

B.1 Proofs

B.1.1 Proof of Lemma 3.1

Proof. Suppose we have any $p_1 = [v_1, r_1]$ and $p_2 = [v_2, r_2] \in \Theta$. First, it is easy to check that $d(p_1, p_1) = 0$. The non-negativity and symmetry of d are trivial.

We show a triangle inequality. Let $p_3 = [v_3, r_3]$ be any point in Θ. By the spherical triangle inequality [Ramsay and Richtmyer, 1995], we have $\arccos(v_1, v_2) \leq \arccos(v_1, v_3) + \arccos(v_2, v_3)$.

Without loss of generality, $d_1(p_1, p_2) \leq d_2(p_1, p_2)$. Then

$$\begin{aligned}
d_1(p_1, p_2) &= \big(\arccos^2(v_1^T v_2) + |r_1 - r_3 + r_3 - r_2|^2\big)^{1/2} \\
&\leq \Big(\big(\arccos(v_1^T v_3) + \arccos(v_2^T v_3)\big)^2 + (|r_1 - r_3| + |r_3 - r_2|)^2\Big)^{1/2} \\
&\leq \big(\arccos^2(v_1^T v_3) + |r_1 - r_3|^2\big)^{1/2} + \big(\arccos^2(v_2^T v_3) + |r_2 - r_3|^2\big)^{1/2} \\
&= d_1(p_1, p_3) + d_1(p_2, p_3)
\end{aligned}$$

and

$$\begin{aligned}
d_2(p_1, p_2) &= \big(\arccos^2(-v_1^T v_2) + |r_1 - (\pi - r_2)|^2\big)^{1/2} \\
&\leq \Big(\big(\arccos(v_1^T v_3) + \arccos(-v_2^T v_3)\big)^2 + (|r_1 - r_3| + |\pi - r_3 - r_2|)^2\Big)^{1/2} \\
&= d_1(p_1, p_3) + d_2(p_2, p_3).
\end{aligned}$$

or $d_1(p_1, p_2) \leq d_1(p_1, p_3) + d_2(p_2, p_3)$. In a similar way,

$$\begin{aligned}
d_1(p_1, p_2) &\leq d_2(p_1, p_3) + d_2(p_2, p_3) \\
d_1(p_1, p_2) &\leq d_2(p_1, p_3) + d_1(p_2, p_3).
\end{aligned}$$

□

B.1.2 Proof of Proposition 3.1

Proof. If E is empty, it is trivial. In the case E non-empty, we first show (BPC) implies (ZC). Let $B_n = \cup_{k=n}^{\infty} E_k$ and $B = \cap_{n=1}^{\infty} \overline{B_n}$. By (BPC), there is a sufficiently large n such that $d(p_n, E) \leq \epsilon$ for any $p_n \in B_n$, that is, there exists $q \in E$ with $d(p, q) \leq \epsilon$.

Given $p \in B$, we can find $p_N \in B_n$ for sufficiently large N such that $d(p_N, p) \leq \epsilon$ and $d(p, E) \leq \epsilon$. Then $d(p, E) = \inf_{q \in E} d(p, q) \leq d(p_n, p) + \inf_{q \in E} d(p_N, q) \leq 2\epsilon$. Since E is a closed set, $p \in E$ as $\epsilon \to 0$. This proves (ZC).

We next show (ZC) implies (BPC). Consider a sequence $p_n \in E_n$ such that $p_n = \arg\max_{p \in E_n} d(p, E)$ and let $r_n = \max_{p \in E_n} d(p, E)$. Then we have either $r_n \to 0$ or $r_n \not\to 0$. If $r_n \to 0$, then (BPC) holds. In the case $r_n \not\to 0$, there exists a sequence $n(k)$ such that $r_{n(k)} \geq r_0 > 0$. If $p_{n(k)}$ has an accumulation point p', then $d(p', E) > 0$, which leads to a contradiction to (ZC). Thus $p_{n(k)}$ has no accumulation point.

As Θ is compact and thus totally bounded, we can find a finite cover $\{A_j\}$ of Θ such that $\mathrm{diam}(A_j) = \epsilon$ for any small $\epsilon > 0$. Since there is no accumulation point, each A_j contains only finitely many $p_{n(k)}$. However, this is contradict to the existence of the subsequence $p_{n(k)}$.

$\square$

B.1.3 Proof of Theorem 3.1

Proof. Let $\Delta(x, p_1, p_2) = \rho(x, p_1) - \rho(x, p_2)$ for $x \in S^m$ and $p_1, p_2 \in \Theta$. Under uniform continuity, for every $p \in \Theta$ and $\epsilon > 0$, there exists $\delta > 0$ such that $|\Delta(X, p, p')| < \epsilon$ a.s. whenever $d(p, p') < \delta$.

For $\omega \in \Omega$ and $p = [v, r] \in \Theta$, denote

$$F_n(p) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i(\omega), p)^2 + j_\tau(r),$$

$$F(p) = \mathbb{E}(\rho(X, p)^2) + j_\tau(r),$$

$$\ell_n = \inf_{p \in \Theta} F_n(p), \quad \ell = \inf_{p \in \Theta} F(p),$$

$$E_n^\tau = \{p \in \Theta : F_n(p) = \ell_n\}, \quad E^\tau = \{p \in \Theta : F(p) = \ell\}.$$

Let $\{p_1, p_2, \ldots\} \subset \Theta$ be dense. For each $k = 1, 2, \ldots$, given $p_k = [v_k, r_k]$, we have $F_n(p_k) \to F(p_k)$ almost surely by the Strong

Law of Large Numbers. We put this in another way as follows. Let $A_k = \{\omega : F_n(p_k) \to F(p_k), n \to \infty\} \subset \Omega$, then $\mathcal{P}(A_k) = 1$. With $A = \cap_{k=1}^{\infty} A_k$, we have $\mathcal{P}(A) = 1$ and

$$F_n(p_k) \to F(p_k)$$

for all $\omega \in A$ and each $k = 1, 2, \ldots$.

For any $p = [v, r], p' = [v', r'] \in \Theta$, we check

$$\begin{aligned} |F_n(p') - F_n(p)| &\leq \frac{1}{n} \sum_{i=1}^{n} \big(\rho(X_i, p') + \rho(X_i, p)\big) \, |\rho(X_i, p') - \rho(X_i, p)| \\ &\quad + |j_\tau(r') - j_\tau(r)| \\ &= \frac{1}{n} \sum_{i=1}^{n} \big(2\rho(X_i, p') + |\Delta(X_i, p, p')|\big) \, |\Delta(X_i, p, p')| \\ &\quad + |j_\tau(r') - j_\tau(r)|. \end{aligned} \tag{B.1}$$

By plugging $p_k = [v_k, r_k]$ into p' in (B.1),

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} \rho(X_j, p_k)^2 + j_\tau(r_k) - \frac{1}{n} \sum_{i=1}^{n} (2\rho(X_i, p_k) + |\Delta(X_i, p, p_k)|) \, |\Delta(X_i, p, p_k)| \\ &\quad - |j_\tau(p_k) - j_\tau(p)| \leq \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, p)^2 + j_\tau(p) \leq \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, p_k)^2 + j_\tau(r_k) \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} (2\rho(X_i, p_k) + |\Delta(X_i, p, p_k)|) \, |\Delta(X_i, p, p_k)| + |j_\tau(r_k) - j_\tau(r)|. \end{aligned}$$

We now consider a situation $p_k \to p$ as $k \to \infty$ for any $p \in \Theta$. Then for arbitrary $\epsilon, \epsilon' > 0$, there exists a sufficiently large k_0 such that for $k > k_0$, (a) by uniform continuity, $|\Delta(X_j, p, p_k)| < \epsilon$, (b) $|j_\tau(r_k) - j_\tau(r)| < \epsilon'$ by (P2) and (P3). As $n \to \infty$, we also have (c) $\frac{1}{n} \sum_{i=1}^{n} \rho(X_j, p_k) \to \mathbb{E}\rho(X, p_k)^2$ by the Strong Law of Large

Numbers. Combining (a)-(c),

$$\begin{aligned}
&\mathbb{E}\rho(X, p_k)^2 + j_\tau(r_k) - \ (2\mathbb{E}\rho(X, p_k) + \epsilon)\,\epsilon - \epsilon' \\
&\quad \le \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \rho(X_i, p)^2 + j_\tau(r) \le \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \rho(X_i, p)^2 + j_\tau(r) \\
&\quad \le \mathbb{E}\rho(X_j, p_k)^2 + j_\tau(r_k) + (2\mathbb{E}\pi(X_i, p_k) + \epsilon)\,\epsilon + \epsilon'.
\end{aligned}$$

Sending $\epsilon, \epsilon' \to 0$ and $n \to \infty$, as $\frac{1}{n}\sum_{i=1}^{n} \rho(X_i, p)^2$ and $j_\tau(r)$ are bounded above and below, for any subsequence $p_{k_s} \to p$, we have

$$\lim_{s\to\infty} \mathbb{E}\rho(X, p_{k_s})^2 + j_\tau(r_{k_s}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \rho(X_i, p)^2 + j_\tau(r).$$

Thus, for all $\omega \in A$,

$$F_n(p) \to F(p). \tag{B.2}$$

Next, plugging p_k into p and p into p' in (B.1),

$$\begin{aligned}
|F_n(p_n) - F_n(p)| \le \frac{1}{n}\sum_{i=1}^{n} &\left(2\rho(X_i, p) + |\Delta(X_i, p_n, p)|\right) |\Delta(X_i, p_n, p)| \\
&+ |j_\tau(r') - j_\tau(r)| \to 0.
\end{aligned}$$

by (a) and (b). Thus $|F_n(p_n) - F(p)| \le |F_n(p_n) - F_n(p)| + |F_n(p) - F(p)| \to 0$ and

$$F_n(p_n) \to F(p) \tag{B.3}$$

for all sequences $p_n \to p$ and $\omega \in A$.

Finally, we assert the theorem.

- ○ When $\cap_{n=1}^{\infty}\overline{\cup_{k=n}^{\infty} E_n^\tau} = \phi$: trivial

- ○ When $\cap_{n=1}^{\infty}\overline{\cup_{k=n}^{\infty}E_n^{\tau}} \neq \phi$: Denote $B_n = \cup_{k=n}^{\infty}E_k^{\tau}$. Let $\{b_n \in B_n\}$ be a sequence such that $b_n \to b \in \cap_{n=1}^{\infty}\overline{B_n}$. For each n, we can find an index k_n such that $b_n \in E_{k_n}^{\tau}$. We want to show that $b \in E^{\tau}$.

  By (B.3), $\ell_{k_n} = F_{k_n}(p_{k_n}) \to F(b) \geq \ell$. At the same time, by (B.2), for any arbitrary $q \in \Theta$, there exists a sequence $\epsilon_n \to 0$ such that $F(q) \geq F_n(q) - \epsilon_n \geq \ell_n - \epsilon_n$. Then $\inf_{q\in\Theta} F(q) = \ell \geq \limsup_{n\to\infty} \ell_n$ and we conclude $\ell_n \to \ell = F(b)$.

□

B.1.4 Proof of Theorem 3.2

Proof. We first introduce the concept of epi-convergence (Rockafellar and Wets,1998).

Definition B.1. *Consider arbitrary functions h and $\{h_k\}$ defined on a metric space (Θ, d) to $\mathbb{R}$ for $k = 1, 2, \ldots$. The sequence $\{h_k\}$ epi-converges to h at $p \in \Theta$ if and only if*

(1) for every sequence $q_k \to q \in \Theta$, $\liminf_{k\to\infty} h_k(q_k) \geq h(q)$

(2) there exists a sequence $q_k \to q \in \Theta$, $\limsup_{k\to\infty} h_k(q_k) \leq h(q)$.

Our first aim is to assure that, for $p = (v, r)$, the function $f_{\tau_k}(p) = \mathbb{E}_X\rho^2(X, p) + j_{\tau}(r)$ epi-converges to $f_0(p) = \mathbb{E}_X\rho^2(X, p)$. To show epi-convergence, we utilize the following proposition, *see* Proposition 7.15 of Rockafellar and Wets (1998) or Royset (2018).

Proposition B.1. *Let a function h and a sequence of functions h_k for $k = 1, 2, \ldots$ be on a metric space (Θ, d) to $\mathbb{R}$. If $\{h_k\}$ are* *continuous and converges uniformly to h, then $\{h_k\}$ epi-converges to f.*

By definition, $f_{\tau_k}(p)$ are continuous on $p = [v, r] \in \Theta$. Also $\{f_{\tau_k}\}$ uniformly converges to f_0, since for all $p \in \Theta$,

$$|f_{\tau_k}(p) - f_0(p)| = |\rho_{\tau_k}(r)| \to 0$$

as $k \to \infty$ by (P4). Thus, $\{f_{\tau_k}\}$ epi-converges to f_0.

Before going into the next step, we define the following set, the population solution set inflated by $\epsilon > 0$. Given $\tau \geq 0$ and $\epsilon > 0$,

$$E^{\tau,\epsilon} = \left\{ p \in \Theta : \mathbb{E}_X \rho^2(X, p) + j_\tau(p) \leq \epsilon + \inf_{q \in \Theta} \mathbb{E}_X \rho^2(X, q) + j_\tau(q) \right\}.$$

Our second aim is to show, given any decreasing sequence $\epsilon_k \to 0$, E^{τ_k,ϵ_k} converges to E in outer limit. We use the following fact, *see* Theorem 7.31 of Rockafellar and Wets (1998).

Proposition B.2. *Let a function h and a sequence of functions h_k, $k = 1, 2, \ldots$ be defined on a metric space (Θ, d) to $\mathbb{R}$. Suppose $\{h_k\}$ epi-converges to h with $-\infty < \inf h < \infty$. Let $\epsilon_k \to 0$ be a decreasing sequence and set $B_k^\epsilon = \{p \in \Theta : h_k(p) \leq \epsilon + \inf h_k\}$ and $B = \{p \in \Theta : \inf h\}$. Then B_k^ϵ converges to B in outer limit as $k \to \infty$.*

In our case, as $-\infty < 0 \leq \mathbb{E}_X \rho^2(X, p) \leq \pi^2/4 < \infty$, we have E^{τ_k,ϵ_k} converges to E in outer limit as $k \to \infty$.

From the fact $E^{\tau_k} \subset E^{\tau_k,\epsilon_k}$, it is easily derived that $\cap_{n=1}^{\infty} \overline{\cup_{k=n}^{\infty} E^{\tau_k}} \subset \cap_{n=1}^{\infty} \overline{\cup_{k=n}^{\infty} E^{\tau_k,\epsilon_k}}$. Therefore, we conclude that E^{τ_k} converges to E in outer limit as $k \to \infty$.

□

B.1.5 Proof of Theorem 3.3

Proof. Let $\Delta(x, p_1, p_2) = \rho(x, p_1) - \rho(x, p_2)$ for $x \in S^m$ and $p_1, p_2 \in \Theta$. Under uniform continuity, for every $p \in \Theta$ and $\epsilon > 0$, there exists $\delta > 0$ such that $|\Delta(X, p, p')| < \epsilon$ a.s. whenever $d(p, p') < \delta$.

For $\omega \in \Omega$ and $p = [v, r] \in \Theta$, denote

$$\begin{aligned}
&F_n^\tau(p) = \frac{1}{n}\sum_{i=1}^n \rho(X_i(\omega), p)^2 + j_\tau(r),\\
&F_n(p) = \frac{1}{n}\sum_{i=1}^n \rho(X_i(\omega), p)^2,\\
&F(p) = \mathbb{E}(\rho(X, p)^2),\\
&\ell_n = \inf_{p\in\Theta} F_n(p), \quad \ell = \inf_{p\in\Theta} F(p),\\
&E_n^\tau = \{p \in \Theta : F_n(p) = \ell_n\}, \quad E^\tau = \{p \in \Theta : F(p) = \ell\}.
\end{aligned}$$

Let $\{p_1, p_2, \ldots\} \subset \Theta$ be dense. For each $k = 1, 2, \ldots$, given $p_k = [v_k, r_k]$, we have $F_n(p_k) \to F(p_k)$ almost surely as $n \to \infty$ by the Strong Law of Large Numbers. By (P4), we also have $j_{\tau_n}(r_k) \to 0$ monotonically as $n \to \infty$. By the continuous mapping theorem, combining two results, for each p_k, $F_n^{\tau_n}(p_k) \to F(p_k)$ almost surely as $n \to \infty$. We put this as follows. Let $A_k = \{\omega : F_n^{\tau_n}(p_k) \to F(p_k), n \to \infty\} \subset \Omega$, then $\mathcal{P}(A_k) = 1$. With $A = \cap_{k=1}^\infty A_k$, we have $\mathcal{P}(A) = 1$ and

$$F_n(p_k) \to F(p_k)$$

for all $\omega \in A$ and each $k = 1, 2, \ldots$.

For any $p = [v, r], p' = [v', r'] \in \Theta$, we check

$$\begin{aligned}
|F_n^{\tau_n}(p') - F_n^{\tau_n}(p)| &\le \frac{1}{n}\sum_{i=1}^{n} \big(\rho(X_i, p') + \rho(X_i, p)\big)\, |\rho(X_i, p') - \rho(X_i, p)| \\
&\quad + |j_{\tau_n}(r') - j_{\tau_n}(r)| \\
&= \frac{1}{n}\sum_{i=1}^{n} \big(2\rho(X_i, p') + |\Delta(X_i, p, p')|\big)\, |\Delta(X_i, p, p')| \\
&\quad + |j_{\tau_n}(r') - j_{\tau_n}(r)|.
\end{aligned} \tag{B.4}$$

By plugging $p_k = [v_k, r_k]$ into p' in (B.4),

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \rho(X_j, p_k)^2 + j_{\tau_n}(r_k) - \frac{1}{n}\sum_{i=1}^{n} \left(2\rho(X_i, p_k) + |\Delta(X_i, p, p_k)|\right) |\Delta(X_i, p, p_k)| \\
&\quad - |j_{\tau_n}(p_k) - j_{\tau_n}(p)| \le \frac{1}{n}\sum_{i=1}^{n} \rho(X_j, p)^2 + j_{\tau_n}(p) \le \frac{1}{n}\sum_{i=1}^{n} \rho(X_j, p_k)^2 + j_{\tau_n}(r_k) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left(2\rho(X_i, p_k) + |\Delta(X_i, p, p_k)|\right) |\Delta(X_i, p, p_k)| + |j_{\tau_n}(r_k) - j_{\tau_n}(r)|.
\end{aligned}$$

We now consider a situation $p_k \to p \in \Theta$ as $k \to \infty$ for some p. Then for arbitrary $\epsilon, \epsilon' > 0$, there exists a sufficiently large k_0 such that for $k > k_0$, (a) by uniform continuity, $|\Delta(X_j, p, p_k)| < \epsilon$, (b) $|j_{\tau_n}(r_k) - j_{\tau_n}(r)| < \epsilon'$ by (P2) and (P3). As $n \to \infty$, we also have (c) $\frac{1}{n}\sum_{i=1}^{n} \rho(X_j, p_k) \to \mathbb{E}\rho(X, p_k)^2$ almost surely by the Strong Law of Large Numbers. Combining (a)-(c),

$$\begin{aligned}
&\mathbb{E}\rho(X, p_k)^2 + j_{\tau_n}(r_k) - \ \left(2\mathbb{E}\rho(X, p_k) + \epsilon\right) \epsilon - \epsilon' \\
&\quad \le \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \rho(X_j, p)^2 + j_{\tau_n}(r) \le \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \rho(X_j, p)^2 + j_{\tau_n}(r) \\
&\quad \le \mathbb{E}\rho(X_j, p_k)^2 + j_{\tau_n}(r_k) + \left(2\mathbb{E}\rho(X_i, p_k) + \epsilon\right) \epsilon + \epsilon'.
\end{aligned}$$

Sending $\epsilon, \epsilon' \to 0$ and $n \to \infty$, as $\frac{1}{n}\sum_{i=1}^{n} \rho(X_i, p)^2$ and $j_{\tau_n}(r)$ are bounded above and below, for any subsequence $p_{k_s} \to p$, we have

$$\lim_{s\to\infty} \mathbb{E}\rho(X, p_{k_s})^2 + j_{\tau_n}(r_{k_s}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \rho(X_i, p)^2 + j_{\tau_n}(r).$$

Thus, for all $\omega \in A$,

$$F_n^{\tau_n}(p) \to F(p). \tag{B.5}$$

Next, plugging p_k into p and p into p' in (B.4),

$$\begin{aligned}|F_n^{\tau_n}(p_n) - F_n^{\tau_n}(p)| \le \frac{1}{n}\sum_{i=1}^{n} \left(2\pi(X_i, p) + |\Delta(X_i, p_n, p)|\right) |\Delta(X_i, p_n, p)| \\ + |\rho_{\tau_n}(r') - \rho_{\tau_n}(r)| \to 0.\end{aligned}$$

by (a) and (b). Thus $|F_n^{\tau_n}(p_n) - F(p)| \le |F_n^{\tau_n}(p_n) - F_n(p)| + |F_n^{\tau_n}(p) - F(p)| \to 0$, and

$$F_n^{\tau_n}(p_n) \to F(p) \tag{B.6}$$

for all sequences $p_n \to p$ and $\omega \in A$.

Finally, we assert the theorem.

- When $\cap_{n=1}^{\infty}\overline{\cup_{k=n}^{\infty} E_n^{\tau}} = \phi$: trivial
- When $\cap_{n=1}^{\infty}\overline{\cup_{k=n}^{\infty} E_n^{\tau}} \neq \phi$: Denote $B_n = \cup_{k=n}^{\infty} E_k^{\tau}$. Let $\{b_n \in B_n\}$ be a sequence such that $b_n \to b \in \cap_{n=1}^{\infty}\overline{B_n}$. For each n, we can find an index k_n such that $b_n \in E_{k_n}^{\tau}$. We want to show that $b \in E^{\tau}$.

  By (B.6), $\ell_{k_n} = F_{k_n}(p_{k_n}) \to F(b) \ge \ell$. At the same time, by (B.5), for any arbitrary $q \in \Theta$, there exists a sequence $\epsilon_n \to 0$ such that $F(q) \ge F_n(q) - \epsilon_n \ge \ell_n - \epsilon_n$. Then $\inf_{q\in\Theta} F(q) = \ell \ge \limsup_{n\to\infty} \ell_n$ and we conclude $\ell_n \to \ell = F(b)$.

□

B.1.6 Proof of Proposition 3.2

We start with deriving series expansion of von Mises-Fisher distribution with respect to spherical harmonics. For convenience, let $p = m + 1$, following the conventional notation S^{p-1}.

Proposition B.3. *The density of von Mises-Fisher distribution on* S^{p-1} *with the mean direction* μ *and the concentration parameter* κ *is given with respect to spherical harmonics as*

$$\begin{aligned} f_{vMF}(x^T\mu;\kappa) &= \frac{1}{\omega_p}\sum_{l=0}^{\infty}\sum_{j=1}^{dim\mathcal{H}_l^p} c_l Y_l^m(\mu) Y_l^m(x) \\ &= \frac{1}{\omega_p}\sum_{l=0}^{\infty}\frac{l+\lambda}{\lambda} c_l C_l^{\lambda}(x^T\mu) \\ &= \frac{1}{\omega_p}\sum_{l=0}^{\infty} c_l \sqrt{\frac{(p-2)_l}{l!}}\sqrt{\frac{l+\lambda}{\lambda}} Y_l^0(x^T\mu), \end{aligned}$$

where $\lambda = (p-2)/2$ *and* $\omega_p = 2\pi^{p/2}/\Gamma(p/2)$ *the surface area of* S^{p-1}, $Y_l^0(\cos\theta)$ *the spherical harmonic function of degree* l *and order* 0 *for* $0 \le \theta \le \pi$, *and*

$$c_l = \frac{\sqrt{\pi}}{(p-2)_l 2^{p-3}} \frac{\Gamma(p+l-2)}{\Gamma((p-1)/2)\Gamma((p-2)/2)} \frac{I_{(p-2)/2+l}(\kappa)}{I_{p/2-1}(\kappa)}.$$

Proof. The probability density function of the von Mises-Fisher distribution for the random p-dimensional unit vector x is given by

$$f_{\mathrm{vMF}}(x^T\mu;\kappa) = C_p(\kappa)\mathrm{exp}(\kappa\mu^T x),$$

with a concentration parameter $\kappa > 0$ and a pole $\mu \in S^{p-1}$. The normalization constant $C_p(\kappa)$ is given by

$$C_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)},$$

where I_ν denotes the modified Bessel function of the first kind at order ν.

By the Funk-Hecke formula (Dai and Xu,2013), we have

$$\int_{S^{p-1}} f_{\mathrm{vMF}}(x^T\mu;\kappa)Y_l^m(x)\mathrm{d}\sigma(x) = c_l Y_l^m(\mu),$$

where

$$c_l = \frac{\omega_{p-1}}{C_l^{(p-2)/2}(1)} \int_{-1}^{1} f_{\mathrm{vMF}}(t;\kappa)C_l^{(p-2)/2}(t)(1-t^2)^{\frac{p-3}{2}}\,\mathrm{d}t.$$

We express c_l with respect to the modified Bessel function

$$c_l = \frac{\sqrt{\pi}}{(p-2)_l 2^{p-3}} \frac{\Gamma(p+l-2)}{\Gamma((p-1)/2)\Gamma((p-2)/2)} \frac{I_{(p-2)/2+l}(\kappa)}{I_{p/2-1}(\kappa)}.$$

using the following lemma.

Lemma B.1. *We have the following formula involving exponential function and Gegenbauer polynomial (Dai and Xu,2013),*

$$\int_{-1}^{1} e^{\kappa t}C_l^{(p-2)/2}(t)(1-t^2)^{(p-3)/2}\,dt = \frac{4\pi}{2^{p/2}l!}\frac{\Gamma(p+l-2)}{\Gamma((p-2)/2)}\kappa^{-(p-2)/2}I_{(p-2)/2+l}(\kappa).$$

Proof. For the formula 7.321 in Gradshteyn and Ryzhik (2014, p.805),

$$\int_{-1}^{1} (1-x^2)^{\nu-1/2}e^{i\alpha x}C_n^\nu(x)\,\mathrm{d}x = \frac{\pi 2^{1-\nu} i^n \Gamma(2\nu+l)}{l!\Gamma(\nu)}\alpha^{-\nu}J_{\nu+l}(\alpha)$$

for $\mathfrak{Re}\,\nu > -1/2$, put $\nu = (p-2)/2$ and $\alpha = -i\kappa$. Here J_ν is the Bessel function of the first kind at order ν. Then, use the identity 9.6.3 in Abramowitz and Stegun (1972, p.375)

$$e^{\pm\nu\pi i/2}I_\nu(z) = J_\nu(ze^{\pm\pi i/2}).$$

□

Then using the addition formula of spherical harmonics, we have

$$\begin{aligned}
f_{\text{vMF}}(x^T\mu;\kappa) &= \frac{1}{\omega_p}\sum_{l=0}^{\infty}\sum_{j=1}^{\dim\mathcal{H}_l^p} c_l Y_l^m(\mu)Y_l^m(x) \\
&= \frac{1}{\omega_p}\sum_{l=0}^{\infty}\frac{l+\lambda}{\lambda}c_l C_l^{\lambda}(x^T\mu) \\
&= \frac{1}{\omega_p}\sum_{l=0}^{\infty}c_l\sqrt{\frac{(p-2)_l}{l!}}\sqrt{\frac{l+\lambda}{\lambda}}Y_l^0(x^T\mu),
\end{aligned}$$

with $\lambda = (p-2)/2$ and ω_p, the surface area of S^{p-1}.

$\square$

Remark B.1. *In Mardia and Jupp (1999, p.168), the normalizing constant is given as*

$$C_p(\kappa) = \left(\frac{\kappa}{2}\right)^{p/2-1}\frac{1}{\Gamma(p/2)I_{p/2-1}(\kappa)}.$$

Remark B.2. *When $p = 3$, this result coincides with Jammalamadaka and Terdik (2019) that*

$$f_{vMF}(x^T\mu;\kappa) = \sum_{l=0}^{\infty}\frac{\sqrt{2l+1}}{4\pi}\frac{I_{l+1/2}(\kappa)}{I_{1/2}(\kappa)}Y_l^0(x^T\mu).$$

The proof of Proposition 3.2 is as follows.

Proof.

$$\begin{aligned}
\mathbb{E}\theta_v &= \int_{S^{p-1}} \arccos(v^T x) C_p(\kappa) \exp(\kappa\mu^T x)\, \mathrm{d}\sigma(x) \\
&= \int_{S^{p-1}} \arccos(v^T x) \sum_{l=0}^{\infty} \frac{1}{\omega_p} \frac{l+\lambda}{\lambda} c_l \sqrt{\frac{(p-2)_l \frac{p-2}{2}}{l!(l+\frac{p-2}{2})}}\, Y_l^0(x^T\mu)\, \mathrm{d}\sigma(x) \\
&= \sum_{l=0}^{\infty} \frac{1}{\omega_p} \frac{l+\lambda}{\lambda} c_l \sqrt{\frac{(p-2)_l \frac{p-2}{2}}{l!(l+\frac{p-2}{2})}} \int_{S^{p-1}} \arccos(v^T x) Y_l^0(x^T\mu)\, \mathrm{d}\sigma(x) \\
&= \sum_{l=0}^{\infty} A(l) Y_l^0(v^T\mu) \frac{\omega_{p-1}}{C_l^{(p-2)/2}(1)} \int_{-1}^{1} \arccos(t) C_l^{(p-2)/2}(t)(1-t^2)^{(p-3)/2}\, \mathrm{d}t,
\end{aligned}$$

where

$$A(l) = \sum_{l=0}^{\infty} \frac{1}{\omega_p} \frac{l+\lambda}{\lambda} c_l \sqrt{\frac{(p-2)_l \frac{p-2}{2}}{l!(l+\frac{p-2}{2})}}.$$

□

Remark B.3. *We compute the integral term*

$$\int_{-1}^{1} \arccos(t) C_l^{(p-2)/2}(t)(1-t^2)^{(p-3)/2}\, dt$$

numerically using the following facts.

When $l = 0$, as $C_0^{(p-2)/2} = 1$, we use the following result.

$$\int_{-1}^{1} \arccos(t)(1-t^2)^{(p-3)/2}\, dt = \begin{cases} \frac{((p-3)!!)^2}{(p-2)!}, & p \text{ odd} \\ \frac{((p-3)!!)^2}{2(p-2)!}, & p \text{ even.} \end{cases}$$

When $l > 0$, we use the Maclaurin series of arccos on the interval $[-1, 1]$,

$$\arccos t = \frac{\pi}{2} - \sum_{n=0}^{\infty} \frac{\begin{pmatrix} 2n \\ n \end{pmatrix}}{4^n(2n+1)} t^{2n+1}, \qquad |t| \le 1.$$

It can be shown from integration by parts that

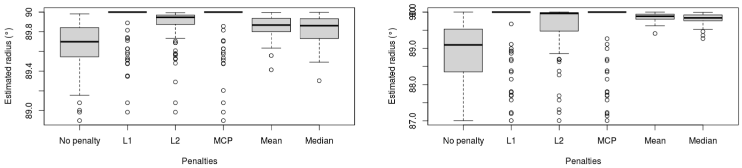
$$\int_{-1}^{1} t^k C_l^\lambda(t)(1-t^2)^{\lambda-1/2}\,dt \begin{cases} >0 & \textit{if } k \geq l \ \textit{ and } k-l \textit{ is even} \\ =0 & \textit{otherwise} \end{cases}$$

and from the formula 7.311.2 in Gradshteyn and Ryzhik (2014, p.802),

$$\int_{-1}^{1} t^{l+2\rho} C_l^\lambda(t)(1-t^2)^{\lambda-1/2}\,dt = \frac{\Gamma(2\lambda+l)\Gamma(2\rho+l+1)\Gamma(\lambda+1/2)\Gamma(\rho+1/2)}{2^l l!\Gamma(2\lambda)\Gamma(2\rho+1)\Gamma(l+\lambda+\rho+1)}$$

B.2 Additional Information on Simulation Study

(a) $r = 90°$, $t_1 = 90°$, $\kappa = 100$ (b) $r = 90°$, $t_1 = 60°$, $\kappa = 100$



(c) $r = 90°$, $t_1 = 90°$, $\kappa = 30$ (d) $r = 90°$, $t_1 = 60°$, $\kappa = 30$

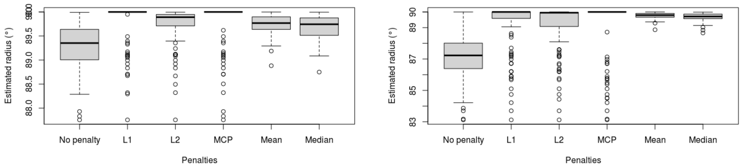


Figure B.2.1: Results of radius estimation when $d = 2$, $r = 90°$ and $n = 1000$.

		Original PNS	L_1	L_2	MCP	Mean (Oracle)	Median (Oracle)
$r = 45°$, $t_1 = 90°$, $\kappa = 100$, $n = 100$	mean (sd)	45.07 (1.44)	45.81 (1.94)	45.81 (1.90)	45.07 (1.44)	45.31 (0.53)	45.34 (0.67)
$r = 45°$, $t_1 = 60°$, $\kappa = 100$, $n = 100$	mean (sd)	44.25 (3.57)	46.96 (5.72)	47.16 (5.81)	44.25 (3.57)	45.33 (0.53)	45.39 (0.60)
$r = 45°$, $t_1 = 90°$, $\kappa = 30$, $n = 100$	mean (sd)	45.49 (3.04)	46.89 (3.72)	47.08 (3.93)	45.49 (3.04)	46.01 (0.97)	46.11 (1.21)
$r = 45°$, $t_1 = 60°$, $\kappa = 100$, $n = 1000$	mean (sd)	45.26 (0.41)	45.49 (0.48)	45.49 (0.49)	45.26 (0.41)	45.30 (0.17)	45.30 (0.22)

Table B.2.1: Results of radius estimation when $d = 2$ and $r = 45°$.

		Original PNS	L_1	L_1 + IoD	MCP	MCP + IoD	Mean (Oracle)	Median (Oracle)
$r = 5°$, $t_1 = 90°$	mean (sd)	8.06 (0.47)	8.15 (0.50)	82.73 (23.24)	8.06 (0.47)	83.46 (22.30)	8.43 (0.43)	8.01 (0.55)
$\kappa = 100$, $n = 100$	$\widehat{r} = \pi/2$	0	0	**91**	0	**92**	0	0
$r = 90°$, $t_1 = 10°$	mean (sd)	9.02 (0.65)	14.27 (19.06)	87.81 (11.57)	12.27 (15.96)	89.19 (8.09)	90.09 (0.53)	90.20 (0.60)
$\kappa = 100$, $n = 100$	$\widehat{r} = \pi/2$	0	5	**94**	4	**99**	0	0

Table B.2.2: Results on mitigating type 2 overfitting phenomenon.

B.3 Additional Information on Real Data Analysis

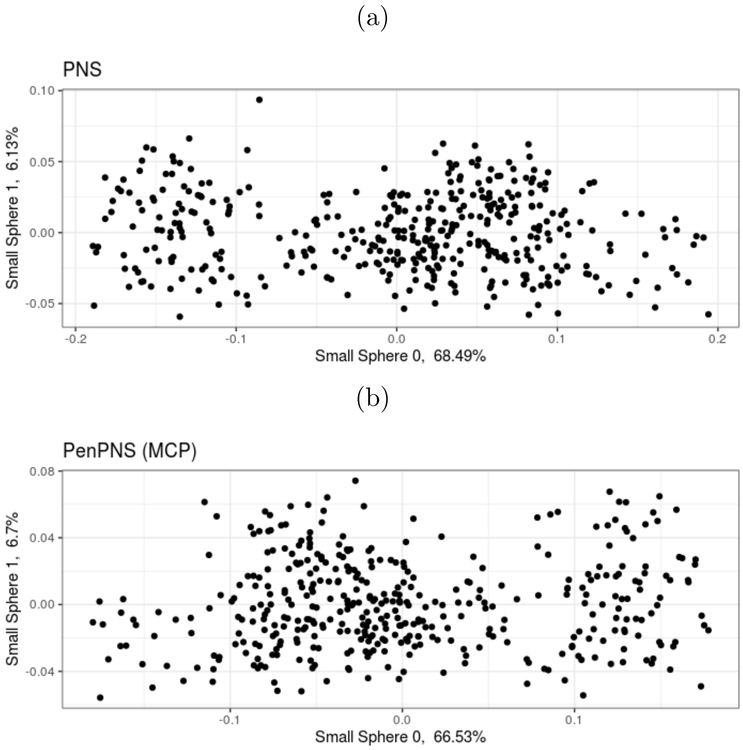
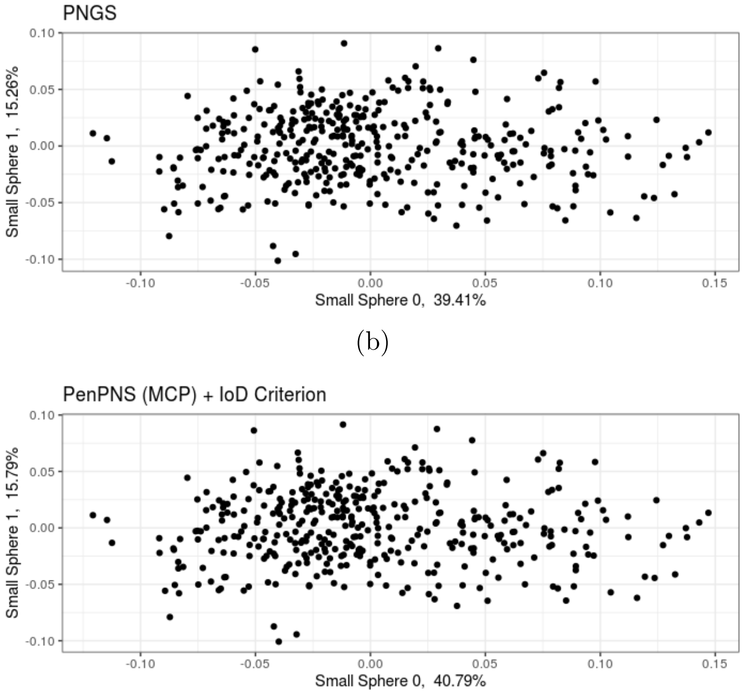


Figure B.3.1: Cephalometric X-ray Image Data : Data points in the coordinates by PNS (top) and PenPNS (MCP) (bottom). The number means the percent variance explained.



(a)

(b)

Figure B.3.2: Cephalometric X-ray Image Data : Data points in the coordinates by PNGS (top) and PenPNS (MCP) + IoD Criterion (bottom). The number means the percent variance explained.

Table B.3.1: Results of the Original PNS on Cephalometric X-ray Image Data. *SSR* stands for 'sum of squared residuals'. *Size* is a size (radius) of $\mathfrak{A}_m$ in $\mathbb{R}^{d+1}$.

Dimension	SSR	Proportion (%)	Radius (°)	Size in $\mathbb{R}^{d+1}$
35→34	9.28e-07	0	89.616	1
34→33	6.438e-06	0.02	86.7994	0.99998
33→32	4.248e-06	0.04	55.9552	0.99842
32→31	1.3369e-05	0.55	15.4737	0.82729
31→30	8.592e-06	0.1	89.3343	0.22072
30→29	1.041e-05	0.11	89.3539	0.2207
29→28	1.3123e-05	0.16	87.0537	0.22069
28→27	1.4962e-05	0.19	86.5	0.2204
27→26	2.7532e-05	0.19	85.1729	0.21999
26→25	2.1458e-05	0.24	89.4131	0.21921
25→24	2.2423e-05	0.27	79.8502	0.21919
24→23	2.4921e-05	0.28	83.2425	0.21576
23→22	2.7892e-05	0.3	87.8096	0.21426
22→21	2.9107e-05	0.33	84.726	0.21411
21→20	3.3793e-05	0.35	84.2716	0.2132
20→19	3.7577e-05	0.37	76.3937	0.21214
19→18	9.0956e-05	0.43	83.8254	0.20618
18→17	5.0163e-05	0.46	85.9081	0.20499
17→16	6.1442e-05	0.57	68.0432	0.20446
16→15	6.0078e-05	0.55	80.5524	0.18963
15→14	6.717e-05	0.58	77.1097	0.18706
14→13	8.2036e-05	0.81	61.6184	0.18235
13→12	9.1071e-05	0.72	82.6931	0.16043
12→11	0.00010021	0.83	75.1032	0.15913
11→10	0.000111185	1.04	65.447	0.15378
10→9	0.000142635	1.04	72.4487	0.13987
9→8	0.000121618	1.2	71.6694	0.13336
8→7	0.000188541	1.4	71.1073	0.1266
7→6	0.000239954	1.46	74.0747	0.11978
6→5	0.000255995	1.69	79.4723	0.11518
5→4	0.000332435	2.28	78.773	0.11324
4→3	0.000345227	2.82	73.0055	0.11107
3→2	0.000461081	4.01	74.0964	0.10622
2→1	0.00097963	6.13	37.3427	0.10216
1→0	0.009077118	68.49	.	0.06197
Sum	**0.01315532**	100	.	.

Table B.3.2: Results of PenPNS (MCP) on Cephalometric X-ray Image Data

Dimension	SSR	Proportion (%)	Radius (°)	Size in $\mathbb{R}^{d+1}$
35→34	9.29e-07	0	90	1
34→33	0.000165633	0.02	90	1
33→32	0.000407293	0.03	90	1
32→31	0.000383708	0.07	90	1
31→30	0.000166148	0.12	90	1
30→29	0.000118163	0.13	90	1
29→28	0.000136102	0.19	90	1
28→27	0.000122323	0.23	90	1
27→26	0.000126636	0.22	90	1
26→25	0.000127404	0.28	90	1
25→24	0.000120823	0.31	90	1
24→23	9.2e-05	0.33	90	1
23→22	7.9903e-05	0.34	90	1
22→21	5.0067e-05	0.37	90	1
21→20	5.5969e-05	0.39	90	1
20→19	5.45e-05	0.43	90	1
19→18	6.8102e-05	0.49	90	1
18→17	6.9583e-05	0.52	90	1
17→16	7.5542e-05	0.58	90	1
16→15	8.1184e-05	0.61	90	1
15→14	8.0205e-05	0.64	90	1
14→13	9.2151e-05	0.78	90	1
13→12	9.7528e-05	0.8	90	1
12→11	0.000107125	0.9	90	1
11→10	0.00011202	0.95	90	1
10→9	0.000124998	1.07	90	1
9→8	0.000130583	1.21	90	1
8→7	0.000213865	1.46	90	1
7→6	0.000225034	1.67	90	1
6→5	0.000234899	1.94	90	1
5→4	0.000307102	2.32	90	1
4→3	0.000345819	3.32	8.8589	1
3→2	0.000433961	4.08	90	0.154
2→1	0.000905904	6.7	22.2347	0.154
1→0	0.002882381	66.53	.	0.05827
Sum	**0.008795588**	100	.	.

Table B.3.3: Results of PNGS on Cephalometric X-ray Image Data

Dimension	SSR	Proportion (%)	Radius (°)	Size in $\mathbb{R}^{d+1}$
35→34	9.4e-08	0	90	1
34→33	2.193e-06	0.03	90	1
33→32	5.238e-06	0.08	90	1
32→31	2.3826e-05	0.35	90	1
31→30	3.0398e-05	0.45	90	1
30→29	3.2492e-05	0.48	90	1
29→28	3.6514e-05	0.53	90	1
28→27	3.7737e-05	0.55	90	1
27→26	3.6922e-05	0.54	90	1
26→25	4.288e-05	0.63	90	1
25→24	4.4287e-05	0.65	90	1
24→23	4.4848e-05	0.66	90	1
23→22	4.6415e-05	0.68	90	1
22→21	4.8862e-05	0.72	90	1
21→20	5.005e-05	0.73	90	1
20→19	5.2152e-05	0.76	90	1
19→18	5.9216e-05	0.87	90	1
18→17	6.3302e-05	0.93	90	1
17→16	6.7462e-05	0.99	90	1
16→15	7.1574e-05	1.05	90	1
15→14	7.4471e-05	1.09	90	1
14→13	8.8939e-05	1.3	90	1
13→12	9.0918e-05	1.33	90	1
12→11	0.000102183	1.5	90	1
11→10	0.00010737	1.57	90	1
10→9	0.000119884	1.76	90	1
9→8	0.000134624	1.97	90	1
8→7	0.000161609	2.37	90	1
7→6	0.000184337	2.7	90	1
6→5	0.000211901	3.1	90	1
5→4	0.0002532	3.71	90	1
4→3	0.000328375	4.81	90	1
3→2	0.000439975	6.45	90	1
2→1	0.001041575	15.26	90	1
1→0	0.002689564	39.41	.	1
Sum	**0.006825388**	100	.	.

Table B.3.4: Results of PenPNS (MCP) + IoD Criterion on Cephalometric X-ray Image Data

Dimension	SSR	Proportion (%)	Radius (°)	Size in $\mathbb{R}^{d+1}$
35→34	9.4e-08	0	90	1
34→33	2.192e-06	0.03	90	1
33→32	3.349e-06	0.05	90	1
32→31	7.641e-06	0.12	90	1
31→30	1.252e-05	0.19	90	1
30→29	1.4332e-05	0.22	90	1
29→28	2.0421e-05	0.31	90	1
28→27	2.5036e-05	0.38	90	1
27→26	2.4269e-05	0.37	90	1
26→25	3.0032e-05	0.46	90	1
25→24	3.3803e-05	0.51	90	1
24→23	3.5842e-05	0.54	90	1
23→22	3.7117e-05	0.56	90	1
22→21	4.0047e-05	0.61	90	1
21→20	4.234e-05	0.64	90	1
20→19	4.6403e-05	0.7	90	1
19→18	5.2768e-05	0.8	90	1
18→17	5.6474e-05	0.86	90	1
17→16	6.2659e-05	0.95	90	1
16→15	6.593e-05	1	90	1
15→14	6.8979e-05	1.05	90	1
14→13	8.4216e-05	1.28	90	1
13→12	8.6101e-05	1.31	90	1
12→11	9.7564e-05	1.48	90	1
11→10	0.000102898	1.56	90	1
10→9	0.000115491	1.75	90	1
9→8	0.000130464	1.98	90	1
8→7	0.000157646	2.39	90	1
7→6	0.00018072	2.74	90	1
6→5	0.000209631	3.18	90	1
5→4	0.000250704	3.8	90	1
4→3	0.000326408	4.95	90	1
3→2	0.00043827	6.65	90	1
2→1	0.001040521	15.79	90	1
1→0	0.002688727	40.79	.	1
Sum	**0.006591607**	100	.	.

Bibliography

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas.* Dover Publications, Inc..

Jushan Bai and Serena Ng (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrika*, **70(1)**, 191-221.

Yoav Benjamini and Yosef Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B*, **57(1)**, 289-300.

Rabi Bhattacharya and Vic Patrangenaru (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. *Annals of Statistics*, **31(1)**, 1-29.

W.M. Boothby (1993). *An introduction to Differentiable Manifolds.* Academic Press.

A. Björck and G.H. Golub (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, **27(123)**, 579-594.

G. Bolt and S. Lunag'omez and C. Nemeth (2022). Distances for comparing multisets and sequences. *ArXiv*:2206.08858v1.

David R. Cox (1966). *The Statistical Analysis of Series of Events*, 1st ed. Springer.

Feng Dai and Yuan Xu. *Approximation Theory and Harmonic Analysis on Spheres and Balls*. 1st edition. Springer.

Dietrich, S. and Oleś, M. and Lu, J and Sellner, L and et al. (2018). Drug-perturbation-based stratification of blood cancer. *Journal of Clinical Investigation*, **128(1)**, 427-445.

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

Ian L. Dryden and Kantilal Vardichand Mardia (1993). *Statistical Shape Analysis*. Wiley.

Ian L. Dryden and Kantilal Vardichand Mardia (2016). *Statistical Shape Analysis: With Applications in R*. 2nd Ed. Wiley.

Carl Eckart and Gale Young (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211-218.

P. Thomas Fletcher and Conglin Lu and Stephen M. Pizer and Sarang Joshi (2004). Principal geodesic analysis for the study of nonlinear statistics of shape . *IEEE Transations on edical Imaging*, **23(8)**, 995-1005.

Qing Feng and Meilei Jiang and J. Hannig and J.S. Marron (2018). Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 241-265.

Ronald Aylmer Fisher (1953). Dispersion on a sphere. *Proc. R. Soc. Lond. A.*, **217 (1130)**, 295-305.

N.I. Fisher (1993). *Statistical Analysis of Circular Data.* Cambridge University Press.

N.I. Fisher and T. Lewis and B.J.J. Embleton (1993). *Statistical Analysis of Spherical Data.* Cambridge University Press.

Maurice Fréchet (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincare*, **10**, 215-310. Presses universitaires de France.

Xing Gao and Sungwon Lee and Sungkyu Jung (2021). Covariate-driven factorization by thresholding for multi-block data. *Biometrics*, **77(3)**, 1011-1023.

Irina Gaynanova and Gen Li (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics*, **75**, 1121–1132.

J.C. Gower (1975). Generalized procrustes analysis. *Psychometrika*, **40**, 33-51.

Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of Integrals, Series, and Products.* 8th edition. Elsevier Inc.

Trevor Hastie. *Principal curves and surfaces.* PhD Thesis, Stanford University.

Harold Hotelling (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417-441, 498-520.

S. Huckemann, H. Ziezold (2006). Principal component analysis for Riemannian manifolds with an application to triangular shape spaces,. *Adv. Appl. Prob. (SGSA)*, **38(2)**, 299-319.

S. Huckemann, T. Hotz, A. Munk (2010). Intrinsic shape analysis: Geodesic principal component analysis for Riemannian manifolds modulo Lie group actions (with discussion). *Statist. Sinica*, **20(1)**, 1-100.

Stephan F. Huckemann (2011). Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *The Annals of Statistics*, **39(2)**, 1098-1124.

H. Karcher (1977). Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.*, **XXX**, 509-541.

David George Kendall (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. London Math. Soc.*, **16**, 81-121.

Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of Differential Geometry*, Vol. II. Wiley, Chichester.

S. Rao Jammalamadaka and György H. Terdik (2019). Harmonic analysis and distribution-free inference for spherical distributions. *Journal of Multivariate Analysis*, **171**, 436-451.

Sungkyu Jung (2012). Analysis of principal nested spheres. *Biometrika*, **99(3)**, 551–568.

Sungkyu Jung and Myung Hee Lee and Jeongyoun Ahn (2018). On the number of principal components in high dimensions. *Biometrika*, **105(2)**, 389-402.

J.M. Lee (2006). *Riemannian Manifolds: An Introduction to Curvature*, volume 176. Springer Science & Business Media.

Kenneth Levenberg (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly of Applied Mathematics*, **2(2)**, 164-168.

Gen Li and Sungkyu Jung (2017). Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, **72**, 1433-1442.

Gen Li and Irina Gaynanova (2018). A general framework for association analysis of heterogeneous data. *The Annals of Applied Statistics*, **12(3)**, 1700-1726.

E.F. Lock and K.A. Hoadley and J.S. Marron and A.B. Nobel (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Application Statistics*, **7(1)**, 523-542.

Kanti V. Mardia and Peter E. Jupp (1999). *Directional Statistics.* 1st ed. John Wiley & Sons, Inc..

Kanti V. Mardia and Henrik Wiechers and Benjamin Eltzner and Stephan F. Huckemann (2013). Principal component analysis and clustering on manifolds. *Journal of Multivariate Analysis*, **188**, 104862.

Donald Marquardt (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics*, **11(2)**, 431–441.

Mshaik, R. and Simonet, J. and Georgievski, A. and et al. (2021). HSP90 inhibitor NVP-BEP800 affects stability of SRC kinases and growth of T-cell and B-cell acute lymphoblastic leukemias. *Blood Cancer Journal*, **11(61)**, doi:10.1038/s41408-021-00450-2, url:https://www.nature.com/articles/s41408-021-00450-2.

Jeremy L. Norris and Melissa A. Farrow and Danielle B. Gutierrez et al. (2017). Integrated, High-Throughput, Multiomics Platform Enables Data-Driven Construction of Cellular Responses and Reveals Global Drug Mechanisms of Action. *Journal of Proteome Research*, **16(3)**, 1364–1375.

Damien Passemier and Jian-Feng Yao (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *Journal of Multivariate Analysis*, **127**, 173-183.

Feng Qi (2010). Bounds for the Ratio of Two Gamma Functions. *Journal of Inequalities and Applications*, 493058. https://doi.org/10.1155/2010/493058

Jason A. Reuter and Damek Spacek and Michael P. Snyder (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, **58(4)**, doi:10.1016/j.molcel.2015.05.004.

R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis.* 1st ed. Springer.

Johannes O. Royset (2018). Approximations and solution estimates in optimization. *Mathematical Programming: Series A and B*, **170(2)**, 479-506.

Christof Schotz (2022). Strong laws of large numbers for generalizations of Frechet mean sets. *A Journal of Theoretical and Applied Statistics*, **56(1)**, 34-52.

Age K. Smilde and Johan A. Westerhuis and Sijmen de Jong (2003). A framework for sequential multiblock component methods. *Journal of Chemometrics*, **17**, 323-337.

I. Subramanian and S. Verma and S. Kumar and A. Jere and K. Anamika (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, doi:10.1177/1177932219899051, url:https://journals.sagepub.com/doi/full/10.1177/1177932219899051.

Katrijn VanDeun and Age K Smilde and Mariet J van der Werf and Henk AL Kiers and Iven VanMechelen (2009). A structured overview of simultaneous component based data integration. *BMC Bioinformatics*, **10, 246**.

Chunxin Wang and Richard J. Youle (2014). The Role of Mitochondria in Apoptosis. *Annu Rev Genet.*, **43**, 95–118.

Ching-Wei Wang and Cheng-Ta Huang and Jia-Hong Lee and Chung-Hsing Li and Sheng-Wei Chang and Ming-Jhih Siao and

Tat-Ming Lai and Bulat Ibragimov and Tomaž Vrtovec and Olaf Ronneberger and Philipp Fischer and Tim F. Cootes and Claudia Lindner (2016). A benchmark for comparison of dental radiography analysis algorithms. *Medical Image Analysis*, **31**, 63–76.

Ke Ye and Lek-Heng Lim (2016). Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal of Matrix Analysis and Applications*, **37(3)**, 1176-1197.

Cun-Hui Zhang (2010). Nearly unbiased varialbe selection under minimax concave penalty. *The Annals of Statistics*, **38(2)**, 894-942.

Youwei Zhang and Tony Hunter (2014). Roles of Chk1 in Cell Biology and Cancer Therapy. *International Journal of Cancer*, **134(5)**, 1013-1023.

Herbert Ziezold (1977). On Expected Figures and a Strong Law of Large Numbers for Random Elements in Quasi-Metric Spaces. *Transaction of the 7th Prague Conference on Information Theory, Statistical Decision Function and Random Processes A*, 591–602.

국문초록

본 학위논문에서는 비유클리드 공간에서의 차원축소법에 대해 다룬다. 비유클리드 공간에선 곡률로 인해 피타고라스 정리와 같은 유클리드 공간에서 널리 사용되는 방법론을 활용하기 힘들다. 자료의 구조를 파악하기 위해 주어진 비유클리드 공간의 기하학적 성질을 이해하는 것이 중요하다. 이 논문에서는 주성분분석이나 인자분석과 같이 널리 사용되는 다변량 자료분석을 비유클리드 공간에 일반화하고자 다음 두 방법론을 제안한다.

제 2장에서는 다중 근원 자료를 분석하고자 Principal Structure Identification (PSI) 방법론을 제안한다.

다중 근원 자료는 같은 관찰 대상으로부터 여러 변수 집단에 대해 채취된 자료이다. 요즘 각광받는 다중 오믹스 자료가 좋은 예이다. 다중 근원 자료를 통합적인 관점에서 분석하기 위해, 장료들 사이의 공통 스코어가 있는지 확인하도록 한다. 이때 공통 스코어는 모든 자료 집단들이 공유하는 것일수도, 일부 자료집단들이 공유하는 것일수도, 혹은 하나의 자료집단에만 귀속되는 것일수도 있다.

우리 방법론이 가지는 가장 큰 특징 중 하나는, 각 자료가 가지는 인자 스코어 선형부공간을 자료 간 연관관계를 밝히는 기하학적 기본 요소로 사용하는 것이다. 공통 스코어 선형부공간을 알아내는데 있어 잡음으로부터 생기는 부정확함을 회피하고자, 자료집단의 인자

스코어 선형부공간의 1차원 깃발 공간을 계산한다. 이 1차원 깃발 공간과 근접한 선형부공간들을 골라내어 스코어를 공유하는 자료집단을 묶어주는 절차적 알고리즘을 제안한다.

제 3장에서는 초구면체 위의 자료를 분석하는 Penalized Principal Nested Spheres (PenPNS) 방법론을 제안한다.

Analysis of Principal Nested Spheres (Jung, 2012)는 초구면체 위의 자료를 분석하는 차원축소 방법론이다. PNS에서는 차원축소가 절차적 과정을 통해 이루어지며, 불필요한 차원이 하나씩 축소된다. 특히 작은 구로 자료를 가늠하여 비측지선 추정을 가능케한 것이 장점이다.

하지만 PNS에게는 과적합이란 단점이 있다. 과적합이란 큰 구를 따라 생성된 데이터가 지나치게 작은 구로 추정되는 현상을 말한다. 우리는 두 종류의 과적합을 다룬다. (1) 모수로서 반지름은 0에서 $\pi/2$까지의 범위를 가지게 되므로, 추정된 반지름이 $\pi/2$보다 큰 경우 추정된 축은 구면의 반대편의 것으로 대체되며 반지름 추정도 그에 따른다. 따라서, 추정된 반지름의 분포는 실제 축과 점들 사이의 각도의 분포가 반으로 접힌 모양을 가지게 된다. 이와 같은 이유로 반지름 추정의 값은 $\pi/2$보다 작을 수 밖에 없다. (2) 만약 자료가 큰 구를 따라 생성되었지만 매우 짧은 구간에서 생성되는 경우, 자료점 구름의 모양은 접시 모양과 가까워지게 된다. 이 경우 접시 모양 구름 안에 들어있는 반지름이 아주 작은 작은 구로 자료가 추정된다.

PenPNS는 과적합 현상을 교정하는 PNS의 개량형이다. 첫번째 과적합 현상을 교정하기 위해, PenPNS는 추정 과정에서 반지름을 정규화한다. 이때 반지름이 $\pi/2$보다 작아질수록 벌점화 항목의 값이 커진다. 두번째 과적합 현상을 교정하기 위해, PenPNS는 교차 검증 오차에 벌점화 항목을 반영한다. 이 벌점화 항목은 퍼짐지표에서 유래하는데, 퍼짐지표는 접시모양 데이터 구름이 작은 반지름의 구로

추정되는 상황에서 그 값이 커진다. 이와 같은 과적합 해결 방법의 효과를 모의실험과 실제 데이터 분석을 통해 보이겠다.

주요어 : 차원축소, 비유클리드 공간, 다중 근원 자료, 초구면체 위 자료, 멀티 오믹스 자료.

학 번 : 2017 – 38449