On the Multiway Principal Component Analysis

Jialin Ouyang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

On the Multiway Principal Component Analysis

Jialin Ouyang

Multiway data are becoming more and more common. While there are many approaches to extending principal component analysis (PCA) from usual data matrices to multiway arrays, their conceptual differences from the usual PCA, and the methodological implications of such differences remain largely unknown. This thesis aims to specifically address these questions. In particular, we clarify the subtle difference between PCA and singular value decomposition (SVD) for multiway data, and show that multiway principal components (PCs) can be estimated reliably in absence of the eigengaps required by the usual PCA, and in general much more efficiently than the usual PCs. Furthermore, the sample multiway PCs are asymptotically independent and hence allow for separate and more accurate inferences about the population PCs. The practical merits of multiway PCA are further demonstrated through numerical, both simulated and real data, examples.

Table of Contents

List of Figures

Acknowledgements

First and foremost, I would like to thank my advisor Prof. Ming Yuan. Throughout my four years as his student, he has been consistently offering me insightful and detailed guidance on problem identification, problem-solving, and presentation. I also learned a lot from his dedicated work ethic. His lessons will continue to inspire and motivate me throughout my life.

I am very grateful to Prof. Arian Maleki, Prof. Richard Davis, Prof. Marco Avella and Prof. Kaizheng Wang for serving on the dissertation committee. Prof. Arian Maleki and Prof. Richard Davis taught my first-year statistical inference and probability classes, which laid a solid foundation for my research. Prof. Arian Maleki and Prof. Richard Davis also served on my oral exam committee and offered encouragement and insightful comments.

My fellow Ph.D. students in the Statistics Department have formed a fun and helpful community, which has made my five-year journey much more pleasurable. I especially want to thank my fellow Ph.D. student and roommate, Dr. Long Zhao. We have had countless fruitful conversations, both technical and non-technical, that have helped me in and out of my research.

Finally, I want to thank my parents and my close friends, for their understanding, encouragement, and support throughout all my ups and downs. They always lend me their attentive ears whenever I need them, and I cannot be too grateful for that.

To my parents.

Chapter 1: Introduction

More and more often in practice, we need to deal with data of rich and complex structures that are more appropriately organized as multiway arrays rather than the usual data matrices. Examples of such multiway data are ubiquitous in many fields such as chemometrics, economics, psychometrics, and signal processing among others (see, e.g., Kroonenberg 2008; Animashree Anandkumar et al. 2014; A. Zhang and Xia 2018; E. Y. Chen, Fan, and E. Li 2020; E. Y. Chen, Xia, et al. 2020; Y. Han, R. Chen, et al. 2020; Xia, A. R. Zhang, and Y. Zhou 2020; Bi et al. 2021; R. Chen, D. Yang, and C.-H. Zhang 2021; R. Han, Willett, and A. R. Zhang 2022). In this thesis, we investigate the methodological implications and statistical properties of principal component analysis (PCA) for this type of data and pinpoint the benefits and challenges of doing so.

PCA is among the most popular statistical methods for multivariate data analysis when data are organized as matrices. See, e.g., Anderson 1984; Jolliffe 2002. With each column vector of a data matrix as an observation, PCA seeks orthogonal linear transformations of these vectors into a new coordinate system so that the variance of each coordinate is maximized successively. It allows us to represent most of the variation in the data by a small number of coordinates and therefore can guide us in reducing the dimensionality. As such, PCA often serves as a critical first step to capture the essential features in a dataset for many downstream analyses and is widely used in many scientific and engineering fields. Moving beyond matrices, for multiway data, each observation itself forms a matrix or more generally a multiway array. For example, when repeated measurements are made across different combinations of location and time, each observation can be more naturally organized as a matrix with each row corresponding to a certain location and each column a time point. To apply PCA to this type of data, it is tempting to neglect the multiway nature of the observations and treat each observation as a vector nonetheless, a practice often referred to as *stringing*. However, as observed in numerous practical applications, appropriately accounting

for the additional structure when applying PCA can greatly enhance interpretability and improve efficiency. See, e.g., Kroonenberg 2008.

There is a long and illustrious history of developing suitable methods for such a purpose and it can be traced back at least to the pioneering work of Tucker, Harshman, and Carroll in the 1960s. Since then, numerous approaches have also been developed. Examples include Kroonenberg and De Leeuw 1980; De Lathauwer, De Moor, and Vandewalle 2000; Vasilescu and Terzopoulos 2002; J. Yang et al. 2004; Kong et al. 2005; D. Zhang and Z.-H. Zhou 2005; H. Lu, Plataniotis, and Venetsanopoulos 2006; H. Lu, Plataniotis, and Venetsanopoulos 2008; X. Li, Pang, and Y. Yuan 2010; Liu, M. Yuan, and Zhao 2017; Taguchi 2018 among many others. See, e.g., H. Lu, Plataniotis, and Venetsanopoulos 2011; Cichocki et al. 2015 for recent surveys of existing techniques. Most of these developments are outside the mainstream statistics literature and often with a strong algorithmic flavor and exploratory data analysis focus. These approaches are intuitive and often yield more interpretable insights than naively applying PCA after stringing. However, their statistical underpinnings are largely unknown. The main goal of this article is to fill in this void. Indeed, as we shall demonstrate, a careful and rigorous statistical treatment allows for a better understanding of the operating characteristics of multiway PCA, leads to improved methodology, and reveals new opportunities and challenges in analyzing multiway data.

More specifically, we focus on a simple and natural approach to multiway PCA: when seeking linear transformations that maximize the variance, we impose the additional constraint that they conform to the multiway structure of the data. Doing so not only allows for enhanced interpretability but also inherits many nice properties of the usual PCA. Just as the usual principal components (PCs) are the eigenvectors of the covariance matrix, the multiway PCs can be identified with certain eigenvectors of the covariance operator. To better understand the impact of multiway structure on our ability to recover and make inferences about the multiway PCs, we also investigate the properties of multiway PCA under a spiked covariance model.

Statistical properties of the usual PCA are well understood in the classical setting where the sample size is large whereas the number of variables is small (see, e.g., Anderson 1984). More

and more often in today's applications, however, the dimensionality can also be large. There are abundant theoretical results concerning the usual PCA in such a high-dimensional setting as well, especially in the context of the spiked covariance model. For example, Johnstone and A. Y. Lu 2009 first demonstrated the critical role of dimensionality in PCA by showing that, with fixed signal strength, the sample PCA is consistent if and only if the number of variables is of a smaller order than the sample size. In another influential paper, Paul 2007 established the asymptotic distribution of sample PCs. Other related treatments include Baik and Silverstein 2006; Nadler 2008; Johnstone and A. Y. Lu 2009; Jung and Marron 2009; Bai and Silverstein 2010; Lee, Zou, and Wright 2010; Benaych-Georges and Nadakuditi 2011; Bai and Yao 2012; Shen et al. 2013; Koltchinskii and Lounici 2014; Koltchinskii, Lounici, et al. 2017; Wang and Fan 2017; Koltchinskii, Löffler, and Nickl 2020 among numerous others. In a sense, our results naturally extend these earlier works to multiway PCA. However, the need to work with higher-order covariance operators rather than covariance matrices creates new and fundamental challenges and requires us to develop a different proof strategy and several new technical tools. More importantly, our analysis also reveals fundamental differences in behavior between the usual PCA and multiway PCA and inspires new methodological development for the latter.

Firstly, we establish the rates of convergence for the sample multiway PCs under mild regularity conditions. These rates explain why it is essential that we account for the inherent data structure when applying PCA to multiway data, and why naively applying PCA after stringing could be problematic. Intuitively, multiway PCA uses fewer parameters than the usual PCA and therefore is easier for estimation. This is described precisely by our result in that the estimation error of multiway PCs is determined by the dimension of each mode of the data array rather than the total number of entries, and therefore multiway PCs can be estimated accurately even if the latter far exceeds the sample size. But a more important observation is that how well a multiway PC can be estimated is determined by the corresponding eigenvalue of the covariance operator, and not the gap between its eigenvalues like the usual PCA. This somewhat surprising finding has far-reaching implications. In particular, it means that for multiway data the PCs can be estimated well even if

their corresponding eigenvalues are not simple.

Moreover, to facilitate making statistical inferences about multiway PCs, we derive asymptotic distributions of the sample multiway PCs. Our results again reveal unexpected but important distinctions between multiway PCA and usual PCA. For example, the estimated multiway PCs are asymptotically independent of each other, and their asymptotic distribution is determined by their corresponding eigenvalues instead of eigengaps. Furthermore, we show that bias correction is important for the sample multiway PCs. Similar to the usual PCA, sample multiway PC can exhibit significant bias when the dimension (of each mode) is high. But there is also another source of bias that may arise due to the inherent ambiguity in ordering the PCs in absence of eigengaps. Nonetheless, we show that both types of bias can be eliminated, enabling us to make inferences about and construct confidence intervals for the multiway PCs.

The thesis is organized as follows. In Chapter 2, we introduce the notion of multiway PCA both at a population level and how it works on a finite sample. Chapter 3 investigates the rates of convergence for the sample multiway PCs. Turning our attention to the asymptotic distribution of multiway PCA in Chapter 4, we show how to make valid inferences about the multiway PCs. In Chapter 5, we propose a polynomial time algorithm to approximate the sample PCs, and the merits of the multiway PCA and our proposed approaches are further demonstrated through numerical experiments, both simulated and real. All the main proofs are relegated to Chapter 6. Other technical results and their proofs are gathered in the Appendix.

Chapter 2: Multiway PCA

Multiway PCA can be viewed through the lens of usual PCA with the additional multiway structure imposed on the PCs. Let $\mathscr{X} \in \mathbb{R}^{d_1 \times \cdots \times d_p}$ be an order-p random array. To simplify, we shall assume in what follows that $\mathscr{X}$ is centered, i.e., $\mathbb{E}\mathscr{X} = 0$, unless otherwise indicated. The idea behind PCA is to look for a linear transformation of $\mathscr{X}$ that maximizes the variance:

$$\max_{\mathscr{W} \in \mathbb{R}^{d_1 \times \cdots \times d_p} : \|\mathscr{W}\|_{\mathrm{F}} = 1} \operatorname{var}(\langle \mathscr{X}, \mathscr{W} \rangle). \tag{2.1}$$

Here $\|\mathscr{W}\|_{\mathrm{F}} = \langle \mathscr{W}, \mathscr{W} \rangle^{1/2}$ and

$$\langle \mathscr{X}, \mathscr{W} \rangle = \sum_{j_1=1}^{d_1} \cdots \sum_{j_p=1}^{d_p} x_{j_1,\ldots,j_p} w_{j_1,\ldots,j_p}.$$

Denote by $\mathscr{U}_1$ the solution to (2.1). The basic premise of multiway PCA is that $\mathscr{U}_1$ conforms to the multiway structure underlying $\mathscr{X}$ in that it is a rank-one tensor and can be expressed as

$$\mathscr{U}_1 = \mathbf{u}_1^{(1)} \otimes \mathbf{u}_1^{(2)} \otimes \cdots \otimes \mathbf{u}_1^{(p)}, \tag{2.2}$$

where $\mathbf{u}_1^{(q)}$ is a unit length vector $\mathbb{R}^{d_q}$ and $\otimes$ stands for the outer product, i.e., the $(i_1, \ldots, i_p)$ entry of $\mathscr{U}_1$ is given by

$$[\mathscr{U}_1]_{i_1,\ldots,i_p} = u_{1i_1}^{(1)} u_{1i_2}^{(2)} \cdots u_{1i_p}^{(p)}.$$

In other words, $\mathscr{U}_1$ is also the solution to

$$\max_{\mathscr{W} \in \Theta} \operatorname{var}(\langle \mathscr{X}, \mathscr{W} \rangle), \tag{2.3}$$

where Θ is the collection of all unit length rank-one tensors of conformable dimensions, i.e.,

$$\Theta = \{\mathscr{W} = \mathbf{w}^{(1)} \otimes \mathbf{w}^{(2)} \otimes \cdots \otimes \mathbf{w}^{(p)} : \mathbf{w}^{(q)} \in \mathbb{R}^{d_q}, \|\mathbf{w}^{(q)}\| = 1, \forall q = 1, \ldots, p\}.$$

Even if the solution to (2.1) is not strictly rank-one as described by (2.2), imposing such a constraint when seeking variance-maximizing transformation can nonetheless be desirable because of the enhanced interpretability: the additional rank-one constraint allows us to separate the effect along each mode, and help address questions such as "who does what to whom and when" which are often central to multiway data analysis. See, e.g., Kroonenberg 2008 for further discussion and numerous motivating examples.

Subsequent PCs can be defined successively:

$$\max_{\substack{\mathscr{W} \in \mathbb{R}^{d_1 \times \cdots \times d_p} : \|\mathscr{W}\|_{\mathrm{F}} = 1 \\ \mathscr{W} \perp \mathscr{U}_l, l = 1, \ldots, k-1}} \operatorname{var}(\langle \mathscr{X}, \mathscr{W} \rangle). \tag{2.4}$$

As before, we shall consider the case when the solution has rank one. A key requirement in defining PCs is that the kth PC is orthogonal to all other PCs, i.e., $\mathscr{W} \perp \mathscr{U}_l$. In vector case, i.e., $p = 1$, this simply means that $\langle \mathscr{W}, \mathscr{U}_l \rangle = 0$. In multiway case, however, there are many different notions of orthogonality. See, e.g., Kolda 2001 for a detailed discussion on this subject. Each notion has its own subtleties and caveats that may have different statistical implications. In this work we shall focus on the notion of *complete orthogonality*: two rank-one tensors $\mathscr{W}_1 = \mathbf{w}_1^{(1)} \otimes \mathbf{w}_1^{(2)} \otimes \cdots \otimes \mathbf{w}_1^{(p)}$ and $\mathscr{W}_2 = \mathbf{w}_2^{(1)} \otimes \mathbf{w}_2^{(2)} \otimes \cdots \otimes \mathbf{w}_2^{(p)}$ are complete orthogonal if and only if $\langle \mathbf{w}_1^{(q)}, \mathbf{w}_2^{(q)} \rangle = (\mathbf{w}_1^{(q)})^\top \mathbf{w}_2^{(q)} = 0$ for all $q = 1, \ldots, p$. More specifically, the kth multiway PC, denoted by $\mathscr{U}_k$, solves

$$\max_{\mathscr{W} \in \Theta : \mathscr{W} \perp_c \mathscr{U}_l, \forall l < k} \operatorname{var}(\langle \mathscr{X}, \mathscr{W} \rangle), \tag{2.5}$$

where $\perp_c$ stands for complete orthogonality.

As in the case of the usual PCA, multiway PCs can also be equivalently defined using the covariance matrix of $\operatorname{vec}(\mathscr{X})$. In fact, it is more convenient to think of a covariance operator when

it comes to multiway data. More specifically, we shall view

$$\Sigma := \mathrm{cov}(\mathscr{X}) = \mathbb{E}(\mathscr{X} \otimes \mathscr{X})$$

as a $d_1 \times d_2 \times \cdots \times d_p \times d_1 \times \cdots \times d_p$ array. Then for any $\mathscr{W} \in \Theta$,

$$\mathrm{var}(\langle \mathscr{X}, \mathscr{W} \rangle) = \langle \Sigma, \mathscr{W} \otimes \mathscr{W} \rangle = \langle \Sigma, \mathbf{w}^{(1)} \otimes \cdots \otimes \mathbf{w}^{(p)} \otimes \mathbf{w}^{(1)} \otimes \cdots \otimes \mathbf{w}^{(p)} \rangle.$$

Write

$$\lambda_k = \mathrm{var}(\langle \mathscr{X}, \mathscr{U}_k \rangle).$$

Because of the symmetry of Σ,

$$\mathscr{U}_1 \otimes \mathscr{U}_1 = \mathrm{argmax}_{\mathbf{w}^{(1)} \otimes \cdots \otimes \mathbf{w}^{(2p)} : \|\mathbf{w}^{(q)}\| = 1} \langle \Sigma, \mathbf{w}^{(1)} \otimes \cdots \otimes \mathbf{w}^{(2p)} \rangle$$

so that $\lambda_1 \mathscr{U}_1 \otimes \mathscr{U}_1$ is also the best rank-one approximation to Σ (see, e.g., Friedland 2013). Similarly,

$$\mathscr{U}_k \otimes \mathscr{U}_k = \mathrm{argmax}_{\substack{\mathbf{w}^{(1)} \otimes \cdots \otimes \mathbf{w}^{(2p)} : \|\mathbf{w}^{(q)}\| = 1 \\ \mathbf{w}^{(1)} \otimes \cdots \otimes \mathbf{w}^{(2p)} \perp_c \mathscr{U}_l \otimes \mathscr{U}_l, \quad l < k}} \langle \Sigma, \mathbf{w}^{(1)} \otimes \cdots \otimes \mathbf{w}^{(2p)} \rangle$$

In vector case, e.g. $p = 1$, $\{(\lambda_k, \mathscr{U}_k) : k \geq 1\}$ are the eigenpairs of the covariance matrix Σ and

$$\Sigma_r := \sum_{k=1}^{r} \lambda_k \mathscr{U}_k \otimes \mathscr{U}_k$$

is the best rank-r approximation to Σ, i.e.,

$$\Sigma_r = \mathrm{argmin}_{A \in \mathbb{R}^{d_1 \times d_1} : \mathrm{rank}(A) \leq r} \|A - \Sigma\|.$$

When $p > 1$, this characterization becomes tenuous because the notion of best low-rank approximation becomes precarious. For matrices, best low-rank approximations can be identified with singular value decomposition thanks to the Eckart-Young theorem. Low-rank approximation to

tensors is much more subtle and the best low-rank approximation may not exist in general. See, e.g., Hackbusch 2012. Nonetheless, by construction, Σ_r is the so-called best rank-r *greedy orthogonal approximation* to Σ. See, e.g., Kolda 2001. In particular, when the multiway structure does manifest itself in a way such that the usual PCs are rank-one tensors, for example, the solution to (2.1) and (2.4) has rank one, then Σ_r is the best low-rank approximation to Σ.

Sample multiway PCs can also be defined in a similar fashion. Specifically, given a sample $\mathcal{X}_1, \ldots, \mathcal{X}_n$ of independent copies of $\mathcal{X}$, $\mathcal{U}_k$s can be estimated by maximizing the sample variances:

$$\widehat{\mathcal{U}}_k := \text{argmax}_{\mathcal{W} \in \Theta : \mathcal{W} \perp_c \mathcal{U}_l, \forall l < k} \frac{1}{n} \sum_{i=1}^{n} \langle \mathcal{X}_i, \mathcal{W} \rangle^2. \tag{2.6}$$

Let

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_i \otimes \mathcal{X}_i$$

be the sample covariance operator. Then $\widehat{\mathcal{U}}_1$ can be defined via the best rank-one approximation to $\widehat{\Sigma}$

$$\widehat{\mathcal{U}}_1 = \text{argmax}_{\mathcal{W} \in \Theta} \langle \widehat{\Sigma}, \mathcal{W} \otimes \mathcal{W} \rangle.$$

And other PCs can also be equivalently defined as

$$\widehat{\mathcal{U}}_k = \text{argmax}_{\mathcal{W} \in \Theta : \mathcal{W} \perp_c \widehat{\mathcal{U}}_l, \forall l < k} \langle \widehat{\Sigma}, \mathcal{W} \otimes \mathcal{W} \rangle.$$

Note that $\widehat{\mathcal{U}}_k$ can also be identified with the best rank-one approximation to a deflated covariance operator:

$$\widehat{\mathcal{U}}_k = \text{argmax}_{\mathcal{W} \in \Theta} \langle \check{\Sigma}_k, \mathcal{W} \otimes \mathcal{W} \rangle,$$

where

$$\check{\Sigma}_k = \widehat{\Sigma} \times_1 \widehat{\mathcal{P}}_k^{(1)} \times_2 \cdots \times_p \widehat{\mathcal{P}}_k^{(p)} \times_{p+1} \widehat{\mathcal{P}}_k^{(1)} \times_{p+2} \cdots \times_{2p} \widehat{\mathcal{P}}_k^{(p)}$$

and $\widehat{\mathcal{P}}_k^{(1)}$ is the projection matrix of the linear subspace spanned by $\{\widehat{\mathbf{u}}_l^{(q)} : 1 \leq l < k\}$. Hereafter $\times_q$ represents the mode q product between a tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_k}$ and a matrix $A \in \mathbb{R}^{m \times d_q}$ so

that $\mathcal{T} \times_q A \in \mathbb{R}^{d_1 \times \ldots d_{q-1} \times m \times d_{q+1} \cdots \times d_k}$ with elements

$$[\mathcal{T} \times_q A]_{i_1 \ldots i_{q-1} j i_{q+1} \ldots i_k} = \sum_{i_q=1}^{d_q} \mathcal{T}_{i_1 \ldots i_q \ldots i_k} A_{j i_q}.$$

Computing the best rank-one approximation to a tensor is a classical problem in numerical linear algebra, and casting the sample multiway PCA as such allows us to take advantage of the many existing algorithms for doing so. In this work, we focus on the statistical properties of multiway PCA. Readers interested in further discussions about the computational aspect are referred to, e.g., T. Zhang and Golub 2001; Hackbusch 2012; Janzamin et al. 2019 and references therein.

Similar to the usual PCs, multiway PCs can be used to construct low-rank approximations of the original data. However, there are also fundamental, albeit sometimes subtle, differences between the two types of PCA. The usual sample PCs coincide with the leading singular vectors of the data matrix after appropriate centering and therefore can be computed via singular value decomposition (SVD). In contrast, multiway PCA is, while closely related to, not equivalent to the best low-rank approximations of the original data array in general. More specifically, consider stacking the observations into a higher-order tensor $\mathbf{X} \in \mathbb{R}^{n \times d_1 \times \cdots \times d_p}$ whose ith frontal slice is $\mathcal{X}_i$. In the case when $\mathcal{X}$ is a vector, i.e., $p = 1$, $\mathbf{X}$ is a matrix and the sample PC, $\widehat{\mathcal{U}}_k$ as defined above, is its kth right singular vector. It is therefore tempting to do the same and estimate $\mathcal{U}_k$s by seeking the best orthogonal low-rank approximation to $\mathbf{X}$ directly:

$$\min_{\substack{\mathscr{W}_1, \ldots, \mathscr{W}_r \in \Theta, \mathbf{a}_1, \ldots, \mathbf{a}_r \in \mathbb{R}^n \\ \mathscr{W}_l \perp_c \mathscr{W}_k, \mathbf{a}_l \perp \mathbf{a}_k, \forall l \neq k}} \|\mathbf{X} - (\mathbf{a}_1 \otimes \mathscr{W}_1 + \cdots + \mathbf{a}_r \otimes \mathscr{W}_r)\|_{\mathrm{F}} \tag{2.7}$$

See, e.g., Harshman and Lundy 1984. This problem, often known as the tensor SVD problem, has attracted a lot of attention in recent years. See, e.g., Richard and Montanari 2014; Hopkins, Shi, and Steurer 2015; Liu, M. Yuan, and Zhao 2017; A. Zhang and Xia 2018; Auddy and M. Yuan 2020. However, the sample multiway PCs are generally not the solution to (2.7). First of all, the difference between the best orthogonal rank-r and rank-$(r-1)$ approximations to $\mathbf{X}$ is

generally not a rank-one tensor and therefore cannot be associated with a multiway PC. See, e.g., Hackbusch 2012. To overcome this challenge, one may consider solving (2.7) in a greedy fashion, i.e, optimizing (2.7) over $\mathscr{W}_k$ and $\mathbf{a}_k$ only while fixing the other ones. In general, however, this still results in a different set of PCs because of the extra orthogonality constraint on $\mathbf{a}_k$s imposed by (2.7). As we shall see, this subtle distinction between multiway PCA and low-rank approximations to a data tensor not only means that a treatment different from that for the tensor SVD is needed for multiway PCA but also leads to different statistical behavior between the two.

Chapter 3: Rates of Convergence

A natural question one first asks is how well $\mathcal{U}_k$ and its components $\mathbf{u}_k^{(q)}$s can be estimated by their sample counterparts. We shall now turn our attention to this question and study the rate of convergence for the sample multiway PCs. On the one hand, we provide further justification for the superiority of multiway PCA to the usual PCA with stringing, in addition to enhanced interpretability. On the other hand, our investigation also leads to new insights into the operating characteristics of sample multiway PCA and its intriguing distinction from the usual PCA. To fix ideas, we shall consider the so-called spiked covariance model as a working model for our theoretical development.

Suppose that a random array $\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_p}$ follows a linear factor model:

$$\mathcal{X} = \sum_{k=1}^{r} \sigma_k \theta_k \mathcal{U}_k + \sigma_0 \mathcal{E}, \tag{3.1}$$

where $(\theta_1, \ldots, \theta_r)^\top \sim N(0, I_r)$ are the random factor loadings, $\mathcal{U}_k$s ($\in \Theta$) are unit length rank-one principal components such that $\mathcal{U}_k \perp_c \mathcal{U}_l$ for any $k \neq l$, and $\mathcal{E}$ is a noise tensor with independent $N(0, 1)$ entries. It is worth pointing out that our results and arguments can be extended beyond normality and applied to general subgaussian distributions. We opt for the normality assumption for ease of presentation. Without loss of generality, we shall also assume that eigenvalues of the signal are nontrivial and sorted in non-increasing order, i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Note that we do not require σ_ks to be distinct. It is not hard to see that the covariance operator of the aforementioned $\mathcal{X}$ is given by

$$\Sigma = \sum_{k=1}^{r} \sigma_k^2 \mathcal{U}_k \otimes \mathcal{U}_k + \sigma_0^2 \mathcal{I},$$

where $\mathcal{I}$ is the identity tensor, i.e., $\mathcal{I}_{j_1 \ldots j_p j'_1 \ldots j'_p} = 1$ if $j_q = j'_q$ for all $q = 1, \ldots, p$ and 0 otherwise. The spiked covariance model such as (3.1) is widely used as a working model to study PCA in the case of vector observations, i.e., $p = 1$. See, e.g., Johnstone 2001 and Paul 2007.

In this chapter, we shall establish the rates of convergence of the sample multiway PCs. To this end, denote by $\angle(\mathbf{w}_1, \mathbf{w}_2)$ the angle between two vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ taking value in $[0, \pi/2]$, and similarly for two arrays $\mathcal{W}_1$ and $\mathcal{W}_2$, $\angle(\mathcal{W}_1, \mathcal{W}_2)$ denotes the angle between their vectorizations $\mathrm{vec}(\mathcal{W}_1)$ and $\mathrm{vec}(\mathcal{W}_2)$.

It is instructive to begin with the classical setting where the dimensionality $d_1, \ldots, d_p$ as well as all other parameters, e.g. σ_0, σ_ks and r, are held fixed as the sample size n diverges. Our first result shows that the sample PC $\widehat{\mathcal{U}}_k$ and its components $\widehat{\mathbf{u}}_k^{(q)}$s are root-$n$ consistent in this regime.

Theorem 1. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent observations following the spiked covariance model* (3.1) *with $p > 1$ such that $\mathcal{U}_k = \mathbf{u}_k^{(1)} \otimes \cdots \otimes \mathbf{u}_k^{(p)}$ and $\sigma_k > 0$. Assume that all parameters are fixed as the sample size n increases. Let $\widehat{\mathcal{U}}_k = \widehat{\mathbf{u}}_k^{(1)} \otimes \cdots \otimes \widehat{\mathbf{u}}_k^{(p)}$ be the sample multiway PC as defined by* (2.6)*. Then there exists a permutation π over $[r] := \{1, \ldots, r\}$ such that*

$$\max_{1 \le q \le p} \sin \angle(\widehat{\mathbf{u}}_k^{(q)}, \mathbf{u}_{\pi(k)}^{(q)}) = O_p(n^{-1/2}), \tag{3.2}$$

for all $k \in [r]$, and hence

$$\sin \angle(\widehat{\mathcal{U}}_k, \mathcal{U}_{\pi(k)}) = O_p(n^{-1/2}), \qquad k = 1, \ldots, r$$

as $n \to \infty$.

The most notable difference between the above result and those for the usual PCA (e.g., Anderson 1984) is that fact that the root-n consistency of the sample multiway PCs does not require that the eigenvalues ($(\sigma_k^2 + \sigma_0^2)$s or equivalently σ_ks) of the covariance matrix be simple, i.e., $\sigma_k \neq \sigma_{k+1}$. Note that, without the multiway structural constraint, the usual PCs are only uniquely defined and hence can possibly be estimated if their corresponding eigenvalues are simple. As Theorem 1 in-

dicates, such a restriction is not necessary for multiway PCA. For multiway PCA, each sample PC is root-n consistent regardless of the other eigenvalues. It is also worth noting that, since we do not require the σ_ks to be distinct, there is no guarantee that $\widehat{\mathcal{U}}_k$ estimates $\mathcal{U}_k$. This is not a deficiency of multiway PCA, but rather a necessity due to the possible indeterminacy of the kth largest eigenvalue. In fact, if $\sigma_{k+1} < \sigma_k < \sigma_{k-1}$, then we can choose $\pi(k) = k$ in Theorem 1. In general, Theorem 1 shows that each of the sample PCs is necessarily a root-n consistent estimate of one of the multiway PCs.

To further understand the operating characteristics and merits of multiway PCA, we now consider the more general case and further highlight the role of dimensionality and signal-to-noise ratio. For brevity, in what follows, we shall assume that $\mathcal{X}$ is "nearly cubic" in that there exist constants $0 < c_1, c_2 < \infty$ such that $c_1 d \le d_1, \ldots, d_p \le c_2 d$ for some natural number d which may diverge with n. General cases can be treated similarly but incur considerably more cumbersome notation and tedious derivation.

Theorem 2. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent observations following the spiked covariance model* (3.1) *with $p > 1$ such that $\mathcal{U}_k = \mathbf{u}_k^{(1)} \otimes \cdots \otimes \mathbf{u}_k^{(p)}$. Let $\widehat{\mathcal{U}}_k = \widehat{\mathbf{u}}_k^{(1)} \otimes \cdots \otimes \widehat{\mathbf{u}}_k^{(p)}$ be the sample multiway PC as defined by* (2.6)*. Suppose that*

$$r \log r \le c_0 \min\{n, d\} \quad \text{and} \quad \left(\frac{\sigma_0}{\sigma_r} + \frac{\sigma_0^2}{\sigma_r^2} \right) \cdot \max\left\{ \sqrt{\frac{d}{n}}, \frac{d}{n} \right\} \le \frac{c_0}{\sqrt{r}}, \tag{3.3}$$

for a sufficiently small constant $c_0 > 0$. Then there exist a constant $C > 0$ and a permutation π over $[r]$ such that

$$\max_{1 \le q \le p} \sin \angle(\widehat{\mathbf{u}}_k^{(q)}, \mathbf{u}_{\pi(k)}^{(q)}) \le C \left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2} \right) \cdot \max\left\{ \sqrt{\frac{d}{n}}, \frac{d}{n} \right\}, \tag{3.4}$$

for all $k \in [r]$, and hence

$$\sin \angle(\widehat{\mathcal{U}}_k, \mathcal{U}_{\pi(k)}) \le C \left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2} \right) \cdot \max\left\{ \sqrt{\frac{d}{n}}, \frac{d}{n} \right\}, \qquad k = 1, \ldots, r,$$

with probability tending to one as n diverges.

Theorem 2 can be viewed as a generalization of Theorem 1. Its proof is rather involved and we shall brief discuss some of the challenges and the main ideas for resolving them. The proof proceeds by induction over k. Special attention is needed to deal with the case when an eigenvalue is not simple or the eigengap is small. This creates difficulty in identifying which multiway PC a sample multiway PC estimates, or equivalently the permutation π. To this end, we shall define

$$\pi(1) = \mathrm{argmax}_{1\le l\le r}\left\{\sigma_l^2\left|\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_1^{(q)}\rangle\right|\right\},$$

and for $k > 1$,

$$\pi(k) := \mathrm{argmax}_{l\notin\pi([k-1])}\left\{\sigma_l^2\left|\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right|\right\}.$$

To remove the influence of eigengaps altogether, we need to carefully quantify the impact of estimation error of $\widehat{\mathcal{U}}_1, \ldots, \widehat{\mathcal{U}}_{k-1}$ on the kth sample multiway PC. To this end, we shall derive bounds for both

$$\max_{1\le q\le p} \sin\angle(\mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}),$$

and

$$\max_{1\le q\le p}\max_{l\notin\pi([k])} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle,$$

and leverage the fact that the latter can be much smaller than the former.

When $d = O(n)$, the convergence rate given in Theorem 2 is

$$\sin\angle(\widehat{\mathcal{U}}_k, \mathcal{U}_{\pi(k)}) \le C\left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2}\right)\cdot\sqrt{\frac{d}{n}};$$

and when $d \gg n$, we have

$$\sin\angle(\widehat{\mathcal{U}}_k, \mathcal{U}_{\pi(k)}) \le C\left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2}\right)\cdot\frac{d}{n}.$$

In particular, $\widehat{\mathcal{U}}_k$ is consistent, e.g.,

$$\sin \angle(\widehat{\mathcal{U}}_k, \mathcal{U}_{\pi(k)}) = o_p(1),$$

whenever $\sigma_{\pi(k)}/\sigma_0 \gg \max\{d/n, (d/n)^{1/4}\}$.

Of particular interest here is the role of dimensionality. The rates of convergence given by Theorem 2 depend on the dimensionality through d rather than the ambient dimension $D := d_1 d_2 \cdots d_p$. This is because multiway PCA restricts PCs to be rank-one tensors and therefore has fewer parameters. Such dimensionality reduction is especially important for multiway data. Consider, for example, the case when σ_0 and σ_ks are fixed, then by virtue of the results from Johnstone and A. Y. Lu 2009, direct application of the usual PCA after stringing necessarily leads to an inconsistent estimate of $\mathcal{U}_k$ whenever $D \gg n$. Yet, our result indicates that multiway PCA is consistent whenever $d \ll n$.

To draw further comparisons with the usual PCA, we now focus on the case when $d \ll n$ and $r, \sigma_0, \sigma_1, \ldots, \sigma_r$ are fixed. As shown by Birnbaum et al. 2013, in this regime, the usual PCA (i.e., $p = 1$) satisfies

$$\sin \angle(\widehat{\mathcal{U}}_k, \mathcal{U}_{\pi(k)}) \asymp \left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2} \right) \cdot \sqrt{\frac{d}{n}} + \frac{1}{\sqrt{n}} \left(\sum_{k' \neq k} \frac{(\sigma_0 + \sigma_{\pi(k)})(\sigma_0 + \sigma_{\pi(k')})}{\sigma_{\pi(k)}^2 - \sigma_{\pi(k')}^2} \right)$$

with probability tending to one. Comparing the above rate with that from Theorem 2, it is clear that the difference between the two lie at the second term on the right hand side. Its presence for the usual PCA dictates that there should be no ties among σ_ks. Even if the σ_ks are all distinct, how well we can estimate a PC crucially depends on the gap between its corresponding eigenvalue and the other eigenvalues when $p = 1$. In contrast, the bounds given by Theorem 2 are determined by $\sigma_{\pi(k)}$ alone and not the eigengap $\min\{\sigma_{\pi(k)-1}^2 - \sigma_{\pi(k)}^2, \sigma_{\pi(k)}^2 - \sigma_{\pi(k)+1}^2\}$ as in the usual PCA case.

It is also instructive to compare the convergence rate for the multiway PCA from Theorem 2

with those for tensor SVD. Recall that

$$\mathbf{X} = \sum_{k=1}^{r} \sigma_k \Theta_k \otimes \mathscr{U}_k + \sigma_0 \mathbf{E},$$

where $\Theta_k = (\theta_{1k}, \ldots, \theta_{nk})^\top$ is a vector containing the n realizations of θ_k and $\mathbf{E}$ is a $n \times d_1 \times \cdots \times d_p$ tensor whose ith frontal slice is $\mathscr{E}_i$. In contrast, Θ_ks are deterministic in a tensor SVD model. If Θ_ks are orthogonal to each other, then $\mathscr{U}_k$ can be estimated at the rate of $\sigma_0 \|\mathbf{E}\| / (\sigma_k \|\Theta_k\|)$ which is of the order $(\sigma_0/\sigma_k) \max\{\sqrt{d/n}, 1\}$. This is a direct consequence of the perturbations bounds from Auddy and M. Yuan 2020 and a similar bound was also derived by Richard and Montanari 2014 in the rank-one case, i.e., $r = 1$. In our case, however, Θ_k and Θ_l are random and in general not orthogonal to each other. As a result, the rates we obtained are different in their dependence on the signal-to-noise ratio σ_k/σ_0. Similar phenomenon has also been observed for the usual PCA (see, e.g., Birnbaum et al. 2013).

Chapter 4: Asymptotic Normality and Bias Correction

We now turn to the distributional properties of multiway PCA. This requires us to further delineate the role of bias in the sample PCs. It is known that the usual PCA is biased when the dimension (D) is large when compared with the sample size. See, e.g., Koltchinskii and Lounici 2014; Koltchinskii, Löffler, and Nickl 2020 and the references therein. The same phenomenon is observed for the sample multiway PCs and a non-negligible bias arise when the dimension of each mode (d) is large when compared with the sample size. In addition, there is a more subtle source of bias for the sample multiway PCs due to the ambiguity in ordering the multiway PCs in the absence of eigengaps. As noted before, the lack of an eigengap means that the kth PC may not necessarily be estimated by the kth sample multiway PC. As a more concrete example, consider the case when $r = 2$ and $\lambda_1 = \lambda_2$. Then $\mathcal{U}_1$ can be estimated by either $\widehat{\mathcal{U}}_1$ or $\widehat{\mathcal{U}}_2$, and as Theorem 2 shows, the rate of convergence remains the same in both cases. But the asymptotic distribution may differ between the two scenarios: $\widehat{\mathcal{U}}_2$ is required to be orthogonal to $\widehat{\mathcal{U}}_1$ and estimating $\mathcal{U}_1$ by $\widehat{\mathcal{U}}_2$ may incur extra bias.

In this section, we shall introduce ways to correct for both types of bias and establish the asymptotic normality of the bias-corrected sample PCs. As is customary in the literature, we shall assume that r and $\sigma_1, \ldots, \sigma_r$ are fixed for brevity. In light of the results from the previous section, the sample PCs are consistent if $d \ll n$ in this setting. We shall therefore focus on this regime in the current section.

4.1 When $d = o(\sqrt{n})$

When d is not too large, the bias is solely due to the possibility of repeated eigenvalues and thus ambiguity of the ordering of PCs. Indeed if σ_ks are distinct, then there is no need for bias correction

when $d = o(\sqrt{n})$ and all of our results in this section will hold for the sample multiway PCs. But in practice, we may not know or want to assume that the eigenvalues are simple. Fortunately, we can remove any possible bias fairly easily by a simple one-step update of the sample PCs. More specifically, we shall consider estimating $\mathbf{u}_{\pi(k)}^{(q)}$ by $\tilde{\mathbf{u}}_k^{(q)}$, the leading eigenvector of

$$\widehat{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(q-1)}, \cdot, \widehat{\mathbf{u}}_k^{(q+1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(q-1)}, \cdot, \widehat{\mathbf{u}}_k^{(q+1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}).$$

The additional step frees up the orthogonality constraints imposed on the kth sample multiway PC and therefore allows us to suppress any adverse influence of $\widehat{\mathcal{U}}_1, \ldots, \widehat{\mathcal{U}}_{k-1}$.

We now consider the asymptotic distribution of the bias-corrected sample PCs. We again start with the classical regime when all parameters are fixed as n increases.

Theorem 3. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent observations following the spiked covariance model* (3.1) *with $p > 1$ such that $\mathcal{U}_k = \mathbf{u}_k^{(1)} \otimes \cdots \otimes \mathbf{u}_k^{(p)}$ and $\sigma_k > 0$. Assume that all parameters are fixed as the sample size n increases. Let $\tilde{\mathbf{u}}_1^{(q)}, \ldots, \tilde{\mathbf{u}}_r^{(q)}$ be defined as above. Then there exists a permutation $\pi : [r] \to [r]$ such that*

$$\begin{aligned} &\sqrt{n}\left[\operatorname{vec}(\tilde{\mathbf{U}}^{(q)}) - \operatorname{vec}(\mathbf{U}_\pi^{(q)})\right] \\ &\xrightarrow{d} N\left(0, \operatorname{diag}\left(\left(\frac{\sigma_0^2}{\sigma_{\pi(1)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(1)}^4}\right)\mathcal{P}_{\mathbf{u}_{\pi(1)}^{(q)}}^{\perp}, \ldots, \left(\frac{\sigma_0^2}{\sigma_{\pi(r)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(r)}^4}\right)\mathcal{P}_{\mathbf{u}_{\pi(r)}^{(q)}}^{\perp}\right)\right), \end{aligned}$$

as $n \to \infty$, where $\tilde{\mathbf{U}}^{(q)} = [\tilde{\mathbf{u}}_1^{(q)}, \ldots, \tilde{\mathbf{u}}_r^{(q)}]$, $\mathbf{U}_\pi^{(q)} = [\mathbf{u}_{\pi(1)}^{(q)}, \ldots, \mathbf{u}_{\pi(r)}^{(q)}]$ and $\mathcal{P}_{\mathbf{u}_k^{(q)}}^{\perp} = I_{d_q} - \mathbf{u}_k^{(q)} \otimes \mathbf{u}_k^{(q)}$.

Theorem 3 indicates that

$$n \cdot \operatorname{var}\left(\tilde{\mathbf{u}}_k^{(q)}\right) \to \frac{\sigma_0^2(\sigma_{\pi(k)}^2 + \sigma_0^2)}{\sigma_{\pi(k)}^4}\left(I - \mathbf{u}_{\pi(k)}^{(q)} \otimes \mathbf{u}_{\pi(k)}^{(q)}\right),$$

and

$$n \cdot \operatorname{cov}\left(\tilde{\mathbf{u}}_k^{(q)}, \tilde{\mathbf{u}}_l^{(q)}\right) \to 0.$$

Namely, all estimates of the multiway PCs are asymptotically normal and independent of each

other. Note also that the asymptotic distribution of $\tilde{\mathbf{u}}_k^{(q)}$ does not depend on other eigenvalues or PCs. In other words, it can be estimated to the same precision as if all other components $\mathscr{U}_l$, $l \neq k$ are known! This is to be contrasted with the usual PCA where the asymptotic distribution of $\mathbf{u}_k^{(q)}$ depends on all other eigenvectors and eigenvalues.

More specifically, it is well known that in vector case, i.e., when $p = 1$, under the additional assumption that $\sigma_1^2, \ldots, \sigma_r^2$ are distinct, the sample PCs satisfy

$$\begin{gathered} n \cdot \operatorname{var}\left(\widehat{\mathbf{u}}_k^{(1)}\right) \\ \to \sum_{1 \le l \le r, l \neq k} \frac{(\sigma_k^2+\sigma_0^2)(\sigma_l^2+\sigma_0^2)}{(\sigma_k^2-\sigma_l^2)^2} \mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)} + \frac{(\sigma_k^2+\sigma_0^2)\sigma_0^2}{\sigma_k^4} \left(I - \sum_{1 \le l \le r} \mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)}\right) \end{gathered}$$

and for any $1 \le l \le r$ and $l \neq k$,

$$n \cdot \operatorname{cov}\left(\widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_l^{(1)}\right) \to -\frac{(\sigma_k^2 + \sigma_0^2)(\sigma_l^2 + \sigma_0^2)}{(\sigma_k^2 - \sigma_l^2)^2} \cdot \mathbf{u}_k^{(1)} \otimes \mathbf{u}_l^{(1)}.$$

See, e.g., Anderson 1984. It is clear that the sample PCs are always correlated with each other. Moreover, note that

$$\frac{\sigma_0^2(\sigma_k^2 + \sigma_0^2)}{\sigma_k^4} \le \frac{(\sigma_k^2 + \sigma_0^2)(\sigma_l^2 + \sigma_0^2)}{(\sigma_k^2 - \sigma_l^2)^2}.$$

and the strict inequality holds for any $k \neq l \le r$. This suggests that the estimated multiway PCs have smaller variations than the usual PCs with the same set of eigenvalues.

We now turn our attention to the more general case when the dimensionality and other parameters are allowed to diverge with n. Because the PCs now may have different dimensions for different sample sizes, it is more natural to consider their linear forms, e.g. $\langle \mathbf{u}_k^{(q)}, \mathbf{v} \rangle$, for some fixed vector $\mathbf{v} \in \mathbb{R}^{d_q}$. If the dimensions are fixed, Theorem 3 immediately suggests that $\langle \tilde{\mathbf{u}}_k^{(q)}, \mathbf{v} \rangle$ estimates $\langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v} \rangle$, and

$$\sqrt{n}\left(\langle \tilde{\mathbf{u}}_k^{(q)}, \mathbf{v} \rangle - \langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v} \rangle\right) \to_d N\left(0, \left(\frac{\sigma_0^2}{\sigma_{\pi(k)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(k)}^4}\right) \|\mathcal{P}^{\perp}_{\mathbf{u}_{\pi(k)}^{(q)}} \mathbf{v}\|^2\right)$$

The following result shows that this continues to hold as long as $d = o(\sqrt{n})$.

Theorem 4. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent observations following the spiked covariance model* (3.1) *with $p > 1$ such that $\mathcal{U}_k = \mathbf{u}_k^{(1)} \otimes \cdots \otimes \mathbf{u}_k^{(p)}$ and $\sigma_k > 0$. Assume that* (3.3) *holds, $d = o(\sqrt{n})$. Then there exists a permutation $\pi : [r] \to [r]$ such that*

$$\sqrt{n}\left(\langle \tilde{\mathbf{u}}_k^{(q)}, \mathbf{v}\rangle - \langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v}\rangle\right) \to_d N\left(0, \left(\frac{\sigma_0^2}{\sigma_{\pi(k)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(k)}^4}\right) \|\mathcal{P}^{\perp}_{\mathbf{u}_{\pi(k)}^{(q)}} \mathbf{v}\|^2\right)$$

as $n \to \infty$.

Theorem 4 shows that the same asymptotic behavior of $\tilde{\mathbf{u}}_j^{(q)}$ as in the fixed dimension case can be expected whenever $d = o(\sqrt{n})$.

4.2 When $d = o(n)$

For higher dimension, the simple bias-correction described above is no longer sufficient and a close inspection reveals that $\tilde{\mathbf{u}}_j^{(q)}$ still incurs a non-negligible bias when $d \gg \sqrt{n}$. Thankfully, both types of bias can be corrected with a sample-splitting approach similar in spirit to the scheme developed by Koltchinskii and Lounici 2014 for the usual PCA.

Without loss of generality, assume that n is an even number and we randomly split the n observations into two halves: $(\mathcal{X}_1, \ldots, \mathcal{X}_{n/2})$ and $(\mathcal{X}_{n/2+1}, \ldots, \mathcal{X}_n)$. Denote by $\widehat{\Sigma}^{[1]}$ and $\widehat{\Sigma}^{[2]}$ the sample covariance operator based on the two halves of data respectively. Similarly, we shall write $\widehat{\mathcal{U}}_k^{[h]} = \widehat{\mathbf{u}}_k^{(1),[h]} \otimes \cdots \otimes \widehat{\mathbf{u}}_k^{(p),[h]}$ the kth sample PC based on the h (= 1 or 2) halves of the data. However, as noted before, $\widehat{\mathcal{U}}_k^{[1]}$ and $\widehat{\mathcal{U}}_k^{[2]}$ may not estimate the same PC. To this end, we shall reorder $\widehat{\mathcal{U}}_k^{[1]}$s and $\widehat{\mathcal{U}}_k^{[2]}$s with $\widehat{\mathcal{U}}_k$s (i.e., the estimators derived from the entire dataset) as reference points. Specifically, without loss of generality, we assume that

$$\widehat{\mathcal{U}}_k^{[1]} = \text{argmin}_{k \le l \le r} \sin \angle \left(\widehat{\mathcal{U}}_l^{[1]}, \widehat{\mathcal{U}}_k\right).$$

The same procedure is applied to relabel $\widehat{\mathcal{U}}_k^{[2]}$s. Note also that the sign of a PC is irrelevant in that

$\mathcal{U}_k$ and $-\mathcal{U}_k$ represent the same transformation. We shall therefore also assume hereafter, without loss of generality, that $\langle \widehat{\mathbf{u}}_k^{(q),[1]}, \widehat{\mathbf{u}}_k^{(q),[2]} \rangle \geq 0$.

Recall that

$$\sigma_k^2 \mathbf{u}_k^{(q)} \mathbf{u}_k^{(q)^\top} + \sigma_0^2 I_{d_q} = \Sigma(\mathbf{u}_k^{(1)}, \ldots, \mathbf{u}_k^{(q-1)}, \cdot, \mathbf{u}_k^{(q+1)}, \ldots, \mathbf{u}_k^{(p)}, \mathbf{u}_k^{(1)}, \ldots, \mathbf{u}_k^{(q-1)}, \cdot, \mathbf{u}_k^{(q+1)}, \ldots, \mathbf{u}_k^{(p)}).$$

We shall then update the sample PC using the above identity with Σ and $\mathbf{u}_k^{(q)}$s estimated from separate halves. Denote by $\check{\mathbf{u}}_k^{(q),[1]}$ the leading eigenvector of

$$\widehat{\Sigma}^{[1]}(\widehat{\mathbf{u}}_k^{(1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),[2]}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(p),[2]}, \widehat{\mathbf{u}}_k^{(1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),[2]}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(p),[2]}),$$

and similarly $\check{\mathbf{u}}_k^{(q),[2]}$ the leading eigenvector of

$$\widehat{\Sigma}^{[2]}(\widehat{\mathbf{u}}_k^{(1),[1]}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),[1]}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),[1]}, \ldots, \widehat{\mathbf{u}}_k^{(p),[1]}, \widehat{\mathbf{u}}_k^{(1),[1]}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),[1]}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),[1]}, \ldots, \widehat{\mathbf{u}}_k^{(p),[1]}).$$

To avoid losing efficiency due to sample splitting, we consider a new estimate $\check{\mathcal{U}}_k = \check{\mathbf{u}}_k^{(1)} \otimes \cdots \otimes \check{\mathbf{u}}_k^{(p)}$ where

$$\check{\mathbf{u}}_k^{(q)} = \frac{\check{\mathbf{u}}_k^{(q),[1]} + \check{\mathbf{u}}_k^{(q),[2]}}{\left\| \check{\mathbf{u}}_k^{(q),[1]} + \check{\mathbf{u}}_k^{(q),[2]} \right\|}.$$

The following theorem shows that we can construct an unbiased estimate of $\langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v} \rangle$ by appropriately rescaling $\langle \check{\mathbf{u}}_k^{(q)}, \mathbf{v} \rangle$, as long as $d = o(n^{2/3})$. In order to extend the asymptotic independence

result to the current setting of linear forms, we consider the asymptotic distribution of

$$\Bigg(\big(1 + b_{k_1}^{(q)}\big) \big\langle \check{\mathbf{u}}_{k_1}^{(q)}, \mathbf{v}_1 \big\rangle - \big\langle \mathbf{u}_{\pi(k_1)}^{(q)}, \mathbf{v}_1 \big\rangle, \big(1 + b_{k_2}^{(q)}\big) \big\langle \check{\mathbf{u}}_{k_2}^{(q)}, \mathbf{v}_2 \big\rangle - \big\langle \mathbf{u}_{\pi(k_2)}^{(q)}, \mathbf{v}_2 \big\rangle, \ldots,$$
$$\big(1 + b_{k_m}^{(q)}\big) \big\langle \check{\mathbf{u}}_{k_m}^{(q)}, \mathbf{v}_m \big\rangle - \big\langle \mathbf{u}_{\pi(k_m)}^{(q)}, \mathbf{v}_m \big\rangle \Bigg),$$

where $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \in \mathbb{R}^{d_q}$.

Theorem 5. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent observations following the spiked covariance model (3.1) with $p > 1$ such that $\mathcal{U}_k = \mathbf{u}_k^{(1)} \otimes \cdots \otimes \mathbf{u}_k^{(p)}$ and $\sigma_k > 0$. Let $\check{\mathcal{U}}_k = \check{\mathbf{u}}_k^{(1)} \otimes \cdots \otimes \check{\mathbf{u}}_k^{(p)}$ be the estimated PC as defined above. Assume r and $\sigma_1, \ldots, \sigma_r$ are fixed, and $d = o(n^{2/3})$. Then there exists a permutation $\pi : [r] \to [r]$ such that*

$$\sqrt{n} \Bigg(\big(1 + b_{k_1}^{(q)}\big) \big\langle \check{\mathbf{u}}_{k_1}^{(q)}, \mathbf{v}_1 \big\rangle - \big\langle \mathbf{u}_{\pi(k_1)}^{(q)}, \mathbf{v}_1 \big\rangle, \big(1 + b_{k_2}^{(q)}\big) \big\langle \check{\mathbf{u}}_{k_2}^{(q)}, \mathbf{v}_2 \big\rangle - \big\langle \mathbf{u}_{\pi(k_2)}^{(q)}, \mathbf{v}_2 \big\rangle, \ldots,$$
$$\big(1 + b_{k_m}^{(q)}\big) \big\langle \check{\mathbf{u}}_{k_m}^{(q)}, \mathbf{v}_m \big\rangle - \big\langle \mathbf{u}_{\pi(k_m)}^{(q)}, \mathbf{v}_m \big\rangle \Bigg) \to_d N(\mathbf{0}, \mathbf{\Gamma}),$$

where

$$\mathbf{\Gamma}_{ij} = \begin{cases} \left(\frac{\sigma_0^2}{\sigma_{\pi(k_i)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(k_i)}^4} \right) \left\langle \mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{\pi(k_i)}} \mathbf{v}_i, \mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{\pi(k_i)}} \mathbf{v}_j \right\rangle, & \text{if } k_i = k_j, \\ 0, & \text{if } k_i \neq k_j, \end{cases}$$

and specifically,

$$\sqrt{n} \left((1 + b_k^{(q)}) \langle \check{\mathbf{u}}_k^{(q)}, \mathbf{v} \rangle - \langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v} \rangle \right) \to_d N \left(0, \left(\frac{\sigma_0^2}{\sigma_{\pi(k)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(k)}^4} \right) \| \mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{\pi(k)}} \mathbf{v} \|^2 \right)$$

as $n \to \infty$ where

$$b_k^{(q)} = \sqrt{1 + \frac{d_q}{n} \left(\frac{\sigma_0^2}{\sigma_{\pi(k)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(k)}^4} \right)} - 1. \tag{4.1}$$

It is worth pointing out that when $d = o(n^{1/2})$, the bias correction factor described by (4.1)

obeys $b_k^{(q)} = o(n^{-1/2})$ and therefore can be neglected. This agrees with our earlier observation and of course also suggests that sample-splitting is unnecessary if $d \ll n^{1/2}$. When $d \gg n^{1/2}$, bias correction becomes essential. In particular, Theorem 5 suggests that, as long as $d = o(n^{2/3})$, an explicit bias correction factor can be applied. For higher dimensions, it is unclear if a similar explicit expression exists for the debiasing factor. Nonetheless, we can derive a suitable bias correction factor for all $d \ll n$ via additional sample splitting.

More specifically, we first randomly split the observations into two halves. The first half of the data is then further split into two equal-sized groups to compute the sample covariance operators $\widehat{\Sigma}^{[1][1]}$ and $\widehat{\Sigma}^{[1][2]}$, then we compute $\widehat{\mathbf{u}}_k^{(q),[1][1]}$ and $\widehat{\mathbf{u}}_k^{(q),[1][2]}$ as the leading eigenvectors of

$$\begin{aligned}
&\widehat{\Sigma}^{[1][1]}(\widehat{\mathbf{u}}_k^{(1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),[2]}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(p),[2]}, \\
&\qquad \widehat{\mathbf{u}}_k^{(1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),[2]}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(p),[2]}), \\
&\widehat{\Sigma}^{[1][2]}(\widehat{\mathbf{u}}_k^{(1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),[2]}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(p),[2]}, \\
&\qquad \widehat{\mathbf{u}}_k^{(1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),[2]}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),[2]}, \ldots, \widehat{\mathbf{u}}_k^{(p),[2]}),
\end{aligned}$$

Similarly, we used the second half of the data to compute $\widehat{\mathbf{u}}_k^{(q),[2][1]}$s, and $\widehat{\mathbf{u}}_k^{(q),[2][2]}$s. As before, we shall sort these estimates in compatible order and sign. Let

$$\widehat{b}_k^{(q)} = \frac{\left\| \check{\mathbf{u}}_k^{(q),[1]} + \check{\mathbf{u}}_k^{(q),[2]} \right\|}{\sqrt{\left\langle \widehat{\mathbf{u}}_k^{(q),[1][1]}, \widehat{\mathbf{u}}_k^{(q),[1][2]} \right\rangle} + \sqrt{\left\langle \widehat{\mathbf{u}}_k^{(q),[2][1]}, \widehat{\mathbf{u}}_k^{(q),[2][2]} \right\rangle}} - 1. \tag{4.2}$$

Theorem 6. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent observations following the spiked covariance model* (3.1) *with $p > 1$ such that $\mathcal{U}_k = \mathbf{u}_k^{(1)} \otimes \cdots \otimes \mathbf{u}_k^{(p)}$ and $\sigma_k > 0$. Let $\check{\mathcal{U}}_k = \check{\mathbf{u}}_k^{(1)} \otimes \cdots \otimes \check{\mathbf{u}}_k^{(p)}$ be the estimated PC as defined above. Assume r and $\sigma_1, \ldots, \sigma_r$ are fixed, and $d = o(n)$. Then there*

exists a permutation $\pi : [r] \to [r]$ *such that*

$$\sqrt{n}\Bigg(\big(1 + \widehat{b}^{(q)}_{k_1}\big)\big\langle \check{\mathbf{u}}^{(q)}_{k_1}, \mathbf{v}_1 \big\rangle - \big\langle \mathbf{u}^{(q)}_{\pi(k_1)}, \mathbf{v}_1 \big\rangle, \big(1 + \widehat{b}^{(q)}_{k_2}\big)\big\langle \check{\mathbf{u}}^{(q)}_{k_2}, \mathbf{v}_2 \big\rangle - \big\langle \mathbf{u}^{(q)}_{\pi(k_2)}, \mathbf{v}_2 \big\rangle, \ldots,$$
$$\big(1 + \widehat{b}^{(q)}_{k_m}\big)\big\langle \check{\mathbf{u}}^{(q)}_{k_m}, \mathbf{v}_m \big\rangle - \big\langle \mathbf{u}^{(q)}_{\pi(k_m)}, \mathbf{v}_m \big\rangle \Bigg) \to_d N\left(\mathbf{0}, \mathbf{\Gamma}\right),$$

where

$$\mathbf{\Gamma}_{ij} = \begin{cases} \left(\frac{\sigma_0^2}{\sigma^2_{\pi(k_i)}} + \frac{\sigma_0^4}{\sigma^4_{\pi(k_i)}} \right) \left\langle \mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{\pi(k_i)}} \mathbf{v}_i, \mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{\pi(k_i)}} \mathbf{v}_j \right\rangle, & \text{if } k_i = k_j, \\ 0, & \text{if } k_i \neq k_j, \end{cases}$$

and specifically,

$$\sqrt{n}\left((1 + \widehat{b}^{(q)}_k)\langle \check{\mathbf{u}}^{(q)}_k, \mathbf{v} \rangle - \langle \mathbf{u}^{(q)}_{\pi(k)}, \mathbf{v} \rangle \right) \to_d N\left(0, \left(\frac{\sigma_0^2}{\sigma^2_{\pi(k)}} + \frac{\sigma_0^4}{\sigma^4_{\pi(k)}} \right) \| \mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{\pi(k)}} \mathbf{v} \|^2 \right),$$

as $n \to \infty$ *where* $\widehat{b}^{(q)}_k$ *is given by* (4.2)*. Moreover,*

$$\widehat{b}^{(q)}_k = \sqrt{1 + \frac{d_q}{n}\left(\frac{\sigma_0^2}{\sigma^2_{\pi(k)}} + \frac{\sigma_0^4}{\sigma^4_{\pi(k)}} \right)} - 1 + O_p\left(\frac{d^{3/2}}{n^{3/2}} \right) + o_p\left(\frac{1}{\sqrt{n}} \right).$$

In light of Theorem 6, the double sample splitting approach can be employed to derive confidence intervals for linear forms of the multiway PCs as long as $d = o(n)$. This robustness, however, comes at the expense of increased computational cost and could incur a loss of efficiency in finite samples. In practice, one may still prefer the explicit bias correction as described by Theorem 5 if d is not very large, or the one-step update if d is small.

4.3 Inference about multiway PCs

The asymptotic normality we showed earlier in the section forms the basis for making inferences about linear forms $\langle \mathbf{u}^{(q)}_{\pi(k)}, \mathbf{v} \rangle$. In particular, one of the most interesting and also simplest

examples of linear forms of PCs is their coordinates, i.e., $\mathbf{v}$ is a column vector of the identity matrix. To derive confidence intervals of or testing hypotheses about $\langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v} \rangle$, however, we need to also estimate its variance. Specifically, its asymptotic distribution depends only on σ_0, $\sigma_{\pi(k)}$, and $\mathbf{u}_k^{(q)}$, all of which can be consistently estimated by their sample counterpart. Let

$$\widehat{\sigma}_0^2 = \frac{1}{\prod_{q=1}^{p}(d_q - r)} \sum_{1 \le i_q \le d_q, 1 \le q \le p} [\check{\Sigma}_{r+1}]_{i_1 \cdots i_p i_1 \cdots i_p}$$

and

$$\widehat{\sigma}_{\pi(k)}^2 = \langle \widehat{\Sigma}, \widehat{\mathcal{U}}_k \otimes \widehat{\mathcal{U}}_k \rangle - \widehat{\sigma}_0^2.$$

The following theorem suggest that the asymptotic normality remains valid if we replace the variance of linear forms $\langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v} \rangle$ with these estimates:

Theorem 7. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent observations following the spiked covariance model* (3.1) *with $p > 1$ such that $\mathcal{U}_k = \mathbf{u}_k^{(1)} \otimes \cdots \otimes \mathbf{u}_k^{(p)}$ and $\sigma_k > 0$. Assume r and $\sigma_1, \ldots, \sigma_r$ are fixed. There exists a permutation $\pi : [r] \to [r]$ such that*

(a) If $d = o(\sqrt{n})$, then

$$\frac{\sqrt{n}\left(\langle \tilde{\mathbf{u}}_k^{(q)}, \mathbf{v} \rangle - \langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v} \rangle\right)}{\sqrt{\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_{\pi(k)}^2} + \frac{\widehat{\sigma}_0^4}{\widehat{\sigma}_{\pi(k)}^4}} \left\| \mathcal{P}^{\perp}_{\tilde{\mathbf{u}}_{\pi(k)}^{(q)}} \mathbf{v} \right\|} \to_d N(0,1),$$

(b) If $d = o(n^{2/3})$, then

$$\frac{\sqrt{n}\left((1 + b_k^{(q)})\langle \check{\mathbf{u}}_k^{(q)}, \mathbf{v} \rangle - \langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v} \rangle\right)}{\sqrt{\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_{\pi(k)}^2} + \frac{\widehat{\sigma}_0^4}{\widehat{\sigma}_{\pi(k)}^4}} \left\| \mathcal{P}^{\perp}_{\check{\mathbf{u}}_{\pi(k)}^{(q)}} \mathbf{v} \right\|} \to_d N(0,1),$$

where $b_k^{(q)}$ is given by (4.1).

(c) If $d = o(n)$, then

$$\frac{\sqrt{n}\left((1+\widehat{b}_k^{(q)})\langle \check{\mathbf{u}}_k^{(q)}, \mathbf{v}\rangle - \langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v}\rangle\right)}{\sqrt{\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_{\pi(k)}^2} + \frac{\widehat{\sigma}_0^4}{\widehat{\sigma}_{\pi(k)}^4}}\left\|\mathcal{P}^{\perp}_{\check{\mathbf{u}}_{\pi(k)}^{(q)}}\mathbf{v}\right\|} \to_d N(0,1),$$

where $\widehat{b}_k^{(q)}$ is given by (4.2).

Theorem 7 is an immediate consequence of Slutsky's Theorem and Theorems 4-6. It allows us to make inference or construct confidence intervals for $\langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v}\rangle$. Consider, for example, testing hypothesis that

$$H_0 : \langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v}\rangle = 0 \qquad vs \qquad H_a : \langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v}\rangle \neq 0,$$

when $d = o(\sqrt{n})$. We can proceed to reject H_0 if and only if

$$\left|\sqrt{n}\left(\langle \tilde{\mathbf{u}}_k^{(q)}, \mathbf{v}\rangle - \langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v}\rangle\right)\right| \geq z_{\alpha/2}\sqrt{\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_{\pi(k)}^2} + \frac{\widehat{\sigma}_0^4}{\widehat{\sigma}_{\pi(k)}^4}}\left\|\mathcal{P}^{\perp}_{\tilde{\mathbf{u}}_{\pi(k)}^{(q)}}\mathbf{v}\right\|,$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. Theorem 7 guarantees this is a level-α test asymptotically. Similarly, we can also construct $(1-\alpha)$ confidence interval for $\langle \mathbf{u}_{\pi(k)}^{(q)}, \mathbf{v}\rangle$:

$$\left(\langle \tilde{\mathbf{u}}_k^{(q)}, \mathbf{v}\rangle \pm \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_{\pi(k)}^2} + \frac{\widehat{\sigma}_0^4}{\widehat{\sigma}_{\pi(k)}^4}}\left\|\mathcal{P}^{\perp}_{\tilde{\mathbf{u}}_{\pi(k)}^{(q)}}\mathbf{v}\right\|\right).$$

In particular, by taking $\mathbf{v} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_{d_q}\}$, we can use the above formula to derive confidence intervals for the coordinates of $\mathbf{u}_{\pi(k)}^{(q)}$. Situations with larger d can also be treated accordingly.

Chapter 5: Approximation Algorithm and Numerical Experiments

To complement our theoretical analyses and further demonstrate the merits of multiway PCA, we first introduce a polynomial time algorithm to approximate the sample PCs, and then conduct several sets of numerical experiments.

5.1 Approximation Algorithm

Sample PCs (2.6) are defined as maximizers of a sequence of optimization problems that are highly non-convex. The definition of the first sample PC is closely related to finding the best rank-1 tensor approximation, which is an NP-hard problem (Hillar and Lim 2013). We propose the following computational feasible procedure called "Sequential Extraction with Multiple Initialization"(SEMI), which consists of polynomial numbers of initialization and two-time iterations, and prove that its output reaches the optimal statistical rate of estimation under further signal-to-noise ratio assumptions.

Given the non-convex nature of our problem, initialization is the key to any successful optimization algorithms. Our model is closely related to CP and orthogonal decomposition of tensors, and the initialization methods in related literature can be divided into two types: random initialization or initialization by matricization. The shortcoming of random initialization is that it does not utilize the low-rank structure of the problem, while any statistical guarantees of initialization by matricization relies on some kind of eigengap conditions (Chang et al. 2021, Y. Han, C.-H. Zhang, and R. Chen 2021). A desirable initialization method should work without any eigengap conditions, since that is one of the main merits of treating $\mathbf{X}$ as an array instead of a matrix. In the following, we propose an initialization method that combines the advantage of random initialization and matricization, which utilizes the low-rank structure while not depending on eigengap

conditions.

To motivate our algorithm, suppose

$$\sigma_1 = \sigma_2 = \cdots = \sigma_{r_0} > \sigma_{r_0+1}.$$

If we consider contracting out mode 2 with mode $p+2$, mode 3 with mode $p+3$, ..., mode p with mode $2p$ for the covariance operator Σ, i.e., denote $\mathrm{contr}_1(\Sigma) \in \mathbb{R}^{d_1 \times d_1}$ as

$$\big[\mathrm{contr}_1(\Sigma)\big]_{ij} := \sum_{i_2,\ldots,i_p} \Sigma_{i,i_2,\ldots,i_p,j,i_2,\ldots,i_p},$$

then we can easily observe that

$$\mathrm{contr}_1(\Sigma) = \sum_{k=1}^{r} \sigma_k^2 \mathbf{u}_k^{(1)} \otimes \mathbf{u}_k^{(1)} + \sigma_0^2 \mathbf{I},$$

so the leading r_0-dimensional eigenspace of $\mathrm{contr}(\Sigma)$ is

$$\mathrm{span}\{\mathbf{u}_1^{(1)}, \ldots, \mathbf{u}_{r_0}^{(1)}\}.$$

In reality, Σ is unknown, so we consider the leading r_0-dimensional eigenspace of $\mathrm{contr}_1(\widehat{\Sigma})$, and the intuition is that it should be close to $\mathrm{span}\{\mathbf{u}_1^{(1)}, \ldots, \mathbf{u}_{r_0}^{(1)}\}$, if $\sigma_{r_0} - \sigma_{r_0+1}$ is large enough.

Denote $\widehat{\mathbf{U}}$ to be the $d_1 \times r_0$ matrix containing the leading r_0 eigenvectors of $\mathrm{contr}_1(\widehat{\Sigma})$. We then use it as the starting point of random initialization. If we consider a random matrix $\mathbf{A} \in \mathbb{R}^{r_0 \times M}$ with independent columns uniformly sampled from $(r_0 - 1)$-sphere , then it is likely that at least one of the columns of

$$\widehat{\mathbf{U}}\mathbf{A}$$

is close enough to one of $\mathbf{u}_k^{(1)}$, $k \in [r_0]$. To fix ideas, let us assume that the initial vector $\widehat{\mathbf{u}}_1^{(1),ini}$ is close to $\mathbf{u}_2^{(1)}$ and satisfies

$$\langle \widehat{\mathbf{u}}_1^{(1),ini}, \mathbf{u}_2^{(1)} \rangle > \frac{4}{5},$$

then we have actually created an eigengap for the subsequent modes:

$$\begin{aligned}&\Sigma \times_1 \widehat{\mathbf{u}}_1^{(1),ini} \times_{p+1} \widehat{\mathbf{u}}^{(1),ini}\\ &= \sum_{k=1}^{r} \tilde{\sigma}_k^2 \mathbf{u}_k^{(2)} \otimes \cdots \otimes \mathbf{u}_k^{(p)} \otimes \mathbf{u}_k^{(2)} \otimes \cdots \otimes \mathbf{u}_k^{(p)} + \sigma_0^2 \mathcal{F},\end{aligned}$$

where $\tilde{\sigma}_k^2 = \sigma_k^2 \langle \widehat{\mathbf{u}}_1^{(1),ini}, \mathbf{u}_k^{(1)} \rangle^2$, because

$$\tilde{\sigma}_2^2 > \frac{16}{25}\sigma_2^2 = \frac{16}{25}\sigma_1^2,$$

and for all other $k \in [r], k \neq 2$, we have

$$\tilde{\sigma}_k^2 < \frac{9}{25}\sigma_1^2,$$

i.e., we have an eigengap of at least $\frac{7}{25}\sigma_1^2$.

In reality, we do not have information on the multiplicity of σ_k's. We choose r_0 to be large enough so that multiplicity of σ_k's is no larger than r_0. Then, we initialize $\widehat{\mathbf{u}}_1^{(1),ini}$ by the columns of

$$\widehat{\mathbf{U}}[\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{r_0}] \in \mathbb{R}^{d_1 \times [(r_0-1)*M+1]}, \tag{5.1}$$

where $\mathbf{A}_1 = [1, 0, \ldots, 0]^\top \in \mathbb{R}^{r_0 \times 1}$ and

$$\mathbf{A}_{r*} = \begin{bmatrix} \tilde{\mathbf{A}}_{r*} \\ \mathbf{0}_{(r_0-r*)\times M} \end{bmatrix} \in \mathbb{R}^{r_0 \times M}, r^* = 2, 3, \ldots, r_0, \tag{5.2}$$

in which $\tilde{\mathbf{A}}_{r*} \in \mathbb{R}^{r^* \times M}$ has independent columns uniformly sampled from the unit sphere in $\mathbb{R}^{r*}$. The reason is that if there is a large enough eigengap $\sigma_{r^*} - \sigma_{r^*+1}$ anywhere between 1 and r_0, then there will exist a column in $\mathbf{U}\mathbf{A}_{r^*}$ close to one of $\{\mathbf{u}_1^{(1)}, \ldots, \mathbf{u}_{r^*}^{(1)}\}$, and by constructing the initializations this way, serving as an initial vector $\widehat{\mathbf{u}}_1^{(1),ini}$.

Then, for each initialization $\widehat{\mathbf{u}}_1^{(1),ini}$, we initialize $\widehat{\mathbf{u}}_1^{(2),0}$ to be the leading eigenvector of

$$\operatorname{contr}_1\left[\widehat{\Sigma} \times_1 (\widehat{\mathbf{u}}_1^{(1),ini})^\top \times_{p+1} (\widehat{\mathbf{u}}_1^{(1),ini})^\top\right],$$

and then initialize $\widehat{\mathbf{u}}_1^{(3),0}$ to be the leading eigenvector of

$$\operatorname{contr}_1\left[\widehat{\Sigma} \times_1 (\widehat{\mathbf{u}}_1^{(1),ini})^\top \times_2 (\widehat{\mathbf{u}}_1^{(2),0})^\top \times_{p+1} (\widehat{\mathbf{u}}_1^{(1),ini})^\top \times_{p+2} (\widehat{\mathbf{u}}_1^{(2),0})^\top\right],$$

and so on. After that we let $\widehat{\mathbf{u}}_1^{(1),0}$ be the leading eigenvector of

$$\widehat{\Sigma}\big(\cdot, \widehat{\mathbf{u}}_1^{(2),0}, \ldots, \widehat{\mathbf{u}}_1^{(p),0}, \cdot, \widehat{\mathbf{u}}_1^{(2),0}, \ldots, \widehat{\mathbf{u}}_1^{(p),0}\big).$$

Among all $r_0 + M$ potential starting points, we choose the one that maximizes

$$\left\langle \widehat{\Sigma}, \widehat{\mathscr{U}}_{1,0} \otimes \widehat{\mathscr{U}}_{1,0} \right\rangle, \text{ where } \widehat{\mathscr{U}}_{1,0} = \widehat{\mathbf{u}}_1^{(1),0} \otimes \cdots \otimes \widehat{\mathbf{u}}_1^{(p),0}.$$

After initializing $\widehat{\mathbf{u}}_1^{(q),0}, q \in [p]$, we iterate through the p modes such that $\widehat{\mathbf{u}}_1^{(q)}$ is updated with the leading eigenvector of

$$\widehat{\Sigma}\left(\widehat{\mathbf{u}}_1^{(1)}, \ldots, \widehat{\mathbf{u}}_1^{(q-1)}, \cdot, \widehat{\mathbf{u}}_1^{(q+1)}, \ldots, \widehat{\mathbf{u}}_1^{(p)}, \widehat{\mathbf{u}}_1^{(1)}, \ldots, \widehat{\mathbf{u}}_1^{(q-1)}, \cdot, \widehat{\mathbf{u}}_1^{(q+1)}, \ldots, \widehat{\mathbf{u}}_1^{(p)}\right),$$

and the iteration is run twice to output $\widehat{\mathscr{U}}_1$.

After finding $\widehat{\mathscr{U}}_1$, we update $\widehat{\Sigma}$ with

$$\check{\Sigma} \leftarrow \widehat{\Sigma} \times_1 \widehat{Q}_1^{(1)} \cdots \times_p \widehat{Q}_1^{(p)} \times_{p+1} \widehat{Q}_1^{(1)} \cdots \times_{2p} \widehat{Q}_1^{(p)},$$

where $\widehat{Q}_1^{(q)} := \mathbf{I} - \widehat{\mathbf{u}}^{(q)}(\widehat{\mathbf{u}}^{(q)})^\top$. Repeat the whole process on $\check{\Sigma}$ to get $\widehat{\mathscr{U}}_2$. Update $\check{\Sigma}$ with

$$\check{\Sigma} \leftarrow \check{\Sigma} \times_1 \widehat{Q}_2^{(1)} \cdots \times_p \widehat{Q}_2^{(p)} \times_{p+1} \widehat{Q}_2^{(1)} \cdots \times_{2p} \widehat{Q}_2^{(p)},$$

and the procedure goes on until we reach the r-th sample PC. We summarize the computational method in Algorithm 1.

Algorithm 1: Sequential Extraction with Multiple Initialization (SEMI)

input : Sample covariance operator $\widehat{\Sigma}$, rank r, initialization parameters r_0 and M.

for $k = 1$ *to* r **do**

Initialization:

Obtain $\widehat{\mathbf{U}}$, the $d_1 \times r_0$ matrix containing the leading r_0 eigenvectors of $\mathrm{contr}_1(\check{\Sigma})$. Generate random matrix $[\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{r_0}] \in \mathbb{R}^{d_1 \times [(r_0-1)*M+1]}$ as defined in (5.2);

for $s = 1$ *to* $(r_0 - 1) * M + 1$ **do**

$$\widehat{\mathbf{u}}_k^{(1),0}(s) \leftarrow \left\{\widehat{\mathbf{U}}[\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{r_0}]\right\}(:, s),$$

for $q = 2$ *to* p **do**

$$\widehat{\mathbf{u}}_k^{(q),0}(s) \leftarrow \mathrm{contr}_1\left[\check{\Sigma} \times_1 [\widehat{\mathbf{u}}_k^{(1),0}(s)]^\top \cdots \times_{q-1} [\widehat{\mathbf{u}}_k^{(q-1),0}(s)]^\top \times_{p+1} [\widehat{\mathbf{u}}_k^{(1),0}(s)]^\top \cdots \times_{p+q-1} [\widehat{\mathbf{u}}_k^{(q-1),0}(s)]^\top\right].$$

end

Update $\widehat{\mathbf{u}}_k^{(1),0}(s)$ once more: Let $\widehat{\mathbf{u}}_k^{(1),0}(s)$ be the leading eigenvector of

$$\check{\Sigma}\big(\cdot, \widehat{\mathbf{u}}_k^{(2),0}(s), \ldots, \widehat{\mathbf{u}}_k^{(p),0}(s), \cdot, \widehat{\mathbf{u}}_k^{(2),0}(s), \ldots, \widehat{\mathbf{u}}_k^{(p),0}(s)\big).$$

Let

$$\widehat{\mathbf{u}}_k^{(1),0} \otimes \cdots \otimes \widehat{\mathbf{u}}_k^{(p),0} = \widehat{\mathscr{U}}_{k,0} \leftarrow \mathrm{argmax}_{s \in [r_0+M]} \left\langle \check{\Sigma}, \widehat{\mathscr{U}}_{k,0}(s) \otimes \widehat{\mathscr{U}}_{k,0}(s) \right\rangle.$$

where $\widehat{\mathscr{U}}_{k,0}(s) = \widehat{\mathbf{u}}_k^{(1),0}(s) \otimes \cdots \otimes \widehat{\mathbf{u}}_k^{(p),0}(s)$.

end

Iteration:

for $t = 0, 1$ **do**

for $q = 2$ *to* p **do**

Let $\widehat{\mathbf{u}}_k^{(q),t+1} \leftarrow$ the leading eigenvector of

$$\check{\Sigma}\big(\widehat{\mathbf{u}}_k^{(1),t}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),t}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),t}, \ldots, \widehat{\mathbf{u}}_k^{(p),t}, \widehat{\mathbf{u}}_k^{(1),t}, \ldots, \widehat{\mathbf{u}}_k^{(q-1),t}, \cdot, \widehat{\mathbf{u}}_k^{(q+1),t}, \ldots, \widehat{\mathbf{u}}_k^{(p),t}\big).$$

end

end

Let

$$\widehat{\mathbf{u}}_k^{(q)} \leftarrow \widehat{\mathbf{u}}_k^{(q),2},\ q \in [p], \qquad \widehat{\mathscr{U}}_k \leftarrow \widehat{\mathbf{u}}_k^{(1)} \otimes \cdots \otimes \widehat{\mathbf{u}}_k^{(p)},$$

$$\check{\Sigma} \leftarrow \check{\Sigma} \times_1 \widehat{Q}_k^{(1)} \cdots \times_p \widehat{Q}_k^{(p)} \times_{p+1} \widehat{Q}_k^{(1)} \cdots \times_{2p} \widehat{Q}_k^{(p)},$$

where $\widehat{Q}_k^{(q)} := \mathbf{I} - \widehat{\mathbf{u}}_k^{(q)} (\widehat{\mathbf{u}}_k^{(q)})^\top$.

end

output: $\widehat{\mathscr{U}}_1, \widehat{\mathscr{U}}_2, \ldots, \widehat{\mathscr{U}}_r$.

Theorem 8. *Assume all conditions of Theorem 2 holds. In Algorithm 1, assume*

$$M \big/ \mathrm{poly}(r_0) \to +\infty,$$

where $\mathrm{poly}(r_0)$ *is a polynomial of* r_0.

Define λ_* *to be the maximum number that satisfies: if* $\sigma_j^2 - \sigma_{j+1}^2 < \lambda_*$ *for* $k \le j \le k + r_0 - 2$, *where* $1 \le k \le r - r_0 + 2$, *then* $\sigma_{j+1}^2 - \sigma_{j+2}^2 \ge \lambda_*$ *(we let* $\sigma_{r+1} = 0$ *for simplicity).*

If λ_* *satisfies*

$$\left(\frac{d^{p/2}}{\sqrt{n}} + \frac{d^{(p+1)/2}}{n}\right)\sigma_0^2 + \sqrt{\frac{r_0}{n}}\left(\sigma_0\sigma_1\sqrt{d} + \sigma_1^2\right) < \frac{c\lambda_*}{\sqrt{r_0}} \tag{5.3}$$

where $c > 0$ *is a small constant, then the output of Algorithm 1,* $\widehat{\mathcal{U}}_1, \widehat{\mathcal{U}}_2, \ldots, \widehat{\mathcal{U}}_r$, *satisfies: there exists a permutation* π *over* $[r]$ *such that*

$$\max_{1\le q\le p} \sin\angle(\widehat{\mathbf{u}}_{\pi(k)}^{(q)}, \mathbf{u}_k^{(q)}) = O_p\left(\sqrt{\frac{d}{n}}\left(\frac{\sigma_0}{\sigma_k} + \frac{\sigma_0^2}{\sigma_k^2}\right)\right), \tag{5.4}$$

for all $k \in [r]$, *and hence*

$$\sin\angle(\widehat{\mathcal{U}}_{\pi(k)}, \mathcal{U}_k) = O_p\left(\sqrt{\frac{d}{n}}\left(\frac{\sigma_0}{\sigma_k} + \frac{\sigma_0^2}{\sigma_k^2}\right)\right).$$

The definition of λ_* in Theorem 8 has a simple meaning: count from anywhere in $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$ for r_0 steps, there is a gap of at least λ_*. This is not an eigengap condition, since we can always choose $r_0 = r$ and by definition, $\lambda_* = \sigma_r^2$, and then the condition in Theorem 8 becomes a condition about σ_r.

Although Algorithm 1 only requires a two-step iteration after the initialization, in real-life scenarios we run a couple times of iterations for all the random starts, and then for the chosen initialization we run initialization until the change is small enough, or maximum number of iterations is reached. The number of multiple starts required by Theorem 8 may seem restrictive, but in real-life scenarios, it is more common that the multiplicity of $\{\sigma_k, k \in [r]\}$ is small, so that the number of multiple starts can be set to a smaller number.

5.2 Simulation Studies

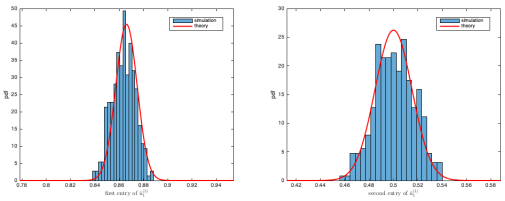
We first present a set of simulation studies to illustrate the finite-sample behavior of the sample PCs. These experiments are specifically designed to assess the role of bias correction, and robustness to deviation from the normal distribution. Throughout this section, unless otherwise noted, samples were generated according to the spike covariance model (3.1) with $p = 2$, e.g., each $\mathcal{X}_i$ is a matrix. Since the two modes are exchangeable, we only focus on the first mode $q = 1$ for brevity. We also fixed the number of spikes at $r = 2$. In each case, we shall set the singular values $\sigma_1 = \sigma_2$. In other words, for each of our examples, the usual PCA (with stringing) will not be able to identify the PCs because of the multiplicity. As mentioned before, without loss of generality and for the sake of brevity, we reordered $\check{\mathbf{u}}_1^{(1)}$ and $\check{\mathbf{u}}_2^{(1)}$ such that $\sin\angle(\check{\mathbf{u}}_1^{(1)}, \mathbf{u}_1) \leq \sin\angle(\check{\mathbf{u}}_2^{(1)}, \mathbf{u}_1)$. In addition, we replaced $\check{\mathbf{u}}_k^{(1)}$ with $-\check{\mathbf{u}}_k^{(1)}$ whenever $\langle \check{\mathbf{u}}_k^{(1)}, \mathbf{u}_k^{(1)} \rangle < 0$. For low-dimensional setup, $\tilde{\mathbf{u}}_1^{(1)}$ and $\tilde{\mathbf{u}}_2^{(1)}$ are treated similarly.

In the first set of experiments, we considered a low-dimensional setup with $d_1 = d_2 = 10$, $n = 200$, $\sigma_1 = \sigma_2 = 2$, and the true PCs were given by

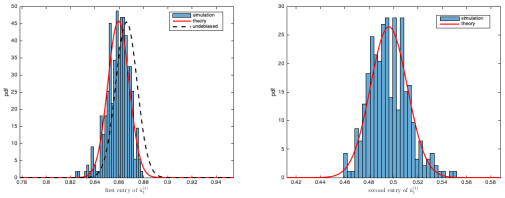
$$\begin{aligned} \mathbf{u}_1^{(1)} &= (\sqrt{3}/2, 1/2, 0, \ldots, 0)^\top, \quad \mathbf{u}_1^{(2)} = (1, 0, \ldots, 0)^\top, \\ \mathbf{u}_2^{(1)} &= (-1/2, \sqrt{3}/2, 0, \ldots, 0)^\top, \quad \mathbf{u}_2^{(2)} = (0, 1, 0, \ldots, 0)^\top. \end{aligned} \tag{5.5}$$

Figure 5.1a reports the histograms of the first two (nonzero) entries of $\tilde{\mathbf{u}}_1^{(1)}$ based on 300 simulation runs. The histograms are overlaid with the asymptotic distributions derived in Theorem 3. The agreement between the two confirms the accuracy of the asymptotic distribution when the dimensionality is low.

To demonstrate the need and effectiveness of bias correction, we increased the dimension to $d_1 = d_2 = 50$. Correspondingly we set $n = 400$ and $\sigma_1 = \sigma_2 = 3$. We repeated the experiment another 300 times and as before, Figure 5.1b reports the histograms of the first two entries of $\check{\mathbf{u}}_1^{(1)}$ along with the asymptotic distribution derived in Theorem 5, plotted in red lines. The dashed black line overlaid with the histogram of the first entries corresponds to the asymptotic distribution



(a) $d_1 = d_2 = 10$



(b) $d_1 = d_2 = 50$

Figure 5.1: Multiway PCA for data generated from normal distribution.

without bias correction as given by Theorem 3. It is clear that in this setting, debiasing is necessary and the bias correction of Theorem 5 indeed leads to a more precise approximation of the finite sample distribution.

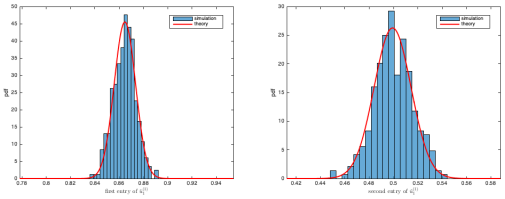
Our next set of simulations aims to explore the robustness of our approach to deviation from normality. To this end, $\{\theta_k, k \in [r]\}$ and the entries of $\mathscr{E}$ were simulated independently from Poisson(1) − 1 (so that they still have mean 0 and variance 1). Again we set $n = 400$ and $\sigma_1 = \sigma_2 = 3$. Figures 5.2a and 5.2b summarize results based on 300 runs, for dimensions $d_1 = d_2 = 10$ and $d_1 = d_2 = 50$, respectively. We overlay them with the theoretical asymptotic distributions given by Theorems 3 and 5. The results are qualitatively similar to those from the previous setting.

5.3 World Bank Data

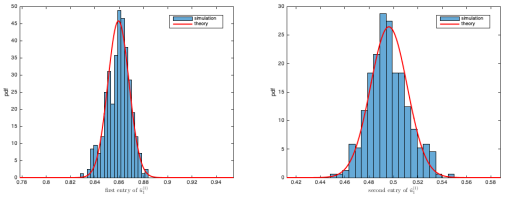
We now consider a real-world data example – the open source global development data from the World Bank[1]. The world Bank offers access to annual country-level data of a number of development indicators. In particular, we shall focus on the following nine most common and important economic and demographic indicators:

`GDP`: gross domestic product (GDP) based on purchasing power parity;

[1] https://data.worldbank.org/



(a) $d_1 = d_2 = 10$



(b) $d_1 = d_2 = 50$

Figure 5.2: Multiway PCA for data generated from Poisson distribution.

`Import`: import volume index (year 2000=100);

`Export`: export volume index (year 2000=100);

`CO2`: total CO2 emissions, in kilo-ton;

`CPI`: Consumer price index (year 2010 = 100);

`Life Span`: Life expectancy at birth;

`Urban Population`: Urban population, percentage of total population;

`Tourism`: number of international inbound tourists;

`Birth Rate`: Birth rate, crude (per 1,000 people).

Yearly data for these indicators have been recorded and we focus on data from Year 2000 through 2018, as considerable data are missing outside this range. We also discarded countries that have more than 5% of missing data in our analysis, resulting in a total of 160 countries under consideration.

These indicators are all positive but of vastly different magnitudes. To this end, a log transformation was first applied. Each log-transformed indicator was then standardized so that the

log-transformed indicator has a mean 0 and a mean absolute deviation 1 for all countries. The use of mean absolute deviation, instead of variance, for standardization allows more robust analysis in the presence of outlying observations. Denote by $\mathbf{X}_{k,t,i}$ the resulting indicator i for country k at time t. There remain a handful of missing values and for convenience, they are replaced with 0 in our analysis. The data tensor $\mathbf{X}$ is of dimensions $160 \times 19 \times 9$. Each frontal slice

$$\mathcal{X}_k = \mathbf{X}_{k,\cdot,\cdot}$$

corresponds to a country and is a 19×9 matrix. Note that its ambient dimension is $19 \times 9 = 171$ and greater than the number of countries so it is problematic to apply the usual PCA with stringing. Accounting for the multiway structure, we can consider the multiway PCs of the form

$$\mathcal{U}_k = \mathbf{u}_k^{(1)} \otimes \mathbf{u}_k^{(2)} \in \mathbb{R}^{19\times 9}.$$

These PC carry a clear meaning: each $\mathcal{U}_k$ represents a shared *development pattern*, where $\mathbf{u}_k^{(1)}$ is the corresponding shared *temporal trend*, and $\mathbf{u}_k^{(2)}$ is the corresponding *comovement pattern*.

Figure 5.3 plots the estimated leading PC along both modes, namely $\check{\mathbf{u}}_1^{(1)}$ and $\check{\mathbf{u}}_k^{(2)}$, together with the 95% confidence intervals for each of their coordinates. It is by far the most significant component, explaining 57.6% of the total variation. It is also evident from the temporal component that the first PC describes a roughly constant growth trend. The only year with a decrease is 2008 when the Global Financial Crisis took place. Correspondingly, except for the entry corresponding to birth rate, all other entries of $\check{\mathbf{u}}_k^{(2)}$ are positive. This suggests a general economic development during this period, with the birth rate in decline.

Similarly, Figure 5.4 shows the second multiway PC in the two modes along with their 95% confidence bands. This PC captures a change of developmental direction at the year of 2008. In particular, CPI, life span, urban population, and tourism steadily decreased prior to 2008 but reversed course after the financial crisis. In contrast, GDP, import, export, CO2 emission and birth rate followed an opposite pattern. There are many plausible explanations for this pattern. It is

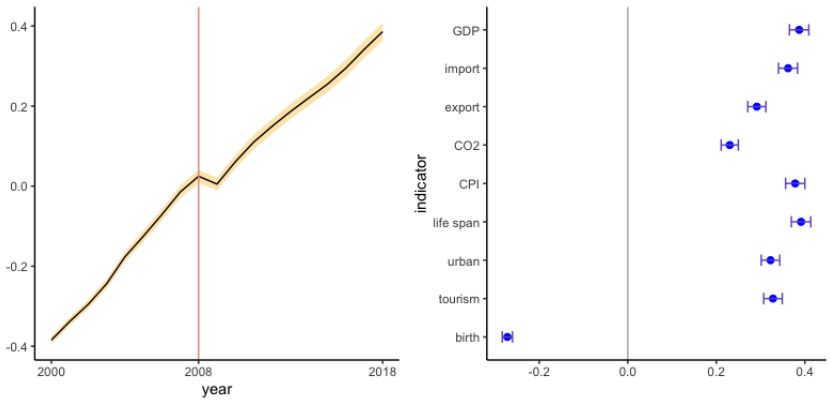


Figure 5.3: The first PC, general economic development: $\mathbf{u}_1^{(1)}$ and $\mathbf{u}_1^{(2)}$ plotted with 95% confidence intervals.

possible that the quantitative easing policies applied by most major economies since 2008 led to growth in the domestic market, thus enhancing the life-quality indicators. It is also possible that the growing inequality after 2008, also caused by quantitative easing among other factors (see, e.g., Montecino and Epstein 2015), has in turn caused the increase in life quality among the upper and the upper middle class. The tourism indicator is the number of international inbound tourists, which most likely is driven by the upper middle class and beyond. The continuous increase in life expectancy in the USA is also reported to be driven primarily by the well-off (see, e.g., Chetty et al. 2016).

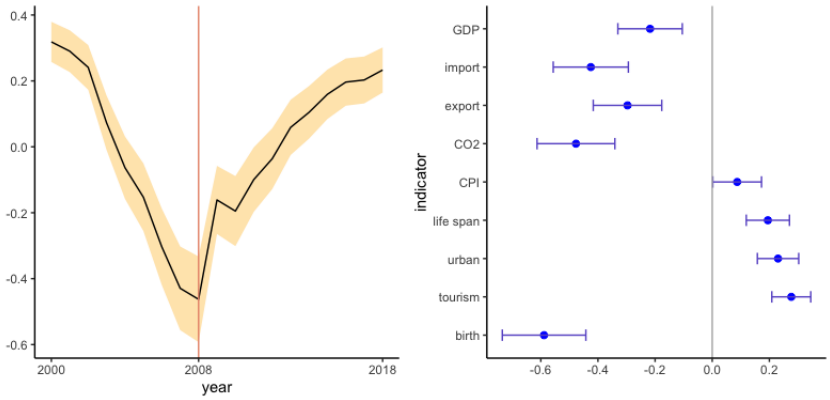


Figure 5.4: The second PC, life quality: $\mathbf{u}_2^{(1)}$ and $\mathbf{u}_2^{(2)}$ plotted with 95% confidence intervals.

Finally, Figure 5.5 shows the third multiway PC. We begin to see much wider confidence intervals as the signal becomes weaker. In fact, only the period around 2008 are significantly

different from zero, and likewise, the entries corresponding to life span, urbanization, and tourism are statistically insignificant. This indicates that these patterns likely focus on the impact of the 2008 financial crisis: it caused an immediate economic downturn but recovered not long after.

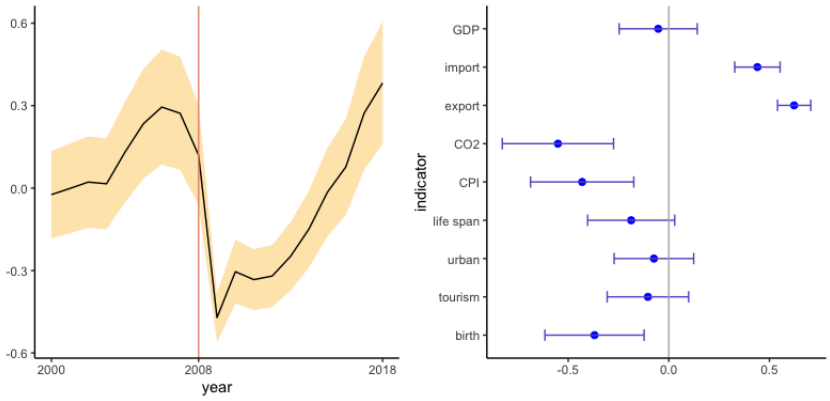


Figure 5.5: The third PC, international trade: $\mathbf{u}_3^{(1)}$ and $\mathbf{u}_3^{(2)}$ plotted with 95% confidence intervals.

5.4 NYC Bike Rental Data

Another data example we considered is the Citibike trip data[2]. In particular, all the Citibike trips from January 1, 2018 to December 31, 2019, on weekdays (522 days in total) that started in Manhattan and lasted for at least 60 seconds were used in our analysis. During this period, there are 35 zip codes in Manhattan with at least one Citibike station. There are a total of 29,515,527 trips and we form a data tensor $\mathbf{Y}$ of dimension $522 \times 24 \times 35$ where Y_{kij} denotes the number of trips starting during the ith hour of the kth day from the jth zip code.

The number of counts at different zip codes are of drastically different magnitudes, and the total counts during the two years also display a clear seasonal trend. To facilitate our analysis, we standardized the counts from each zip code j at each day k so that they have mean 0 and mean absolute deviation 1. As in the previous example, a direct application of the usual PCA can be misleading as the ambient dimension of the daily observation is $24 \times 35 = 840$ and greater than the

[2]https://ride.citibikenyc.com/system-data

number of days. Nonetheless, it is helpful to consider multiway PCs of the form

$$\mathcal{U}_k = \mathbf{u}_k^{(1)} \otimes \mathbf{u}_k^{(2)} \in \mathbb{R}^{24\times 35},$$

where $\mathbf{u}_k^{(1)}$ captures the time-of-the-day effect of bike rental, and $\mathbf{u}_k^{(2)}$ the location pattern.

Figure 5.6 plots the first multiway PC. The spatial pattern clearly indicates that this represents an overall pattern across Manhattan with all 35 entries of $\mathbf{u}_k^{(2)}$ being estimated as positive. The temporal pattern indicates that bike rental strongly coincides with the rush hours with two peaks during the morning and afternoon rush hours. The blank area downtown is zip code 10006, the big blank rectangular is Central Park, the small blank underneath is zip code 10020, and the blank area to the north of Central Park has zip codes 10030 and 10031. At the time of the recorded period, no Citibike station existed in these areas.

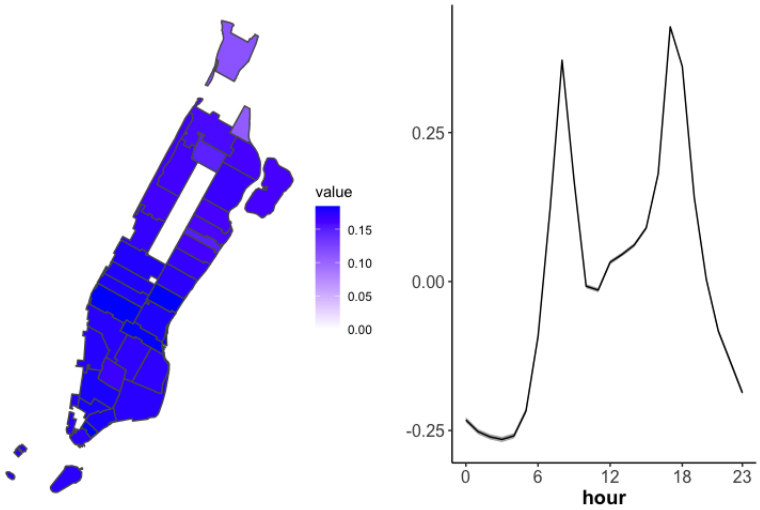


Figure 5.6: The first PC: overall pattern.

The second PC, as shown in Figure 5.7, reveals differences in rental patterns across neighborhoods. While the first PC suggests increased rental activities both in the morning and afternoon rush hours, the second PC captures the difference between morning and evening rental patterns as indicated by the positive peak during the evening rush hours and the negative peak during the morning rush hours. As such, a neighborhood with positive loadings may see more evening rentals

than morning rentals. These are the downtown Financial District, Lower Manhattan, and Midtown, largely corresponding to the business area of Manhattan. On the other hand, zip codes corresponding to negative loadings represent mostly residential areas of Manhattan, including the East Village, Upper West Side, and Upper East Side.

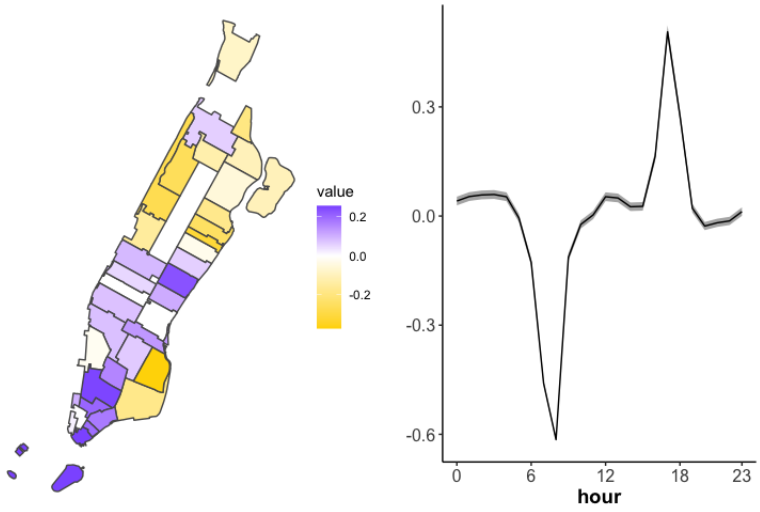


Figure 5.7: The second PC: rush hour differences.

Figure 5.8 depicts the third PC. The temporal pattern has a narrow and tall peak during the afternoon rush hours suggesting that this PC captures the subtle spatial difference during this time of the day. In particular, the zip codes with large positive values (purple color) are the area around Wall Street (the small purple block in Lower Manhattan), the area around Grand Central Terminal, and an area in Upper East Side. The negative zip codes in this pattern include the areas around SoHo, Greenwich Village, and Harlem.

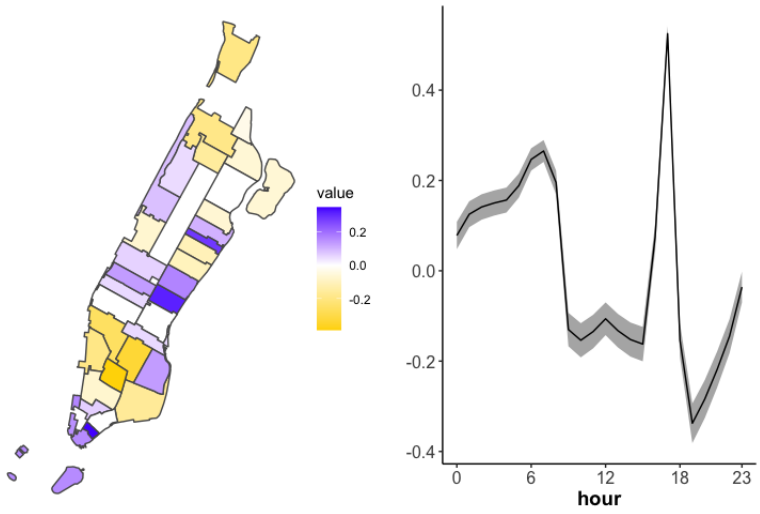


Figure 5.8: The third PC: afternoon rush hour details.

Chapter 6: Proofs

6.1 Notations and Preliminary Bounds

Write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. For a positive integer n, let $[n] := \{1, 2, \ldots, n\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$, denote $\|\mathbf{x}\|$ to be its ℓ_2-norm, $\|\mathbf{x}\|_1$ to be its ℓ_1-norm, and $\|\mathbf{x}\|_\infty = \max_i |x_i|$ to be its ℓ_∞-norm. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if $\exists C, \exists M$, such that $\forall n > M$, $|a_n| \leq C|b_n|$. Write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. For two sequences of real-valued random variables X_n and Y_n, write $X_n = O_p(Y_n)$ if $X_n = R_n Y_n$ and R_n is uniformly tight. Write $X_n = o_p(Y_n)$ if $X_n = R_n Y_n$ and $R_n \xrightarrow{p} 0$. For linear subspace U of $\mathbb{R}^d$, denote P_U and $P_U^\perp$ to be the orthogonal projection onto U and its orthogonal complement $U^\perp$, respectively. For a non-zero vector $u \in \mathbb{R}^d$, denote $P_u := P_{\text{span}\{u\}}$ and $P_u^\perp := P^\perp_{\text{span}\{u\}}$.

For an order-k tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_k}$, define its *tensor operator norm* as:

$$\|\mathcal{T}\| := \sup_{\mathbf{u}_j \in \mathbb{R}^{d_j}, \|\mathbf{u}_j\|=1} \mathcal{T}(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k). \tag{6.1}$$

Specifically, when $k = 2$ so that $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2}$ is a matrix, $\|\mathcal{T}\|$ is the matrix spectral norm of $\mathcal{T}$. For tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_k}$, write

$$\|\mathcal{T}\|_{\max} = \max_{i_1, \ldots, i_k} \left|\mathcal{T}_{i_1, \ldots, i_k}\right|$$

to be its ℓ_∞-norm.

With a slight abuse of notation, the mode q product of $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_k}$ with a vector $\mathbf{a} \in \mathbb{R}^{d_q}$, denoted by $\mathcal{T} \times_q \mathbf{a}$, is defined as an order-$(k-1)$ tensor of size $d_1 \times \ldots d_{q-1} \times d_{q+1} \cdots \times d_k$, with

elements

$$[\mathcal{T} \times_q \mathbf{a}]_{i_1 \ldots i_{q-1} i_{q+1} \ldots i_k} = \sum_{i_q=1}^{d_q} \mathcal{T}_{i_1 \ldots i_q \ldots i_k} \mathbf{a}_{i_q}.$$

Write

$$\widehat{\Sigma}_\theta = \frac{1}{n} \sum_{i=1}^{n} \theta_i \otimes \theta_i, \qquad \widehat{\Sigma}_{\mathscr{E}} = \frac{1}{n} \sum_{i=1}^{n} \mathscr{E}_i \otimes \mathscr{E}_i,$$

and

$$\widehat{\Sigma}_{\theta,\mathscr{E}} = \frac{1}{n} \sum_{i=1}^{n} \theta_i \otimes \mathscr{E}_i,$$

the sample covariance matrices of θ, $\mathscr{E}$ and between them respectively. Correspondingly denote by Σ_θ, $\Sigma_{\mathscr{E}}$ and $\Sigma_{\theta,\mathscr{E}}$ their population counterpart. It is clear $\Sigma_{\theta,\mathscr{E}} = 0$. Recall also that

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \mathscr{X}_i \otimes \mathscr{X}_i.$$

and

$$\Sigma = \sum_{l=1}^{r} \sigma_l \mathbf{u}_l^{(1)} \otimes \cdots \otimes \mathbf{u}_l^{(p)} \otimes \mathbf{u}_l^{(1)} \otimes \cdots \otimes \mathbf{u}_l^{(p)} + \sigma_0^2 \mathcal{J}$$

are the sample and population covariance matrices of $\mathscr{X}$.

The proof relies on the following technical lemmas.

Lemma 1. *There exists a numerical constant $C > 0$ such that for any $t \geq 1$,*

$$\|\widehat{\Sigma} - \Sigma\| \leq C(\sigma_1^2 + \sigma_0^2) \max\left\{\sqrt{\frac{d}{n}}, \frac{d}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\right\},$$

$$\|\widehat{\Sigma}_{\theta,\mathscr{E}}\| \leq C\sigma_0 \max\left\{\sqrt{\frac{d}{n}}, \frac{d}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\right\},$$

$$\|\widehat{\Sigma}_{\mathscr{E}} - \Sigma_{\mathscr{E}}\| \leq C\sigma_0^2 \max\left\{\sqrt{\frac{d}{n}}, \frac{d}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\right\},$$

and

$$\|\widehat{\Sigma}_\theta - \Sigma_\theta\| \leq C \max\left\{\sqrt{\frac{r}{n}}, \frac{r}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\right\},$$

with probability at least $1 - e^{-t}$.

Note that we shall use C to denote a constant that may take different values at each appearance. We shall also make use the following bounds:

Lemma 2. *There exists a numerical constant* $C > 0$ *such that for any* $t > 0$*,*

$$\|\widehat{\Sigma}_\theta - \Sigma_\theta\|_{\max} \leq C \max\left\{\sqrt{\frac{\log r}{n}}, \frac{\log r}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\right\},$$

$$\max_{1\leq l_1,l_2\leq r}\left|\widehat{\Sigma}_{\mathscr{E},\theta}(\mathbf{u}_{l_1}^{(1)}, \mathbf{u}_{l_2}^{(2)}, \ldots, \mathbf{u}_{l_2}^{(p)}, \mathbf{e}_{l_2})\right| \leq C\sigma_0 \max\left\{\sqrt{\frac{\log r}{n}}, \frac{\log r}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\right\}$$

where $\mathbf{e}_{l_2}$ *is the* l_2*th canonical basis of* $\mathbb{R}^r$*, and*

$$\max_{1\leq l\leq r}\left|(\widehat{\Sigma}_{\mathscr{E}} - \Sigma_{\mathscr{E}})(\mathbf{u}_l^{(1)}, \ldots, \mathbf{u}_l^{(p)}, \mathbf{u}_l^{(1)}, \ldots, \mathbf{u}_l^{(p)})\right| \leq C\sigma_0^2 \max\left\{\sqrt{\frac{\log r}{n}}, \frac{\log r}{n}, \sqrt{\frac{t}{n}}, \frac{t}{n}\right\},$$

with probability at least $1 - e^{-t}$.

Both Lemmas are well known and follow immediately from an application of union bounds and χ^2 tail bounds. See, e.g., Vershynin 2010.

6.2 Proof of Theorems 1 and 2

Theorem 1 follows immediately from Theorem 2 and it suffices to prove the latter. For brevity, we shall focus on the case when $d \leq n$ and r diverges with n. Denote by $\mathcal{E}$ the event that

$$\|\widehat{\Sigma}_\theta - \Sigma_\theta\| \leq C\sqrt{\frac{r}{n}}, \quad \text{and} \quad \sigma_0^{-2}\|\widehat{\Sigma}_{\mathscr{E}} - \Sigma_{\mathscr{E}}\|, \sigma_0^{-1}\|\widehat{\Sigma}_{\theta,\mathscr{E}}\| \leq C\sqrt{\frac{d}{n}}$$

and

$$\|\widehat{\Sigma}_\theta - \Sigma_\theta\|_{\max},\ \sigma_0^{-1}\max_{1\leq l_1,l_2\leq r}\left|\widehat{\Sigma}_{\mathscr{E},\theta}(\mathbf{u}_{l_1}^{(1)}, \mathbf{u}_{l_2}^{(2)}, \ldots, \mathbf{u}_{l_2}^{(p)}, \mathbf{e}_{l_2})\right| \leq C\sqrt{\frac{\log r}{n}}$$

By Lemmas 1 and 2, $\mathcal{E}$ holds with probability tending to one. It suffices to proceed conditional on the event $\mathcal{E}$.

As noted, the kth sample PCs may not correspond to the kth population PCs because we do not assume the existence of eigengap and σ_ks may not even be distinct. Nonetheless, we can match the sample PCs with population PCs as follows. Define

$$\pi(1) = \text{argmax}_{1\le l\le r}\left\{\sigma_l^2\left|\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_1^{(q)}\rangle\right|\right\},$$

and for $k > 1$,

$$\pi(k) := \text{argmax}_{l\notin\pi([k-1])}\left\{\sigma_l^2\left|\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right|\right\}.$$

The goal is to show that with high probability,

$$\eta_k := \max_{1\le q\le p}\sin\angle(\mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}) \le C\left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2}\right)\max\left\{\sqrt{\frac{d}{n}}, \frac{d}{n}\right\} =: \delta_k, \tag{6.2}$$

for $k = 1, \ldots, r$. Our proof proceeds by induction over k. To facilitate the induction, we shall also prove that

$$\begin{aligned}\tilde{\eta}_k &:= \max_{1\le q\le p}\max_{l\notin\pi([k])}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle \\ &\le C\left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2}\right)^2\max\left\{\frac{d}{n}, \frac{d^2}{n^2}\right\} + C\left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2}\right)\sqrt{\frac{\log r}{n}} \\ &=: \tilde{\delta}_k.\end{aligned} \tag{6.3}$$

In addition to (6.2) and (6.3), we shall also prove that

$$\sigma_{\pi(k)}^2 \ge \max_{l\notin\pi([k])}\sigma_l^2(1 - C\delta_k^2). \tag{6.4}$$

This immediately implies that

$$\sum_{l=1}^{k}\tilde{\delta}_k^2 \le C\delta_k^2 \quad \text{and} \quad \max_{1\le l\le k}\{\sigma_{\pi(l)}\delta_l\} \le C\sigma_{\pi(k)}\delta_k,$$

by the taking the constant c_0 in (3.3) small enough. We shall make use of these bounds repeatedly.

As noted, we shall proceed by induction over k. In particular, we shall denote by $\delta_0 = \tilde{\delta}_0 = 0$ so that the the base case holds trivially when $k = 0$. Now assume the induction hypotheses (6.2) and (6.3) holds for $1, \ldots, k-1$. We want to show that they continue to hold for k. The general architect of the argument is similar to that for the base case, but additional challenges arise with the need to control the impact of estimation error of $\widehat{\mathbf{u}}_l^{(q)}$s for $1 \le l < k$.

Denote by $\widehat{\mathcal{P}}^{(q)}$ the projection matrix onto the linear space spanned by $\{\widehat{\mathbf{u}}_1^{(q)}, \ldots, \widehat{\mathbf{u}}_{k-1}^{(q)}\}$, for $q \in [p]$. Note that in the case when $k = 1$, $\widehat{\mathcal{P}}^{(q)} = \mathbf{0}_{d\times d}$. Then

$$\begin{aligned}(\widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) &= \operatorname{argmax}_{\substack{\|\mathbf{w}^{(q)}\|\le 1, \langle \mathbf{w}^{(q)}, \widehat{\mathbf{u}}_l^{(q)}\rangle = 0, \\ \forall l \le k-1, q\in[p]}} \widehat{\Sigma}(\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(p)}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(p)}) \\ &= \operatorname{argmax}_{\|\mathbf{w}^{(q)}\|=1, \forall q\in[p]} \widehat{\Sigma}(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)}, \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)})\end{aligned}$$

where $\widehat{\mathcal{P}}_\perp^{(q)} = I - \widehat{\mathcal{P}}^{(q)}$. Observe that

$$\widehat{\mathcal{P}}_\perp^{(q)} \tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(q-1)}, \cdot, \widehat{\mathbf{u}}_k^{(q+1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) \propto \widehat{\mathbf{u}}_k^{(q)},$$

where $\tilde{\Sigma} = \widehat{\Sigma} - \sigma_0^2 \mathcal{I}$. This implies that

$$\langle \widehat{\mathbf{u}}_k^{(q)}, \mathbf{w}\rangle = \frac{\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(q-1)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}, \widehat{\mathbf{u}}_k^{(q+1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})}{\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})}.$$

In particular, for $l \ne k$,

$$\sin\angle(\widehat{\mathbf{u}}_k^{(q)}, \mathbf{u}_l^{(q)}) = \frac{\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(q-1)}, \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q+1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})}{\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})},$$

and

$$\sin\angle(\widehat{\mathbf{u}}_k^{(q)}, \mathbf{u}_k^{(q)}) = \frac{\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(q-1)}, \widehat{\mathcal{P}}_\perp^{(q)}(\widehat{\mathbf{u}}_k^{(q)} - \langle \widehat{\mathbf{u}}_k^{(q)}, \mathbf{u}_{\pi(k)}^{(q)}\rangle \mathbf{u}_{\pi(k)}^{(q)}), \widehat{\mathbf{u}}_k^{(q+1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})}{\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})},$$

We shall derive lower bounds for the nominators and an upper bound for the denominator. It suffices to consider the case $q = 1$. Other indices can be treated in an identical fashion.

Lower Bound for $\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})$ Denote by $\tilde{\Sigma} = \Sigma - \sigma_0^2 \mathcal{J}$. Observe that

$$\begin{aligned}
& \tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) \\
\geq & \max_{l \notin \pi([k-1])} \tilde{\Sigma}(\widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_l^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{u}_l^{(p)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_l^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{u}_l^{(p)}) \\
\geq & \max_{l \notin \pi([k-1])} (\Sigma - \sigma_0^2 \mathcal{J}) \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_l^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{u}_l^{(p)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_l^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{u}_l^{(p)}) \\
& - \sup_{\|\mathbf{w}^{(q)}\| \leq 1, 1 \leq q \leq p} \left| (\widehat{\Sigma} - \Sigma)(\widehat{\mathcal{P}}_\perp^{(1)} \mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{w}^{(p)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{w}^{(p)}) \right|. \qquad (6.5)
\end{aligned}$$

Next we bound the two terms on the rightmost hand side.

Starting with the first term, note that for any $l \notin \pi([k-1])$,

$$\begin{aligned}
& (\Sigma - \sigma_0^2 \mathcal{J})(\widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_l^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{u}_l^{(p)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_l^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{u}_l^{(p)}) \\
= & \sum_{1 \leq l' \leq r} \left[\sigma_{l'}^2 \prod_{q=1}^{p} \langle \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{u}_l^{(q)}, \mathbf{u}_{l'}^{(q)} \rangle^2 \right] \\
\geq & \sigma_l^2 \prod_{q=1}^{p} \langle \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{u}_l^{(q)}, \mathbf{u}_l^{(q)} \rangle^2 \\
= & \sigma_l^2 \prod_{q=1}^{p} \|\widehat{\mathcal{P}}_\perp^{(q)} \mathbf{u}_l^{(q)}\|^4.
\end{aligned}$$

By the induction hypothesis (6.3),

$$\|\widehat{\mathcal{P}}^{(q)} \mathbf{u}_l^{(q)}\|^2 = \sum_{1 \leq l' < k} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_{l'}^{(q)} \rangle^2 \leq \sum_{1 \leq l' < k} \tilde{\delta}_{l'}^2 \leq C \delta_k^2.$$

By taking the constant c_0 of (3.3) small enough, we can ensure that

$$\max_{l \notin \pi([k-1])} (\Sigma - \sigma_0^2 \mathcal{J})(\widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_l^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{u}_l^{(p)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_l^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{u}_l^{(p)}) \geq (1 - C\delta_k^4) \tau^2. \qquad (6.6)$$

where

$$\tau^2 = \max_{l \notin \pi([k-1])} \sigma_l^2.$$

Next we derive a bound for

$$\sup_{\|\mathbf{w}^{(q)}\| \le 1, 1 \le q \le p} (\Sigma - \widehat{\Sigma})(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)}, \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)}).$$

Note that

$$\begin{aligned}
& (\Sigma - \widehat{\Sigma})(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)}, \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)}) \\
= & \sum_{l_1=1}^{r}\sum_{l_2=1}^{r} \sigma_{l_1}\sigma_{l_2}\left(\widehat{\Sigma}_{\theta,l_1l_2} - \Sigma_{\theta,l_1l_2}\right)\left(\prod_{q=1}^{p}\langle \mathbf{u}_{l_1}^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\rangle\right)\left(\prod_{q=1}^{p}\langle \mathbf{u}_{l_2}^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\rangle\right) \\
& + \frac{2}{n}\sum_{i=1}^{n}\sum_{l=1}^{r}\sigma_l\theta_{il}\left(\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\rangle\right)\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)}) \\
& + \left(\frac{1}{n}\sum_{i=1}^{n}[\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)})]^2 - \sigma_0^2\prod_{q=1}^{p}\|\widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\|^2\right). \qquad (6.7)
\end{aligned}$$

We bound the three terms on the right hand side separately.

The first term can be bounded by

$$\begin{aligned}
& \left|\sum_{l_1=1}^{r}\sum_{l_2=1}^{r}\sigma_{l_1}\sigma_{l_2}\left(\frac{1}{n}\sum_{i=1}^{n}\theta_{il_1}\theta_{il_2}\right)\left(\prod_{q=1}^{p}\langle \mathbf{u}_{l_1}^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\rangle\right)\left(\prod_{q=1}^{p}\langle \mathbf{u}_{l_2}^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\rangle\right)\right. \\
& \qquad \left. - \sum_{l=1}^{r}\sigma_l^2\left(\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\rangle\right)^2\right| \\
\le & \|\widehat{\Sigma}_\theta - I_r\|\left[\sum_{l=1}^{r}\sigma_l^2\left(\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\rangle\right)^2\right] \\
= & \|\widehat{\Sigma}_\theta - I_r\|\left[\sum_{l\in\pi([k-1])}\sigma_l^2\left(\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\rangle\right)^2 + \sum_{l\notin\pi([k-1])}\sigma_l^2\left(\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\rangle\right)^2\right],
\end{aligned}$$

Recall that for any $l \in \pi([k-1])$,

$$|\langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{w}^{(q)} \rangle| = |\langle \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{u}_l^{(q)}, \mathbf{w}^{(q)} \rangle| \le \|\widehat{\mathcal{P}}_\perp^{(q)} \mathbf{u}_l^{(q)}\| \le \delta_l.$$

Therefore,

$$\begin{aligned}\sum_{l \in \pi([k-1])} \sigma_l^2 \left(\prod_{q=1}^p \langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{w}^{(q)} \rangle \right)^2 &\le \max_{l \in \pi([k-1])} \sigma_l^2 \left(\prod_{q=2}^p \langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{w}^{(q)} \rangle \right)^2 \\ &\le \max_{1 \le l < k} \{\sigma_{\pi(l)}^2 \delta_l^{2p-2}\} \le \max_{1 \le l < k} \{\sigma_{\pi(l)}^2 \delta_l^2\} \le \tau^2,\end{aligned}$$

by taking c_0 of (3.3) small enough. On the other hand,

$$\sum_{l \notin \pi([k-1])} \sigma_l^2 \left(\prod_{q=1}^p \langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{w}^{(q)} \rangle \right)^2 \le \max_{l \notin \pi([k-1])} \sigma_l^2 = \tau^2.$$

This implies that

$$\left| \sum_{l_1=1}^r \sum_{l_2=1}^r \sigma_{l_1} \sigma_{l_2} \left(\widehat{\Sigma}_{\theta, l_1 l_2} - \Sigma_{\theta, l_1 l_2} \right) \left(\prod_{q=1}^p \langle \mathbf{u}_{l_1}^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{w}^{(q)} \rangle \right) \left(\prod_{q=1}^p \langle \mathbf{u}_{l_2}^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{w}^{(q)} \rangle \right) \right| \le C\tau^2 \sqrt{\frac{r}{n}}. \quad (6.8)$$

Similarly, the second term can be bounded by

$$\begin{aligned}&\left| \frac{1}{n} \sum_{i=1}^n \sum_{l=1}^r \sigma_l \theta_{il} \left(\prod_{q=1}^p \langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{w}^{(q)} \rangle \right) \mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)} \mathbf{w}^{(1)}, \ldots, \widehat{\mathcal{P}}_\perp^{(p)} \mathbf{w}^{(p)}) \right| \\ \le\ & \|\widehat{\Sigma}_{\theta, \mathscr{E}}\| \left[\sum_{l=1}^r \sigma_l^2 \left(\prod_{q=1}^p \langle \mathbf{u}_l^{(q)}, \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{w}^{(q)} \rangle \right)^2 \right]^{1/2} \\ \le\ & C\tau\sigma_0 \sqrt{\frac{d}{n}}. \end{aligned} \quad (6.9)$$

Finally, the third term can be bounded by

$$\left|\frac{1}{n}\sum_{i=1}^{n}[\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)},\ldots,\widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)})]^2-\sigma_0^2\prod_{q=1}^{p}\|\widehat{\mathcal{P}}_\perp^{(q)}\mathbf{w}^{(q)}\|^2\right|\le\|\widehat{\Sigma}_{\mathscr{E}}-\Sigma_{\mathscr{E}}\|\le C\sigma_0^2\sqrt{\frac{d}{n}}. \quad (6.10)$$

Combing (6.7)-(6.10), we get

$$\begin{aligned}&\sup_{\|\mathbf{w}^{(q)}\|\le 1,1\le q\le p}\left|(\widehat{\Sigma}-\Sigma)(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)},\ldots,\widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}^{(1)},\ldots,\widehat{\mathcal{P}}_\perp^{(p)}\mathbf{w}^{(p)})\right|\\ \le\;& C\tau^2\sqrt{\frac{r}{n}}+C(\sigma_0\tau+\sigma_0^2)\sqrt{\frac{d}{n}}.\end{aligned}$$

Together with (6.5), this implies

$$\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)},\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\ge\tau^2\left(1-C\delta_k^4-C\sqrt{\frac{r}{n}}\right)-C(\sigma_0\tau+\sigma_0^2)\sqrt{\frac{d}{n}}-C\sigma_0^2\delta_k^2, \quad (6.11)$$

by taking c_0 of (3.3) small enough.

Upper Bounds for $\tilde{\Sigma}(\widehat{\mathcal{P}}_\perp^{(1)}(\widehat{\mathbf{u}}_k^{(1)} - \langle \widehat{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)}\rangle \mathbf{u}_{\pi(k)}^{(1)}), \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})$ Observe that for any $\mathbf{w}$ orthogonal to $\mathbf{u}_{\pi(k)}^{(1)}$, we have

$$
\begin{aligned}
& \tilde{\Sigma}(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) \\
= \; & \sigma_{\pi(k)}^2 \left(\frac{1}{n}\sum_{i=1}^n \theta_{i\pi(k)}^2\right)\left(\prod_{q=2}^p \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle \mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle \\
& + \sum_{l\neq\pi(k)} \sigma_{\pi(k)}\sigma_l \left(\frac{1}{n}\sum_{i=1}^n \theta_{i\pi(k)}\theta_{il}\right)\left(\prod_{q=2}^p \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p \langle \mathbf{u}_{l}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle \mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle \\
& + \sum_{l\neq\pi(k)} \sigma_l\sigma_{\pi(k)} \left(\frac{1}{n}\sum_{i=1}^n \theta_{il}\theta_{i\pi(k)}\right)\left(\prod_{q=2}^p \langle \mathbf{u}_{l}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle \mathbf{u}_{l}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle \\
& + \sum_{l_1,l_2\neq\pi(k)} \sigma_{l_1}\sigma_{l_2} \left(\frac{1}{n}\sum_{i=1}^n \theta_{il_1}\theta_{il_2}\right)\left(\prod_{q=2}^p \langle \mathbf{u}_{l_1}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p \langle \mathbf{u}_{l_2}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle \mathbf{u}_{l_1}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle \\
& + \frac{1}{n}\sum_{i=1}^n \left[\sum_{l=1}^r \sigma_l\theta_{il}\left(\prod_{q=2}^p \langle \mathbf{u}_{l}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})\langle \mathbf{u}_{l}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle\right] \\
& + \frac{1}{n}\sum_{i=1}^n \left[\sum_{l=1}^r \mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})\sigma_l\theta_{il}\left(\prod_{q=1}^p \langle \mathbf{u}_{l}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right] \\
& + \frac{1}{n}\sum_{i=1}^n \mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) - \sigma_0^2\langle \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}, \widehat{\mathbf{u}}_k^{(p)}\rangle. \qquad (6.12)
\end{aligned}
$$

Again each term on the right hand side needs to be bounded carefully.

The first term on the right hand side of (6.12) can be bounded by

$$
\begin{aligned}
& \left|\sigma_{\pi(k)}^2 \left(\frac{1}{n}\sum_{i=1}^n \theta_{i\pi(k)}^2\right)\left(\prod_{q=2}^p \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle \mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle\right| \\
& \qquad \leq \tau^2 \|\widehat{\Sigma}_\theta\|_{\max} |\langle \mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_\perp \mathbf{w}\rangle|.
\end{aligned}
$$

In particular, when $\mathbf{w} = \widehat{\mathbf{u}}_k^{(1)} - \langle \widehat{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)} \rangle \mathbf{u}_{\pi(k)}^{(1)}$, we have

$$\begin{aligned} |\langle \mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{w} \rangle| &= \left| \langle \widehat{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)} \rangle (1 - \langle \mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_{\pi(k)}^{(1)} \rangle) \right| \\ &\leq 1 - \langle \mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{u}_{\pi(k)}^{(1)} \rangle \\ &= \|\widehat{\mathcal{P}}^{(1)} \mathbf{u}_{\pi(k)}^{(1)}\|^2 \leq \sum_{1 \leq l < k} \tilde{\delta}_l^2 \leq C \delta_k^2, \end{aligned}$$

by taking c_0 small enough. Thus,

$$\left| \sigma_{\pi(k)}^2 \left(\frac{1}{n} \sum_{i=1}^n \theta_{i\pi(k)}^2 \right) \left(\prod_{q=2}^p \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle \right) \left(\prod_{q=1}^p \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle \right) \langle \mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{w} \rangle \right| \leq C \tau^2 \delta_k^2. \tag{6.13}$$

The second term on the right hand side of (6.12) can be bounded as follows:

$$\begin{aligned} &\left| \sum_{l \neq \pi(k)} \sigma_{\pi(k)} \sigma_l \left(\frac{1}{n} \sum_{i=1}^n \theta_{i\pi(k)} \theta_{il} \right) \left(\prod_{q=2}^p \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle \right) \left(\prod_{q=1}^p \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle \right) \langle \mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_\perp^{(1)} \mathbf{w} \rangle \right| \\ \leq\ & \tau \|\widehat{\Sigma}_\theta - I\| \left(\sum_{l \neq \pi(k)} \sigma_l^2 \prod_{q=1}^p \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle^2 \right)^{1/2} \\ \leq\ & \tau \|\widehat{\Sigma}_\theta - I\| \left(\sum_{l \neq \pi(k)} \langle \mathbf{u}_l^{(1)}, \widehat{\mathbf{u}}_k^{(1)} \rangle^2 \right)^{1/2} \max_{l \neq \pi(k)} \left\{ \sigma_l \prod_{q=2}^p |\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle| \right\} \\ =\ & \tau \delta_k \|\widehat{\Sigma}_\theta - I\| \max_{l \neq \pi(k)} \left\{ \sigma_l \prod_{q=2}^p |\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle| \right\} \\ \leq\ & \tau \delta_k \|\widehat{\Sigma}_\theta - I\| \max \left\{ \max_{l \in \pi([k-1])} \left\{ \sigma_l \prod_{q=2}^p |\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle| \right\}, \max_{l \notin \pi([k])} \left\{ \sigma_l \prod_{q=2}^p |\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle| \right\} \right\}. \end{aligned}$$

Note that

$$\max_{l \in \pi([k-1])} \left\{ \sigma_l \prod_{q=2}^p |\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle| \right\} \leq \max_{l \in \pi([k-1])} \left\{ \sigma_l \prod_{q=2}^p |\langle \widehat{\mathcal{P}}_\perp^{(q)} \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle| \right\} \leq \max_{1 \leq l < k} \{ \sigma_{\pi(l)} \delta_l^{p-1} \} \leq 2\tau \delta_k^{p-1},$$

and

$$\max_{l\notin\pi([k])}\left\{\sigma_l\prod_{q=2}^{p}|\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle|\right\}\le\tau\delta_k^{p-1}.$$

We get

$$\left|\sum_{l\neq\pi(k)}\sigma_{\pi(k)}\sigma_l\left(\frac{1}{n}\sum_{i=1}^{n}\theta_{i\pi(k)}\theta_{il}\right)\left(\prod_{q=2}^{p}\langle\mathbf{u}_{\pi(k)}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right|\le C\tau^2\delta_k^p\sqrt{\frac{r}{n}}. \tag{6.14}$$

And similarly, when $\mathbf{w}=\widehat{\mathbf{u}}_k^{(1)}-\langle\widehat{\mathbf{u}}_k^{(1)},\mathbf{u}_{\pi(k)}^{(1)}\rangle\mathbf{u}_{\pi(k)}^{(1)}$, we bound the third term on the right hand side of (6.12) by

$$\begin{aligned}
&\left|\sum_{l\neq\pi(k)}\sigma_l\sigma_{\pi(k)}\left(\frac{1}{n}\sum_{i=1}^{n}\theta_{il}\theta_{i\pi(k)}\right)\left(\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^{p}\langle\mathbf{u}_{\pi(k)}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle\right|\\
\le\;&\tau\|\widehat{\Sigma}_\theta-I\|\left(\sum_{l\neq\pi(k)}\sigma_l^2\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle^2\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)^{1/2}\\
\le\;&\tau\|\widehat{\Sigma}_\theta-I\|\left(\sum_{l\neq\pi(k)}\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle^2\right)^{1/2}\max_{l\neq\pi(k)}\left\{\sigma_l\prod_{q=2}^{p}|\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right\}\\
\le\;&C\tau^2\delta_k^p\sqrt{\frac{r}{n}},
\end{aligned}\tag{6.15}$$

where in the last inequality we used the fact that

$$\sum_{l\neq\pi(k)}\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle^2\le\|\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\|^2\le\|\mathbf{w}\|^2\le\delta_k^2;$$

the fourth term by

$$\begin{aligned}
&\left|\sum_{l_1,l_2\neq\pi(k)}\sigma_{l_1}\sigma_{l_2}\left(\frac{1}{n}\sum_{i=1}^{n}\theta_{il_1}\theta_{il_2}\right)\left(\prod_{q=2}^{p}\langle\mathbf{u}_{l_1}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^{p}\langle\mathbf{u}_{l_2}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_{l_1}^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle\right|\\
\le\;&\|\widehat{\Sigma}_\theta\|\left(\sum_{l\neq\pi(k)}\sigma_l^2\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle^2\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)^{1/2}\left(\sum_{l\neq\pi(k)}\sigma_l^2\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)^{1/2}\\
\le\;&C\tau^2\delta_k^{2(p-1)};
\end{aligned}\tag{6.16}$$

the fifth term by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l\neq\pi(k)}\sigma_l\theta_{il}\left(\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle\right]\\
=&\sum_{l\neq\pi(k)}\left\{\sigma_l\left(\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle\left[\frac{1}{n}\sum_{i=1}^{n}\theta_{il}\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\right]\right\}\\
\leq&\|\widehat{\Sigma}_{\theta,\mathscr{E}}\|\cdot\left[\sum_{l\neq\pi(k)}\left(\sigma_l^2\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w}\rangle^2\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)\right]^{1/2}\\
\leq&C\sigma_0\tau\delta_k^{p-1}\sqrt{\frac{d}{n}};
\end{aligned}\tag{6.17}$$

and the sixth term by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l=1}^{r}\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\sigma_l\theta_{il}\left(\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right]\\
\leq&\frac{1}{n}\sum_{i=1}^{n}\left[\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\sigma_{\pi(k)}\theta_{i\pi(k)}\left(\prod_{q=1}^{p}\langle\mathbf{u}_{\pi(k)}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right]\\
&+\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l\neq\pi(k)}\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\sigma_l\theta_{il}\left(\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right]\\
\leq&\tau\|\widehat{\Sigma}_{\theta,\mathscr{E}}\|+\|\widehat{\Sigma}_{\theta,\mathscr{E}}\|\left(\sum_{l\neq\pi(k)}\sigma_l^2\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)^{1/2}\\
\leq&C\sigma_0\tau\sqrt{\frac{d}{n}}.
\end{aligned}\tag{6.18}$$

Finally the last term can be bounded by

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})-\sigma_0^2\langle\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{w},\widehat{\mathbf{u}}_k^{(1)}\rangle\right|\leq\|\widehat{\Sigma}_{\mathscr{E}}-\Sigma_{\mathscr{E}}\|\leq C\sigma_0^2\sqrt{\frac{d}{n}}.$$

Together with (6.13)-(6.18), we get

$$\tilde{\Sigma}(\widehat{\mathcal{P}}_{\perp}^{(1)}(\widehat{\mathbf{u}}_k^{(1)} - \langle\widehat{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)}\rangle\mathbf{u}_{\pi(k)}^{(1)}), \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) \leq C\tau^2\delta_k^2 + C(\sigma_0^2 + \sigma_0\tau)\sqrt{\frac{d}{n}}.$$

Upper Bounds for $\max_{l\notin\pi([k])} \tilde{\Sigma}(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})$ To derive the helper bound (6.3), we also need an upper bound for

$$\max_{l\notin\pi([k])} \tilde{\Sigma}(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}).$$

We shall follow a similar step by bound each term on the right hand side of (6.12), but now with $\mathbf{w} = \mathbf{u}_m^{(1)}$ ($m \notin \pi([k])$).

Specifically, the first term can be bounded by

$$\left|\sigma_{\pi(k)}^2\left(\frac{1}{n}\sum_{i=1}^n \theta_{i\pi(k)}^2\right)\left(\prod_{q=2}^p \langle\mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p \langle\mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)}\rangle\right| \leq \tau^2\|\widehat{\Sigma}_\theta\|_{\max}|\langle\mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)}\rangle|.$$

Note that

$$\langle\mathbf{u}_{\pi(k)}^{(1)}, \mathbf{u}_m^{(1)}\rangle = \langle\mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)}\rangle + \langle\mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)}\rangle = 0.$$

We get

$$|\langle\mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)}\rangle| = |\langle\mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)}\rangle| \leq \|\widehat{\mathcal{P}}\mathbf{u}_{\pi(k)}^{(1)}\|\|\widehat{\mathcal{P}}\mathbf{u}_m^{(1)}\| \leq \sum_{1\leq l<k} \tilde{\delta}_l^2 \leq C\delta_k^2,$$

by Cauchy-Schwartz inequality. This implies that

$$\left|\sigma_{\pi(k)}^2\left(\frac{1}{n}\sum_{i=1}^n \theta_{i\pi(k)}^2\right)\left(\prod_{q=2}^p \langle\mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p \langle\mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_{\pi(k)}^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)}\rangle\right| \leq C\tau^2\delta_k^2.$$

by taking c_0 small enough.

The second term can also be bounded by

$$\begin{aligned}
&\left|\sum_{l\neq\pi(k)}\sigma_{\pi(k)}\sigma_l\left(\frac{1}{n}\sum_{i=1}^n\theta_{i\pi(k)}\theta_{il}\right)\left(\prod_{q=2}^p\langle\mathbf{u}_{\pi(k)}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_{\pi(k)}^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle\right| \\
\leq\ & \tau\|\widehat{\Sigma}_\theta - I\|\,|\langle\mathbf{u}_{\pi(k)}^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle|\left(\sum_{l\neq\pi(k)}\sigma_l^2\prod_{q=1}^p\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)^{1/2} \\
\leq\ & \tau^2\delta_k^{p+1}\sqrt{\frac{r}{n}}.
\end{aligned}$$

We bound the third term by

$$\begin{aligned}
&\left|\sum_{l\neq\pi(k)}\sigma_l\sigma_{\pi(k)}\left(\frac{1}{n}\sum_{i=1}^n\theta_{il}\theta_{i\pi(k)}\right)\left(\prod_{q=2}^p\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p\langle\mathbf{u}_{\pi(k)}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle\right| \\
\leq\ & \left|\sigma_m\sigma_{\pi(k)}\left(\frac{1}{n}\sum_{i=1}^n\theta_{im}\theta_{i\pi(k)}\right)\left(\prod_{q=2}^p\langle\mathbf{u}_m^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p\langle\mathbf{u}_{\pi(k)}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_m^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle\right| \\
&+\left|\sum_{l\notin\{\pi(k),m\}}\sigma_l\sigma_{\pi(k)}\left(\frac{1}{n}\sum_{i=1}^n\theta_{il}\theta_{i\pi(k)}\right)\left(\prod_{q=2}^p\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^p\langle\mathbf{u}_{\pi(k)}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle\right|
\end{aligned}$$

The first term on the right hand side can be further bounded by

$$\tau^2\|\widehat{\Sigma}_\theta - \Sigma_\theta\|_{\max}\prod_{q=2}^p\left|\langle\mathbf{u}_m^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right| \leq C\tau^2\delta_k^{p-1}\sqrt{\frac{\log r}{n}}.$$

Now consider the second term:

$$
\begin{aligned}
&\left| \sum_{l \notin \{\pi(k), m\}} \sigma_l \sigma_{\pi(k)} \left(\frac{1}{n} \sum_{i=1}^{n} \theta_{il} \theta_{i\pi(k)} \right) \left(\prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle \right) \left(\prod_{q=1}^{p} \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle \right) \langle \mathbf{u}_l^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)} \mathbf{u}_m^{(1)} \rangle \right| \\
\leq \;& \tau \|\widehat{\Sigma}_\theta - I\| \left(\sum_{l \notin \{\pi(k), m\}} \sigma_l^2 \langle \mathbf{u}_l^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)} \mathbf{u}_m^{(1)} \rangle^2 \prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle^2 \right)^{1/2} \\
\leq \;& \tau \|\widehat{\Sigma}_\theta - I\| \left(\sum_{l \in \pi([k-1])} \sigma_l^2 \langle \mathbf{u}_l^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)} \mathbf{u}_m^{(1)} \rangle^2 \prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle^2 \right. \\
&\qquad \left. + \sum_{l \notin \pi([k]) \cup \{m\}} \sigma_l^2 \langle \mathbf{u}_l^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)} \mathbf{u}_m^{(1)} \rangle^2 \prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle^2 \right)^{1/2}.
\end{aligned}
$$

The first term in the bracket on the rightmost hand side can be bounded by

$$
\begin{aligned}
&\sum_{l \in \pi([k-1])} \sigma_l^2 \|\widehat{\mathcal{P}}_{\perp}^{(1)} \mathbf{u}_l^{(1)}\|^2 \prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle^2 \\
\leq \;& \left(\sum_{l \in \pi([k-1])} \langle \mathbf{u}_l^{(2)}, \widehat{\mathbf{u}}_k^{(2)} \rangle^2 \right) \left(\max_{l \in \pi([k-1])} \sigma_l^2 \|\widehat{\mathcal{P}}_{\perp}^{(1)} \mathbf{u}_l^{(1)}\|^2 \prod_{q=3}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle^2 \right) \\
\leq \;& \delta_k^2 \left(\max_{l \in \pi([k-1])} \sigma_l^2 \delta_l^2 \delta_k^{2(p-2)} \right) \leq C \tau^2 \delta_k^{2p};
\end{aligned}
$$

the second term by

$$
\|\widehat{\mathcal{P}}^{(1)} \mathbf{u}_m^{(1)}\|^2 \left(\max_{l \notin \pi([k]) \cup \{m\}} \sigma_l^2 \prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle^2 \right) \leq C \tau^2 \delta_k^{2p},
$$

so that

$$
\begin{aligned}
&\left| \sum_{l \neq \pi(k)} \sigma_l \sigma_{\pi(k)} \left(\frac{1}{n} \sum_{i=1}^{n} \theta_{il} \theta_{i\pi(k)} \right) \left(\prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle \right) \left(\prod_{q=1}^{p} \langle \mathbf{u}_{\pi(k)}^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle \right) \langle \mathbf{u}_l^{(1)}, \widehat{\mathcal{P}}_{\perp}^{(1)} \mathbf{u}_m^{(1)} \rangle \right| \\
\leq \;& C \left(\tau^2 \delta_k^{p-1} \sqrt{\frac{\log r}{n}} + \tau^2 \delta_k^p \sqrt{\frac{r}{n}} \right).
\end{aligned}
$$

Similar to before, the fourth term on the right hand side of (6.12) can be bounded by

$$\begin{aligned}
&\left|\sum_{l_1,l_2\neq\pi(k)}\sigma_{l_1}\sigma_{l_2}\left(\frac{1}{n}\sum_{i=1}^{n}\theta_{il_1}\theta_{il_2}\right)\left(\prod_{q=2}^{p}\langle\mathbf{u}_{l_1}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\left(\prod_{q=1}^{p}\langle\mathbf{u}_{l_2}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_{l_1}^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle\right| \\
\leq\ & \|\widehat{\Sigma}_\theta\|\left(\sum_{l\neq\pi(k)}\sigma_l^2\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle^2\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)^{1/2}\left(\sum_{l\neq\pi(k)}\sigma_l^2\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)^{1/2} \\
\leq\ & C\tau^2\delta_k^{2(p-1)};
\end{aligned}$$

and the fifth term by

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l\neq\pi(k)}\sigma_l\theta_{il}\left(\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle\right]\right| \\
=\ & \left|\sum_{l\neq\pi(k)}\left\{\sigma_l\left(\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle\left[\frac{1}{n}\sum_{i=1}^{n}\theta_{il}\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\right]\right\}\right| \\
\leq\ & \|\widehat{\Sigma}_{\theta,\mathscr{E}}\|\cdot\left(\sum_{l\neq\pi(k)}\sigma_l^2\langle\mathbf{u}_l^{(1)},\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}\rangle^2\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)^{1/2} \\
\leq\ & C\sigma_0\tau\delta_k^{p-1}\sqrt{\frac{d}{n}}.
\end{aligned}$$

We now turn to the sixth term on the right hand side of (6.12). Write

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l=1}^{r}\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\sigma_l\theta_{il}\left(\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right] \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l\neq\pi(k)}\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\sigma_l\theta_{il}\left(\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right] \\
&+\frac{1}{n}\sum_{i=1}^{n}\left[\mathscr{E}_i(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\sigma_{\pi(k)}\theta_{i\pi(k)}\left(\prod_{q=1}^{p}\langle\mathbf{u}_{\pi(k)}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right].
\end{aligned}$$

The first term again can be bounded by

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l\neq\pi(k)}\mathscr{E}_i(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\sigma_l\theta_{il}\left(\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right]\right| \\
\leq\ & \|\widehat{\Sigma}_{\theta,\mathscr{E}}\|\left(\sum_{l\neq\pi(k)}\sigma_l^2\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)^{1/2}\leq C\sigma_0\tau\delta_k^{p-1}\sqrt{\frac{d}{n}}.
\end{aligned}$$

For the second term, note that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathscr{E}_i(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\sigma_{\pi(k)}\theta_{i\pi(k)}\left(\prod_{q=1}^{p}\langle\mathbf{u}_{\pi(k)}^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right)\right]\right| \\
\leq\ & \tau\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathscr{E}_i(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\theta_{i\pi(k)}\right]\right| \\
\leq\ & \tau\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathscr{E}_i(\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\theta_{i\pi(k)}\right]\right|+\tau\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathscr{E}_i(\widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\theta_{i\pi(k)}\right]\right| \qquad (6.19)
\end{aligned}$$

where the second inequality follows from triangular inequality. As before,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left[\mathscr{E}_i(\widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\theta_{i\pi(k)}\right]\right|\leq\|\widehat{\Sigma}_{\theta,\mathscr{E}}\|\|\widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)}\|\leq C\sigma_0\delta_k\sqrt{\frac{d}{n}}.$$

To bound the first term on the rightmost hand side of (6.19), write

$$\widehat{\mathbf{u}}_k^{(q)}=\alpha_q\mathbf{u}_{\pi(k)}+\mathbf{v}^{(q)}$$

where $\alpha_q=\langle\mathbf{u}_{\pi(k)},\widehat{\mathbf{u}}_k\notin\{\pi(k),m\}\notin\{\pi(k),m\}\notin\{\pi(k),m\}^{(q)}\rangle$ and $\|\mathbf{v}^{(q)}\|=\sqrt{1-\alpha_q^2}\leq\delta_k$.

Then

$$
\begin{aligned}
& \mathscr{E}_i(\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) \\
= \;& \alpha_p \mathscr{E}_i(\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p-1)}, \mathbf{u}_{\pi(k)}^{(p)}) + \mathscr{E}_i(\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p-1)}, \mathbf{v}^{(p)}) \\
= \;& \alpha_p \alpha_{p-1} \mathscr{E}_i(\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p-2)}, \mathbf{u}_{\pi(k)}^{(p-1)}, \mathbf{u}_{\pi(k)}^{(p)}) \\
& + \alpha_p \mathscr{E}_i(\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p-2)}, \mathbf{v}^{(p-1)}, \mathbf{u}_{\pi(k)}^{(p)}) \\
& + \mathscr{E}_i(\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p-1)}, \mathbf{v}^{(p)}) \\
= \;& \cdots\cdots \\
= \;& \left(\prod_{q=2}^{p} \alpha_q \right) \mathscr{E}_i(\mathbf{u}_m^{(1)}, \mathbf{u}_{\pi(k)}^{(2)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}) \\
& + \cdots + \mathscr{E}_i(\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p-1)}, \mathbf{v}^{(p)}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^{n} \left[\mathscr{E}_i(\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) \theta_{i\pi(k)} \right] \right| \\
\leq \;& \left| \frac{1}{n} \sum_{i=1}^{n} \left[\mathscr{E}_i(\mathbf{u}_m^{(1)}, \mathbf{u}_{\pi(k)}^{(2)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}) \theta_{i\pi(k)} \right] \right| + \cdots + \left| \frac{1}{n} \sum_{i=1}^{n} \left[\mathscr{E}_i(\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p-1)}, \mathbf{v}^{(p)}) \theta_{i\pi(k)} \right] \right| \\
\leq \;& \left| \frac{1}{n} \sum_{i=1}^{n} \left[\mathscr{E}_i(\mathbf{u}_m^{(1)}, \mathbf{u}_{\pi(k)}^{(2)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}) \theta_{i\pi(k)} \right] \right| + C_p \delta_k \|\widehat{\Sigma}_{\theta,\mathscr{E}}\| \\
\leq \;& C \left(\sigma_0 \sqrt{\frac{\log r}{n}} + \sigma_0 \delta_k \sqrt{\frac{d}{n}} \right),
\end{aligned}
$$

so that the six term on the rightmost hand side of (6.12) can be upper bounded by

$$
\left| \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{l=1}^{r} \mathscr{E}_i(\widehat{\mathcal{P}}_{\perp}^{(1)} \mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) \sigma_l \theta_{il} \left(\prod_{q=1}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_k^{(q)} \rangle \right) \right] \right| \leq C \left(\sigma_0 \tau \sqrt{\frac{\log r}{n}} + \sigma_0 \tau \delta_k \sqrt{\frac{d}{n}} \right).
$$

Finally consider the seventh term:

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})-\langle\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(1)}\rangle\right| \\
\leq\ &\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})-\langle\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(1)}\rangle\right| \\
&+\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})-\langle\widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(1)}\rangle\right|.
\end{aligned}$$

Similar to before, the second term can be bounded by

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})-\sigma_0^2\langle\widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(1)}\rangle\right| \\
\leq\ &\|\widehat{\mathcal{P}}^{(1)}\mathbf{u}_m^{(1)}\|\|\widehat{\Sigma}_{\mathscr{E}}-\Sigma_{\mathscr{E}}\|\leq C\sigma_0^2\delta_k\sqrt{\frac{d}{n}};
\end{aligned}$$

the first term by

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})-\sigma_0^2\langle\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(1)}\rangle\right| \\
\leq\ &\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\mathbf{v}^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})-\sigma_0^2\langle\mathbf{u}_m^{(1)},\mathbf{v}^{(1)}\rangle\right| \\
&+\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\mathbf{u}_{\pi(k)}^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\right| \\
\leq\ &C\delta_k\|\widehat{\Sigma}_{\mathscr{E}}-\Sigma_{\mathscr{E}}\|+\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\mathbf{u}_m^{(1)},\mathbf{u}_{\pi(k)}^{(2)},\cdots,\mathbf{u}_{\pi(k)}^{(p)})\mathscr{E}_i(\mathbf{u}_{\pi(k)}^{(1)},\cdots,\mathbf{u}_{\pi(k)}^{(p)})\right| \\
&+\cdots+\left|\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\mathscr{E}_i(\mathbf{u}_{\pi(k)}^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p-1)},\mathbf{v}^{(p)})\right| \\
\leq\ &C\left(\sigma_0^2\delta_k\sqrt{\frac{d}{n}}+\sigma_0^2\sqrt{\frac{\log r}{n}}\right).
\end{aligned}$$

Putting all seven upper bounds together, we have

$$\max_{l\notin\pi([k])}\tilde{\Sigma}(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_m^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)},\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\leq C\left(\tau^2\delta_k^2+(\tau^2\delta_k^{p-1}+\sigma_0\tau+\sigma_0^2)\sqrt{\frac{\log r}{n}}\right).$$

Finishing Up We first verify (6.4). Note that

$$\begin{aligned}
&(\Sigma-\Sigma_{\mathscr{E}})(\widehat{\mathbf{u}}_k^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)},\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\\
=&\ \sum_{l=1}^{r}\left(\sigma_l^2\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle^2\right)\\
\leq&\ \max_{1\leq l\leq r}\left\{\sigma_l^2\left|\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right|\right\}\cdot\left(\sum_{l=1}^{r}\left|\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right|\right)\\
\leq&\ \max_{1\leq l\leq r}\left\{\sigma_l^2\left|\prod_{q=1}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_k^{(q)}\rangle\right|\right\}.
\end{aligned}$$

where the last inequality follows from Cauchy-Schwartz inequality. Therefore, by definition,

$$\sigma_{\pi(k)}^2\geq(\Sigma-\Sigma_{\mathscr{E}})(\widehat{\mathbf{u}}_k^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)},\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\tag{6.20}$$

On the other hand,

$$\begin{aligned}
&(\Sigma-\Sigma_{\mathscr{E}})(\widehat{\mathbf{u}}_k^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)},\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\\
\geq&\ \tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)},\widehat{\mathbf{u}}_k^{(2)},\ldots,\widehat{\mathbf{u}}_k^{(p)},\widehat{\mathbf{u}}_k^{(1)},\ldots,\widehat{\mathbf{u}}_k^{(p)})\\
&-\sup_{\|\mathbf{w}^{(q)}\|\leq1,1\leq q\leq p}\left|(\widehat{\Sigma}-\Sigma)(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{w}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{w}^{(p)},\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{w}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{w}^{(p)})\right|\\
\geq&\ \tilde{\Sigma}(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_{l_*}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{u}_{l_*}^{(p)},\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_{l_*}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{u}_{l_*}^{(p)})\\
&-\sup_{\|\mathbf{w}^{(q)}\|\leq1,1\leq q\leq p}\left|(\widehat{\Sigma}-\Sigma)(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{w}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{w}^{(p)},\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{w}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{w}^{(p)})\right|\\
\geq&\ (\Sigma-\Sigma_{\mathscr{E}})(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_{l_*}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{u}_{l_*}^{(p)},\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{u}_{l_*}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{u}_{l_*}^{(p)})\\
&-2\sup_{\|\mathbf{w}^{(q)}\|\leq1,1\leq q\leq p}\left|(\widehat{\Sigma}-\Sigma)(\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{w}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{w}^{(p)},\widehat{\mathcal{P}}_{\perp}^{(1)}\mathbf{w}^{(1)},\ldots,\widehat{\mathcal{P}}_{\perp}^{(p)}\mathbf{w}^{(p)})\right|
\end{aligned}$$

where l_* is the index such that $\sigma_{l_*}^2 = \tau^2$. Following the same derivation as before, we have

$$(\Sigma - \Sigma_{\mathscr{E}})(\widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) \geq \tau^2 \left(1 - C\sqrt{\frac{r}{n}} - C\delta_k\right).$$

Together with (6.20), we get

$$\sigma_{\pi(k)}^2 \geq \tau^2 \left(1 - C\sqrt{\frac{r}{n}} - C\delta_k\right).$$

Combing the lower bound for $\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})$ and upper bound for $\tilde{\Sigma}(\widehat{\mathcal{P}}_\perp^{(1)}(\widehat{\mathbf{u}}_k^{(1)} - \langle \widehat{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)} \rangle \mathbf{u}_{\pi(k)}^{(1)}), \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})$, we have

$$\begin{aligned}
\eta_k &\leq \frac{C\tau^2\delta_k^2 + C(\sigma_0^2 + \sigma_0\tau)\sqrt{\frac{d}{n}}}{\tau^2\left(1 - C\delta_k^4 - C\sqrt{\frac{r}{n}}\right) - C(\sigma_0\tau + \sigma_0^2)\sqrt{\frac{d}{n}} - C\sigma_0^2\delta_k^2} \\
&\leq C\delta_k^2 + C\left(\frac{\sigma_0}{\tau} + \frac{\sigma_0^2}{\tau^2}\right)\sqrt{\frac{d}{n}} \\
&\leq C\left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2}\right)\sqrt{\frac{d}{n}} = \delta_k.
\end{aligned}$$

Similarly, Combing the lower bound for $\tilde{\Sigma}(\widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})$ and upper bound for $\max_{l \notin \pi([k])} \tilde{\Sigma}(\widehat{\mathcal{P}}_\perp^{(1)}\mathbf{u}_m^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \widehat{\mathbf{u}}_k^{(1)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})$, we get

$$\begin{aligned}
\tilde{\eta}_k &\leq \frac{C\left(\tau^2\delta_k^2 + (\tau^2\delta_k^{p-1} + \sigma_0\tau + \sigma_0^2)\sqrt{\frac{\log r}{n}}\right)}{\tau^2\left(1 - C\delta_k^4 - C\sqrt{\frac{r}{n}}\right) - C(\sigma_0\tau + \sigma_0^2)\sqrt{\frac{d}{n}} - C\sigma_0^2\delta_k^2} \\
&\leq C\delta_k^2 + C\left(\delta_k + \frac{\sigma_0}{\tau} + \frac{\sigma_0^2}{\tau^2}\right)\sqrt{\frac{\log r}{n}} \\
&\leq C\delta_k^2 + C\left(\frac{\sigma_0}{\sigma_{\pi(k)}} + \frac{\sigma_0^2}{\sigma_{\pi(k)}^2}\right)\sqrt{\frac{\log r}{n}} = \tilde{\delta}_k.
\end{aligned}$$

6.3 Proof of Theorems 3 and 4

Note that Theorem 3 can be viewed as special case of Theorem 4 and it suffices to prove Theorem 4. As before, we only need to consider the case when $q = 1$. Write

$$\mathbf{w} = \langle \mathbf{w}, \mathbf{u}_{\pi(k)}^{(1)} \rangle \mathbf{u}_{\pi(k)}^{(1)} + \tilde{\mathbf{w}}.$$

Then

$$\langle \tilde{\mathbf{u}}_k^{(1)}, \mathbf{w} \rangle - \langle \mathbf{u}_{\pi(k)}^{(1)}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{u}_{\pi(k)}^{(1)} \rangle \left(\langle \tilde{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)} \rangle - 1 \right) + \langle \tilde{\mathbf{u}}_k^{(1)}, \tilde{\mathbf{w}} \rangle.$$

Under the assumption $d = o(n^{1/2})$, by Lemma 1 and Theorem 2, it is not hard to see that

$$\sin \angle(\tilde{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)}) = O_p\left(\sqrt{\frac{d}{n}}\right), \tag{6.21}$$

so

$$1 - \langle \tilde{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)} \rangle \le 1 - \langle \tilde{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)} \rangle^2 = \sin^2 \angle(\tilde{\mathbf{u}}_k^{(1)}, \mathbf{u}_{\pi(k)}^{(1)}) = O_p\left(\frac{d}{n}\right) = o_p(n^{-1/2}).$$

Therefore it suffices to prove that

$$\sqrt{n}\left(\langle \tilde{\mathbf{u}}_k^{(1)}, \tilde{\mathbf{w}} \rangle - \langle \mathbf{u}_{\pi(k)}^{(1)}, \tilde{\mathbf{w}} \rangle \right) \to_d N\left(0, \|\tilde{\mathbf{w}}\|^2 \left(\frac{\sigma_0^2}{\sigma_{\pi(k)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(k)}^4} \right)\right).$$

Recall that $\tilde{\mathbf{u}}_k^{(1)}$ is the leading eigenvector of

$$\widehat{\Sigma}(\cdot, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \cdot, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}),$$

which is the same as the leading eigenvector of

$$\tilde{\Sigma}(\cdot, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \cdot, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}),$$

so

$$\tilde{\mathbf{u}}_k^{(1)} = \frac{\tilde{\Sigma}(\cdot, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})}{\tilde{\Sigma}(\tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})},$$

which implies

$$\langle \tilde{\mathbf{u}}_k^{(1)}, \tilde{\mathbf{w}} \rangle = \frac{\tilde{\Sigma}(\tilde{\mathbf{w}}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})}{\tilde{\Sigma}(\tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)})}.$$

We start with the nominator. Following an identical argument as that for (6.12) in Subsection 6.2, we have

$$\begin{aligned}
&\tilde{\Sigma}(\tilde{\mathbf{w}}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) - \tilde{\Sigma}(\tilde{\mathbf{w}}, \mathbf{u}_{\pi(k)}^{(2)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}, \mathbf{u}_{\pi(k)}^{(1)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}) \\
\leq\, & C\left(\tau^2\delta_k^2 + \tau^2\delta^{p+1}\sqrt{\frac{r}{n}} + \tau^2\delta^{p-1}\sqrt{\frac{r}{n}} + \tau^2\delta_k^{2(p-1)} + \sigma_0\tau\delta_k\sqrt{\frac{d}{n}}\right) = o(n^{-1/2})
\end{aligned}$$

with probability tending to one. So

$$\begin{aligned}
&\tilde{\Sigma}(\tilde{\mathbf{w}}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) \\
=&\tilde{\Sigma}(\tilde{\mathbf{w}}, \mathbf{u}_{\pi(k)}^{(2)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}, \mathbf{u}_{\pi(k)}^{(1)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}) + o_p(n^{-1/2}) \\
=&\sigma_{\pi(k)}\frac{1}{n}\sum_{i=1}^{n}\theta_{\pi(k)i}\mathscr{E}_i(\tilde{\mathbf{w}}, \mathbf{u}_{\pi(k)}^{(2)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\mathbf{u}_{\pi(k)}^{(1)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)})\mathscr{E}_i(\tilde{\mathbf{w}}, \mathbf{u}_{\pi(k)}^{(2)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}) - \mathbf{u}_{\pi(k)}^{(1)} + o_p(n^{-1/2}) \\
\to_d & N(0, (\sigma_0^4 + \sigma_0^2\sigma_{\pi(k)}^2)\|\tilde{\mathbf{w}}\|^2).
\end{aligned}$$

On the other hand, similar to the previous section,

$$\tilde{\Sigma}(\tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}, \tilde{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)}, \ldots, \widehat{\mathbf{u}}_k^{(p)}) = \sigma_{\pi(k)}^2(1 + o_p(1)).$$

Theorem 4 then follows from Slutsky's Theorem.

The claim in Theorem 3 that

$$\sqrt{n}\left[\mathrm{vec}(\tilde{\mathbf{U}}^{(q)}) - \mathrm{vec}(\mathbf{U}_{\pi}^{(q)})\right]$$
$$\xrightarrow{d} N\left(0, \mathrm{diag}\left(\left(\frac{\sigma_0^2}{\sigma_{\pi(1)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(1)}^4}\right)\boldsymbol{\mathcal{P}}_{\mathbf{u}_{\pi(1)}^{(q)}}^{\perp}, \ldots, \left(\frac{\sigma_0^2}{\sigma_{\pi(r)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(r)}^4}\right)\boldsymbol{\mathcal{P}}_{\mathbf{u}_{\pi(r)}^{(q)}}^{\perp}\right)\right)$$

follows from Theorem 4 and the fact that

$$\sigma_{\pi(k)}\frac{1}{n}\sum_{i=1}^{n}\theta_{\pi(k)i}\mathscr{E}_i(\cdot, \mathbf{u}_{\pi(k)}^{(2)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}) + \frac{1}{n}\sum_{i=1}^{n}\mathscr{E}_i(\mathbf{u}_{\pi(k)}^{(1)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)})\mathscr{E}_i(\cdot, \mathbf{u}_{\pi(k)}^{(2)}, \ldots, \mathbf{u}_{\pi(k)}^{(p)}) - \mathbf{u}_{\pi(k)}^{(1)}$$

are independent for any $k_1 \neq k_2$.

6.4 Proof of Theorems 5 and 6

Theorem 5 is a special case of Theorem 6 and it suffices to prove the latter. We first need to introduce a number of notations. Denote

$$\bar{\mathbb{P}}(\cdot) := \mathbb{P}(\cdot|\mathscr{X}_{n/2+1}, \mathscr{X}_{n/2+2}, \ldots, \mathscr{X}_n), \tag{6.22}$$

and

$$\bar{\mathbb{E}}(\cdot) := \mathbb{E}(\cdot|\mathscr{X}_{n/2+1}, \mathscr{X}_{n/2+2}, \ldots, \mathscr{X}_n), \tag{6.23}$$

the conditional probability and expectation given $\{\mathscr{X}_{n/2+1}, \mathscr{X}_{n/2+2}, \ldots, \mathscr{X}_n\}$, respectively.

Write

$$\widehat{\boldsymbol{\mathcal{P}}}_j^{(q)[1]} := \check{\mathbf{u}}_j^{(q)[1]} \otimes \check{\mathbf{u}}_j^{(q)[1]}, \qquad \boldsymbol{\mathcal{P}}_j^{(q)} = \mathbf{u}_j^{(q)} \otimes \mathbf{u}_j^{(q)} = \boldsymbol{\mathcal{P}}_{\mathbf{u}_j^{(q)}},$$

$$C_j^{(q)} := \frac{1}{\sigma_j^2}\left(I_{d_q} - \mathbf{u}_j^{(q)} \otimes \mathbf{u}_j^{(q)}\right),$$

$$z_{jk} := \mathscr{X}_k \times_1 \widehat{\mathbf{u}}_j^{(1)[2]} \cdots \times_{q-1} \widehat{\mathbf{u}}_j^{(q-1)[2]} \times_{q+1} \widehat{\mathbf{u}}_j^{(q+1)[2]} \cdots \times_p \widehat{\mathbf{u}}_j^{(p)[2]}, \qquad k = 1, \ldots, n/2,$$

and

$$M_j^{(q)[1]} := \sum_{l=1}^{r} \sigma_l^2 \left(\prod_{q' \neq q} \langle \mathbf{u}_l^{(q')}, \widehat{\mathbf{u}}_j^{(q')[2]} \rangle \right)^2 \mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)} + I_{d_q} = \sum_{l=1}^{r} \tilde{\sigma}_l^2 \mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)} + I_{d_q},$$

where

$$\tilde{\sigma}_l^2 = \sigma_l^2 \left(\prod_{q' \neq q} \langle \mathbf{u}_l^{(q')}, \widehat{\mathbf{u}}_j^{(q')[2]} \rangle \right)^2.$$

Furthermore, let

$$C_j^{(q)[1]} = \sum_{l \neq j} \frac{1}{\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2} \mathbf{u}_l^{(q)} \otimes \mathbf{u}_l^{(q)} + \frac{1}{\tilde{\sigma}_j^2} \left(I_{d_q} - \sum_{l=1}^{r} \mathbf{u}_l^{(q)} \otimes \mathbf{u}_l^{(q)} \right),$$

$$\mathcal{L}_j^{(q)[1]} = \frac{2}{n} \sum_{k=1}^{n/2} \left(C^{(q)[1]} z_{jk} \otimes \mathcal{P}_j^{(q)} z_{jk} + \mathcal{P}_j^{(q)} z_{jk} \otimes C^{(q)[1]} z_{jk} \right),$$

$$\mathcal{S}_j^{(q)[1]} = \widehat{\mathcal{P}}_j^{(q)[1]} - \mathcal{P}_j^{(q)} - \mathcal{L}_j^{(q)[1]}.$$

We use calligraphic capital letters for $\mathcal{P}_j^{(q)}$, $\widehat{\mathcal{P}}_j^{(q)[1]}$, $C_j^{(q)}$, $C_j^{(q)[1]}$, $\mathcal{L}_j^{(q)[1]}$ and $\mathcal{S}_j^{(q)[1]}$ to remind the readers that they are matrices with specific definitions.

Define

$$b_j^{(q)[1]} := \left\langle \bar{\mathbb{E}}(\mathcal{S}_j^{(q)[1]}) \mathbf{u}_j^{(q)}, \mathbf{u}_j^{(q)} \right\rangle. \tag{6.24}$$

and $b_j^{(q)[2]}$ is similarly defined. Finally, we define

$$b_j^{(q)} = \frac{\left\| \check{\mathbf{u}}_j^{(q),[1]} + \check{\mathbf{u}}_j^{(q),[2]} \right\|}{\sqrt{1 + b_j^{(q)[1]}} + \sqrt{1 + b_j^{(q)[2]}}} - 1.. \tag{6.25}$$

The proof is rather involved and we shall break it into several steps.

Step 1. We shall represent linear forms of $\check{\mathbf{u}}_j^{(q),[1]}$ as bilinear forms of $\widehat{\mathcal{P}}_j^{(q)[1]}$, and prove that

$$\sqrt{n}\Bigg(\big(1+b_{k_1}^{(q)}\big)\big\langle\check{\mathbf{u}}_{j_1}^{(q)},\mathbf{v}_1\big\rangle-\big\langle\mathbf{u}_{\pi(j_1)}^{(q)},\mathbf{v}_1\big\rangle,\big(1+b_{j_2}^{(q)}\big)\big\langle\check{\mathbf{u}}_{j_2}^{(q)},\mathbf{v}_2\big\rangle-\big\langle\mathbf{u}_{\pi(j_2)}^{(q)},\mathbf{v}_2\big\rangle,\ldots,$$
$$\big(1+b_{k_m}^{(q)}\big)\big\langle\check{\mathbf{u}}_{j_m}^{(q)},\mathbf{v}_m\big\rangle-\big\langle\mathbf{u}_{\pi(j_m)}^{(q)},\mathbf{v}_m\big\rangle\Bigg)\to_d N\left(\mathbf{0},\mathbf{\Gamma}\right), \tag{6.26}$$

where

$$\mathbf{\Gamma}_{i_1i_2}=\begin{cases}\left(\frac{\sigma_0^2}{\sigma^2_{\pi(j_{i_1})}}+\frac{\sigma_0^4}{\sigma^4_{\pi(j_{i_1})}}\right)\left\langle\mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{\pi(j_{i_1})}}\mathbf{v}_{i_1},\mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{\pi(j_{i_1})}}\mathbf{v}_{i_2}\right\rangle, & \text{if } j_{i_1}=j_{i_2},\\ 0, & \text{if } j_{i_1}\neq j_{i_2},\end{cases}$$

This also directly implies

$$\sqrt{n}\left(\big(1+b_j^{(q)}\big)\langle\check{\mathbf{u}}_j^{(q)},\mathbf{v}\rangle-\langle\mathbf{u}_{\pi(j)}^{(q)},\mathbf{v}\rangle\right)\to_d N\left(0,\left(\frac{\sigma_0^2}{\sigma^2_{\pi(j)}}+\frac{\sigma_0^4}{\sigma^4_{\pi(j)}}\right)\|\mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{\pi(j)}}\mathbf{v}\|^2\right). \tag{6.27}$$

Step 2. We prove that

$$\langle\check{\mathbf{u}}_j^{(q)},\mathbf{u}_{\pi(j)}^{(q)}\rangle=\frac{1}{\sqrt{1+\frac{d_q}{n}\left(\frac{\sigma_0^2}{\sigma^2_{\pi(j)}}+\frac{\sigma_0^4}{\sigma^4_{\pi(j)}}\right)}}+O_p\left(\frac{d^{3/2}}{n^{3/2}}\right). \tag{6.28}$$

Notice that by letting $\mathbf{v}=\mathbf{u}_{\pi(j)}^{(q)}$, (6.27) implies

$$\big(1+b_j^{(q)}\big)\langle\check{\mathbf{u}}_j^{(q)},\mathbf{u}_{\pi(j)}^{(q)}\rangle-1=o_p(n^{-1/2}).$$

Combine it with (6.28), we immediately have

$$b_j^{(q)} = \sqrt{1 + \frac{d_q}{n}\left(\frac{\sigma_0^2}{\sigma_{\pi(j)}^2} + \frac{\sigma_0^4}{\sigma_{\pi(j)}^4}\right)} - 1 + O_p\left(\frac{d^{3/2}}{n^{3/2}}\right) + o_p\left(\frac{1}{\sqrt{n}}\right),$$

Step 3. Finally, we show that

$$\sqrt{n}\left(\widehat{b}_j^{(q)} - b_j^{(q)}\right) \xrightarrow{p} 0. \tag{6.29}$$

For simplicity, in the rest of the proof we shall assume without loss of generality that the permutation π that matches $\widehat{\mathbf{u}}_j^{(q)}$ to $\mathbf{u}_{\pi(j)}^{(1)}$ is the identity. We shall also make repeated use of the following facts, oftentimes without explicit mentioning.

Similar to the proof of Theorem 2, write

$$\widehat{\Sigma}_{\mathscr{E}}^{[1]} = \frac{2}{n}\sum_{i=1}^{n/2} \mathscr{E}_i \otimes \mathscr{E}_i, \qquad \widehat{\Sigma}_{\theta,\mathscr{E}}^{[1]} = \frac{2}{n}\sum_{i=1}^{n/2} \theta_i \otimes \mathscr{E}_i.$$

By Lemmas 1 and 2, with probability tending to one,

$$\sigma_0^{-2}\|\widehat{\Sigma}_{\mathscr{E}}^{[1]} - \mathscr{I}\| \le C\sqrt{\frac{d}{n}},\ \sigma_0^{-1}\|\widehat{\Sigma}_{\theta,\mathscr{E}}^{[1]}\| \le C\sqrt{\frac{d}{n}}. \tag{6.30}$$

Let

$$\Delta^{(q),[2]} = \widehat{\mathbf{u}}_j^{(q),[2]} - \mathbf{u}_j^{(q)}.$$

By Theorem 2, we have

$$\max_{q\in[p]} \|\Delta^{(q),[2]}\| = O_p\left(\sqrt{\frac{d}{n}}\right)$$

Moreover, under the assumption $d = o(n)$, by Lemma 1 and Theorem 2, it is not hard to see that

$$\sin\angle(\check{\mathbf{u}}_j^{(1)[1]}, \mathbf{u}_j^{(1)}) = O_p\left(\sqrt{\frac{d}{n}}\right).$$

Denote

$$\delta^{(1)[2]} := \max\left\{\max_{q\in[p]} \|\Delta^{(q),[2]}\|, \|\check{\mathbf{u}}_j^{(1)[1]} - \mathbf{u}_j^{(1)}\|\right\},$$

combine the two bounds above, we have that under the assumptions for Theorem 6,

$$\delta^{(1)[2]} = O_p\left(\sqrt{\frac{d}{n}}\right). \tag{6.31}$$

6.4.1 Step 1.

Without loss of generality, for this step we assume $\sigma_0 = 1$. We only need to prove for the case $q = 1$, so within this step, we shall also write $\mathcal{P}_j = \mathcal{P}_j^{(1)} = \mathcal{P}_{\mathbf{u}_j^{(1)}}$ and $C_j = C_j^{(1)} = \frac{1}{\sigma_j^2}\left(I - \mathcal{P}_j\right)$ for simplicity.

Define

$$\eta_{jk}(\mathbf{v}) := \langle y_{jk}, \mathcal{P}_j \mathbf{u}_j^{(1)}\rangle\langle y_{jk}, C_j\mathbf{v}\rangle,$$

where

$$y_{jk} := \mathcal{X}_k \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)} = \sigma_j\theta_{jk}\mathbf{u}_j^{(1)} + \mathcal{E}_k \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}.$$

Recall that $\widehat{\mathcal{P}}_j^{(1)[1]} := \check{\mathbf{u}}_j^{(1),[1]} \otimes \check{\mathbf{u}}_j^{(1),[1]}$ and $\widehat{\mathcal{P}}_j^{(1)[2]} := \check{\mathbf{u}}_j^{(1),[2]} \otimes \check{\mathbf{u}}_j^{(1),[2]}$. Define

$$\rho_j(\mathbf{v})^{[1]} := \left\langle ((\widehat{\mathcal{P}}_j^{(1)[1]} - (1 + b_j^{(1)[1]})\mathcal{P}_j)\mathbf{u}_j^{(1)}, \mathbf{v}\right\rangle, \mathbf{v} \in \mathbb{H}. \tag{6.32}$$

$$\rho_j(\mathbf{v})^{[2]} := \left\langle ((\widehat{\mathcal{P}}_j^{[2]} - (1 + b_j^{(1)[2]})\mathcal{P}_j)\mathbf{u}_j^{(1)}, \mathbf{v}\right\rangle, \mathbf{v} \in \mathbb{H}. \tag{6.33}$$

Equation (6.6) in Koltchinskii and Lounici 2014 provides the representation of linear forms of

$\check{\mathbf{u}}_j^{(1),[1]}$ in terms of $\rho_j(\mathbf{v})^{[1]}$ and $\rho_j(\mathbf{v})^{[2]}$:

$$\begin{aligned}
&\sqrt{n}\left\langle \frac{1}{2}\left[\check{\mathbf{u}}_j^{(1),[1]} + \check{\mathbf{u}}_j^{(1),[2]} - \left(\sqrt{1+b_j^{(1)[1]}} + \sqrt{1+b_j^{(1)[2]}}\right)\mathbf{u}_j^{(1)}\right], \mathbf{v}\right\rangle \\
&= \sum_{h=1}^{2} \sqrt{n}\Bigg[\frac{\rho_j(\mathbf{v})^{[h]}}{\sqrt{1+b_j^{(1)[h]}+\rho_j(\mathbf{v})^{[h]}}} \\
&\quad + \frac{\sqrt{1+b_j^{(1)[h]}}}{\sqrt{1+b_j^{(1)[h]}+\rho_j(\mathbf{v})^{[h]}}\left(\sqrt{1+b_j^{(1)[h]}+\rho_j(\mathbf{v})^{[h]}}+\sqrt{1+b_j^{(1)[h]}}\right)} \rho_j(u_j^{(1)})^{[h]}\langle \mathbf{u}_j^{(1)}, \mathbf{v}\rangle\Bigg]. \quad (6.34)
\end{aligned}$$

We shall make use of the following lemma:

Lemma 3. *Under the same assumptions in Theorem 6, for any* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{d_q}$,

$$\begin{aligned}
&\sqrt{\frac{n}{2}}\left\langle \left[\widehat{\mathcal{P}}_j^{(q)[1]} - (1+b_j^{(q)[1]})\mathcal{P}_j^{(q)}\right]\mathbf{u}, \mathbf{v}\right\rangle \\
&\quad - \sqrt{\frac{2}{n}}\sum_{k=1}^{n/2}\left[\langle y_{jk}^{(q)}, \mathcal{P}_j^{(q)}\mathbf{v}\rangle\langle y_{jk}^{(q)}, C_j^{(q)}\mathbf{u}\rangle + \langle y_{jk}^{(q)}, \mathcal{P}_j^{(q)}\mathbf{u}\rangle\langle y_{jk}^{(q)}, C_j^{(q)}\mathbf{v}\rangle\right] = o_p\left(\|\mathbf{u}\|\|\mathbf{v}\|\right).
\end{aligned}$$

Further more, there exists universal constant C such that $\mathbb{P}(|b_j^{(q)[1]}| \le Cd/n) \to 1$ *as* $n \to \infty$.

Following from Lemma 3, observe that $C_j\mathbf{u}_j^{(1)} = 0$, we have

$$\sqrt{\frac{n}{2}}\rho_j(\mathbf{v})^{[1]} - \sqrt{\frac{2}{n}}\sum_{k=1}^{\frac{n}{2}}\eta_{jk}(\mathbf{v}) \xrightarrow{p} 0, \quad (6.35)$$

$$\sqrt{\frac{n}{2}}\rho_j(\mathbf{v})^{[2]} - \sqrt{\frac{2}{n}}\sum_{k=\frac{n}{2}+1}^{n}\eta_{jk}(\mathbf{v}) \xrightarrow{p} 0. \quad (6.36)$$

If we can show that for any $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \in \mathbb{R}^{d_1}$,

$$\frac{1}{\sqrt{n}}\left(\sum_{k=1}^{n}\eta_{j_1k}(\mathbf{v}_1), \cdots \sum_{k=1}^{n}\eta_{j_mk}(\mathbf{v}_m)\right) \xrightarrow{d} N\left(0, \mathbf{\Gamma}\right), \quad (6.37)$$

where

$$\mathbf{\Gamma}_{i_1 i_2} = \begin{cases} \left(\frac{1}{\sigma^2_{j_{i_1}}} + \frac{1}{\sigma^4_{j_{i_1}}}\right) \left\langle \mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{j_{i_1}}} \mathbf{v}_{i_1}, \mathcal{P}^{\perp}_{\mathbf{u}^{(q)}_{j_{i_1}}} \mathbf{v}_{i_2} \right\rangle, & \text{if } j_{i_1} = j_{i_2}, \\ 0, & \text{if } j_{i_1} \neq j_{i_2}, \end{cases}$$

then combining with the facts that $b^{(1)[h]}_j \xrightarrow{p} 0$ and (6.34), (6.35) and (6.36), we have:

$$\begin{aligned} &\sqrt{n}\Bigg(\left\langle \frac{1}{2}\left[\check{\mathbf{u}}^{(1),[1]}_{j_1} + \check{\mathbf{u}}^{(1),[2]}_{j_1} - \left(\sqrt{1 + b^{(1)[1]}_{j_1}} + \sqrt{1 + b^{(1)[2]}_{j_1}} \right) \mathbf{u}^{(1)}_{j_1} \right], \mathbf{v}_1 \right\rangle, \dots, \\ &\qquad \left\langle \frac{1}{2}\left[\check{\mathbf{u}}^{(1),[1]}_{j_m} + \check{\mathbf{u}}^{(1),[2]}_{j_m} - \left(\sqrt{1 + b^{(1)[1]}_{j_m}} + \sqrt{1 + b^{(1)[2]}_{j_m}} \right) \mathbf{u}^{(1)}_{j_m} \right], \mathbf{v}_m \right\rangle \Bigg) \\ =&\frac{1}{\sqrt{n}} \left(\sum_{k=1}^{n} \eta_{j_1 k}(\mathbf{v}_1), \cdots \sum_{k=1}^{n} \eta_{j_m k}(\mathbf{v}_m) \right) + o_p(1) \\ \xrightarrow{d}& N(0, \mathbf{\Gamma}) . \end{aligned}$$

Recall that $b^{(1)[h]}_j \xrightarrow{p} 0$, by Slutsky's Theorem,

$$\begin{aligned} &\sqrt{n}\left((1 + b^{(1)}_{j_1}) \langle \check{\mathbf{u}}^{(1)}_{j_1}, \mathbf{v}_1 \rangle - \langle \mathbf{u}^{(1)}_{j_1}, \mathbf{v}_1 \rangle, \dots, (1 + b^{(1)}_{j_m}) \langle \check{\mathbf{u}}^{(1)}_{j_m}, \mathbf{v}_m \rangle - \langle \mathbf{u}^{(1)}_{j_m}, \mathbf{v}_m \rangle \right) \\ \xrightarrow{d}& N(0, \mathbf{\Gamma}) . \end{aligned}$$

The claim (6.27) then follows.

We shall now prove (6.37). Note that $\mathcal{P}_j \mathbf{u}^{(1)}_j = \mathbf{u}^{(1)}_j$, and $C_j = \frac{1}{\sigma^2_j} \mathcal{P}^{\perp}_j$, we have

$$\eta_{jk}(\mathbf{v}) = \frac{1}{\sigma^2_j} \left(\sigma_j \theta_{jk} + \mathscr{E}_k \times_1 \mathbf{u}^{(1)}_j \times_2 \mathbf{u}^{(2)}_j \cdots \times_p \mathbf{u}^{(p)}_j \right) \left(\mathscr{E}_k \times_1 (\mathcal{P}^{\perp}_j \mathbf{v}) \times_2 \mathbf{u}^{(2)}_j \cdots \times_p \mathbf{u}^{(p)}_j \right).$$

Both $\sigma_j \theta_{jk} + \mathscr{E}_k \times_1 \mathbf{u}^{(1)}_j \times_2 \mathbf{u}^{(2)}_j \cdots \times_p \mathbf{u}^{(p)}_j$ and $\mathscr{E}_k \times_1 (\mathbb{I}_{d_1} \mathcal{P}^{\perp}_j \mathbf{v}) \times_2 \mathbf{u}^{(2)}_j \cdots \times_p \mathbf{u}^{(p)}_j$ are centered Gaussian random variables. Moreover, they are independent since θ_{jk} is independent with $\mathscr{E}_k$, and

$\mathbf{u}_j^{(1)} \perp \mathcal{P}_j^{\perp}\mathbf{v}$. So

$$\mathbb{E}\eta_{jk}(\mathbf{v}) = 0.$$

More generally, Gaussian variable $\sigma_j\theta_{jk} + \mathscr{E}_k \times_1 \mathbf{u}_j^{(1)} \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}$ and Gaussian vector $\mathscr{E}_k \times_1 (\mathcal{P}_j^{\perp}) \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}$ are independent for the exact same reason. So direct calculation gives:

$$\begin{aligned}
&\operatorname{cov}(\eta_{jk}(\mathbf{v}_1), \eta_{jk}(\mathbf{v}_2)) \\
=&\mathbb{E}\Bigg[\frac{1}{\sigma_j^4}\left(\sigma_j\theta_{jk} + \mathscr{E}_k \times_1 \mathbf{u}_j^{(1)} \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)^2 \\
&\left(\mathscr{E}_k \times_1 (\mathbb{I}_{d_1}\mathcal{P}_j^{\perp}\mathbf{v}_1) \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)\left(\mathscr{E}_k \times_1 (\mathbb{I}_{d_1}\mathcal{P}_j^{\perp}\mathbf{v}_2) \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)\Bigg] \\
=&\frac{1}{\sigma_j^4}\mathbb{E}\left(\sigma_j\theta_{jk} + \mathscr{E}_k \times_1 \mathbf{u}_j^{(1)} \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)^2 \\
&\mathbb{E}\Bigg[\left(\mathscr{E}_k \times_1 (\mathbb{I}_{d_1}\mathcal{P}_j^{\perp}\mathbf{v}_1) \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)\left(\mathscr{E}_k \times_1 (\mathbb{I}_{d_1}\mathcal{P}_j^{\perp}\mathbf{v}_2) \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)\Bigg] \\
=&\frac{1+\sigma_j^2}{\sigma_j^4}\langle \mathbb{I}_{d_1}\mathcal{P}_j^{\perp}\mathbf{v}_1, \mathbb{I}_{d_1}\mathcal{P}_j^{\perp}\mathbf{v}_2\rangle \\
\to&\frac{1+\sigma_j^2}{\sigma_j^4}\langle \mathcal{P}_j^{\perp}\mathbf{v}_1, \mathcal{P}_j^{\perp}\mathbf{v}_2\rangle.
\end{aligned}$$

On the other hand, for $j_1 \neq j_2$, $\{\mathbf{u}_j^{(q)}\}$ are orthogonal with each other, so $\operatorname{cov}(\eta_{j_1k}(\mathbf{v}_1), \eta_{j_2k}(\mathbf{v}_2)) = 0$ for $\forall j_1 \neq j_2 \in [r]$.

With the fact that $\eta_{jk}(\mathbf{v})$ are i.i.d for $k = 1, 2, \ldots, n$, to finish this part of the proof with CLT, it remains to check the Lindeberg condition for CLT, which reduced to

$$\frac{\mathbb{E}\eta_{jk}(\mathbf{v})^2\mathbb{I}\left(|\eta_{jk}(\mathbf{v})| \geq \tau\sqrt{n}\mathbb{E}^{1/2}\eta_{jk}(\mathbf{v})^2\right)}{\mathbb{E}\eta_{jk}(\mathbf{v})^2} \to 0 \text{ as } n \to \infty$$

for all $\tau > 0$. Note that

$$\frac{\mathbb{E}\eta_{jk}(\mathbf{v})^2 \mathbb{I}\left(|\eta_{jk}(\mathbf{v})| \geq \tau\sqrt{n}\mathbb{E}^{1/2}\eta_{jk}(\mathbf{v})^2\right)}{\mathbb{E}\eta_{jk}(\mathbf{v})^2} \leq \frac{\mathbb{E}\eta_{jk}(\mathbf{v})^4}{\tau^2 n \left(\mathbb{E}\eta_{jk}(\mathbf{v})^2\right)^2}.$$

Since

$$\begin{aligned}
&\mathbb{E}\eta_{jk}(\mathbf{v})^4 \\
=&\mathbb{E}\left(\sigma_j\theta_{jk} + \mathscr{E}_k \times_1 \mathbf{u}_j^{(1)} \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)^4 \mathbb{E}\left(\mathscr{E}_k \times_1 (\mathcal{P}_j^{\perp}\mathbf{v}) \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)^4,
\end{aligned}$$

and

$$\begin{aligned}
&\left(\mathbb{E}\eta_{jk}(\mathbf{v})^2\right)^2 \\
=&\mathbb{E}^2\left(\sigma_j\theta_{jk} + \mathscr{E}_k \times_1 \mathbf{u}_j^{(1)} \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)^2 \mathbb{E}^2\left(\mathscr{E}_k \times_1 (\mathcal{P}_j^{\perp}\mathbf{v}) \times_2 \mathbf{u}_j^{(2)} \cdots \times_p \mathbf{u}_j^{(p)}\right)^2,
\end{aligned}$$

with the fact that for a centered normal random variable ξ, $\mathbb{E}\xi^4 = 3\mathbb{E}^2\xi^2$, we get

$$\frac{\mathbb{E}\eta_{jk}(\mathbf{v})^4}{\tau^2 n \left(\mathbb{E}\eta_{jk}(\mathbf{v})^2\right)^2} = \frac{1}{\tau^2 n} \to 0,$$

and (6.37) follows.

6.4.2 Step 2.

Again, it suffices to consider the case $q = 1$. We shall now argue that

$$\langle \check{\mathbf{u}}_j^{(1)}, \mathbf{u}_j^{(1)} \rangle = \frac{1}{\sqrt{1 + \frac{d_1}{n}\left(\frac{\sigma_0^2}{\sigma_j^2} + \frac{\sigma_0^4}{\sigma_j^4}\right)}} + O_p\left(\frac{d^{3/2}}{n^{3/2}}\right). \tag{6.38}$$

Write $\tilde{\Sigma}^{[1]} = \widehat{\Sigma}^{[1]} - \mathscr{I}$, $\tilde{\Sigma}^{[2]} = \widehat{\Sigma}^{[2]} - \mathscr{I}$, and let

$$\grave{\mathbf{u}}_j^{(1),[1]} = \tilde{\Sigma}^{[1]}(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}, \breve{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}),$$

and

$$\grave{\mathbf{u}}_j^{(1),[2]} = \tilde{\Sigma}^{[2]}(\cdot, \widehat{\mathbf{u}}_j^{(2)[1]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[1]}, \breve{\mathbf{u}}_j^{(1),[2]}, \widehat{\mathbf{u}}_j^{(2)[1]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[1]}).$$

Since $\breve{\mathbf{u}}_j^{(1),[1]}$ is the leading eigenvector of

$$\widehat{\Sigma}^{[1]}(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}, \cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}),$$

it is also the leading eigenvector of

$$\tilde{\Sigma}^{[1]}(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}, \cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}),$$

so

$$\breve{\mathbf{u}}_j^{(1),[1]} = \frac{\grave{\mathbf{u}}_j^{(1),[1]}}{\|\grave{\mathbf{u}}_j^{(1),[1]}\|},$$

and similarly

$$\breve{\mathbf{u}}_j^{(1),[2]} = \frac{\grave{\mathbf{u}}_j^{(1),[2]}}{\|\grave{\mathbf{u}}_j^{(1),[2]}\|}.$$

Write

$$\mathbf{z}_j^{(1)[1]} := \grave{\mathbf{u}}_j^{(1),[1]} - \langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \rangle \mathbf{u}_j^{(1)} - \mathbf{y}_j^{(1)[1]},$$

where

$$\mathbf{y}_j^{(1)[1]} := \mathcal{P}_{\mathbf{u}_j^{(1)}}^{\perp} \left[\frac{2}{n} \sum_{k=1}^{n/2} \mathscr{E}_k(\cdot, \mathbf{u}_j^{(2)}, \ldots, \mathbf{u}_j^{(p)}) \sigma_j \theta_{jk} + \frac{2}{n} \sum_{k=1}^{n/2} \mathscr{E}_k(\cdot, \mathbf{u}_j^{(2)}, \ldots, \mathbf{u}_j^{(p)}) \mathscr{E}_k(\mathbf{u}_j^{(1)}, \ldots, \mathbf{u}_j^{(p)}) \right]. \tag{6.39}$$

$\mathbf{z}_j^{(1)[2]}$ and $\mathbf{y}_j^{(1)[2]}$ are defined similarly so that

$$\grave{\mathbf{u}}_j^{(1),[1]} = \langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \rangle \mathbf{u}_j^{(1)} + \mathbf{y}_j^{(1)[1]} + \mathbf{z}_j^{(1)[1]},$$

$$\grave{\mathbf{u}}_j^{(1),[2]} = \langle \grave{\mathbf{u}}_j^{(1),[2]}, \mathbf{u}_j^{(1)} \rangle \mathbf{u}_j^{(1)} + \mathbf{y}_j^{(1)[2]} + \mathbf{z}_j^{(1)[2]}.$$

We shall treat the three terms in $\grave{\mathbf{u}}_j^{(1),[1]}$ and $\grave{\mathbf{u}}_j^{(1),[1]}$ separately to show that

$$\langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \rangle = \sigma_j^2 + O_p\left(\frac{d}{n}\right), \ \langle \grave{\mathbf{u}}_j^{(1),[2]}, \mathbf{u}_j^{(1)} \rangle = \sigma_j^2 + O_p\left(\frac{d}{n}\right), \tag{6.40}$$

$$\left\| \mathbf{y}_j^{(1)[1]} + \mathbf{y}_j^{(1)[2]} \right\| = 2\sqrt{\frac{d_1}{n}}\sqrt{\sigma_0^2\sigma_j^2 + \sigma_0^4} + O_p\left(\frac{\sqrt{d}}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right), \tag{6.41}$$

$$\begin{aligned} \left\| \mathbf{y}_j^{(1)[1]} \right\| &= \sqrt{\frac{2d_1}{n}}\sqrt{\sigma_0^2\sigma_j^2 + \sigma_0^4} + O_p\left(\frac{\sqrt{d}}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right), \\ \left\| \mathbf{y}_j^{(1)[2]} \right\| &= \sqrt{\frac{2d_1}{n}}\sqrt{\sigma_0^2\sigma_j^2 + \sigma_0^4} + O_p\left(\frac{\sqrt{d}}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right), \end{aligned} \tag{6.42}$$

and

$$\mathbf{z}_j^{(1)[1]} = O_p\left(\frac{d}{n}\right), \ \mathbf{z}_j^{(1)[2]} = O_p\left(\frac{d}{n}\right). \tag{6.43}$$

We first show that equation (6.38) follows from the bounds given in (6.40)-(6.43). We start

with $\|\dot{\mathbf{u}}_j^{(1),[1]}\|$. Combining (6.40)-(6.43), we have

$$\begin{aligned}
&\|\dot{\mathbf{u}}_j^{(1),[1]}\| \\
=&\sqrt{\left\langle \dot{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \right\rangle^2 + \left\| \mathbf{y}_j^{(1)[1]} + \mathbf{z}_j^{(1)[1]} \right\|^2} \\
=&\sqrt{\left[\sigma_j^2 + O_p\left(\frac{d}{n}\right)\right]^2 + \left[O_p\left(\sqrt{\frac{d}{n}}\right) + O_p\left(\frac{d}{n}\right)\right]^2} \\
=&\sigma_j^2 + O_p\left(\frac{d}{n}\right),
\end{aligned} \tag{6.44}$$

where the second equality follows from (6.40), (6.42), (6.43), and the last equality holds because $d = o(n)$. Similarly we can derive that

$$\|\dot{\mathbf{u}}_j^{(1),[2]}\| = \sigma_j^2 + O_p\left(d/n\right).$$

Hence,

$$\begin{aligned}
&\left\| \frac{\mathbf{y}_j^{(1)[1]}}{\|\dot{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\mathbf{y}_j^{(1)[2]}}{\|\dot{\mathbf{u}}_j^{(1),[2]}\|} \right\| \\
=&\left\| \frac{1}{\sigma_j^2}(\mathbf{y}_j^{(1)[1]} + \mathbf{y}_j^{(1)[2]}) + \left(\frac{1}{\|\dot{\mathbf{u}}_j^{(1),[1]}\|} - \frac{1}{\sigma_j^2}\right)\mathbf{y}_j^{(1)[1]} + \left(\frac{1}{\|\dot{\mathbf{u}}_j^{(1),[2]}\|} - \frac{1}{\sigma_j^2}\right)\mathbf{y}_j^{(1)[2]} \right\| \\
=&2\sqrt{\frac{d_1}{n}\left(\frac{\sigma_0^2}{\sigma_j^2} + \frac{\sigma_0^4}{\sigma_j^4}\right)} + O_p\left(\frac{\sqrt{d}}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{\sqrt{d}}{n}\right) \cdot O_p\left(\frac{1}{\sqrt{n}}\right) \\
=&2\sqrt{\frac{d_1}{n}\left(\frac{\sigma_0^2}{\sigma_j^2} + \frac{\sigma_0^4}{\sigma_j^4}\right)} + O_p\left(\frac{\sqrt{d}}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right),
\end{aligned} \tag{6.45}$$

where the second equality follows from (6.41), (6.42) and (6.44).

Moreover, (6.43), (6.44) imply that

$$\left\| \frac{\mathbf{z}_j^{(1)[1]}}{\|\dot{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\mathbf{z}_j^{(1)[2]}}{\|\dot{\mathbf{u}}_j^{(1),[2]}\|} \right\| = O_p\left(\frac{d}{n}\right), \tag{6.46}$$

and (6.40) and (6.44) imply that

$$\frac{\langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \rangle}{\|\grave{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\langle \grave{\mathbf{u}}_j^{(1),[2]}, \mathbf{u}_j^{(1)} \rangle}{\|\grave{\mathbf{u}}_j^{(1),[2]}\|} = 2 + O_p\left(\frac{d}{n}\right). \tag{6.47}$$

Therefore,

$$\begin{aligned}
&\langle \breve{\mathbf{u}}_j^{(1)}, \mathbf{u}_j^{(1)} \rangle \\
=&\frac{\frac{\langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \rangle}{\|\grave{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\langle \grave{\mathbf{u}}_j^{(1),[2]}, \mathbf{u}_j^{(1)} \rangle}{\|\grave{\mathbf{u}}_j^{(1),[2]}\|}}{\sqrt{\left(\frac{\langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \rangle}{\|\grave{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\langle \grave{\mathbf{u}}_j^{(1),[2]}, \mathbf{u}_j^{(1)} \rangle}{\|\grave{\mathbf{u}}_j^{(1),[2]}\|}\right)^2 + \left(\frac{\mathbf{y}_j^{(1)[1]}}{\|\grave{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\mathbf{y}_j^{(1)[2]}}{\|\grave{\mathbf{u}}_j^{(1),[2]}\|} + \frac{\mathbf{z}_j^{(1)[1]}}{\|\grave{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\mathbf{z}_j^{(1)[2]}}{\|\grave{\mathbf{u}}_j^{(1),[2]}\|}\right)^2}} \\
=&1\Bigg/\sqrt{1 + \left(\frac{\frac{\mathbf{y}_j^{(1)[1]}}{\|\grave{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\mathbf{y}_j^{(1)[2]}}{\|\grave{\mathbf{u}}_j^{(1),[2]}\|} + \frac{\mathbf{z}_j^{(1)[1]}}{\|\grave{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\mathbf{z}_j^{(1)[2]}}{\|\grave{\mathbf{u}}_j^{(1),[2]}\|}}{\frac{\langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \rangle}{\|\grave{\mathbf{u}}_j^{(1),[1]}\|} + \frac{\langle \grave{\mathbf{u}}_j^{(1),[2]}, \mathbf{u}_j^{(1)} \rangle}{\|\grave{\mathbf{u}}_j^{(1),[2]}\|}}\right)^2} \\
=&1\Bigg/\sqrt{1 + \left(\frac{2\sqrt{\frac{d_1}{n}\left(\frac{\sigma_0^2}{\sigma_j^2} + \frac{\sigma_0^4}{\sigma_j^4}\right)} + O_p\left(\frac{\sqrt{d}}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{d}{n}\right)}{2 + O_p\left(\frac{d}{n}\right)}\right)^2} \\
=&1\Bigg/\sqrt{1 + \left\{\left[\sqrt{\frac{d_1}{n}\left(\frac{\sigma_0^2}{\sigma_j^2} + \frac{\sigma_0^4}{\sigma_j^4}\right)} + O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{d}{n}\right)\right] \cdot \left[1 + O_p\left(\frac{d}{n}\right)\right]\right\}^2} \\
=&\frac{1}{\sqrt{1 + \left[\sqrt{\frac{d_1}{n}\left(\frac{\sigma_0^2}{\sigma_j^2} + \frac{\sigma_0^4}{\sigma_j^4}\right)} + O_p\left(\frac{d}{n}\right)\right]^2}} \\
=&\frac{1}{\sqrt{1 + \frac{d_1}{n}\left(\frac{\sigma_0^2}{\sigma_j^2} + \frac{\sigma_0^4}{\sigma_j^4}\right)}} + O_p\left(\frac{d^{3/2}}{n^{3/2}}\right),
\end{aligned}$$

where the third equation follows from (6.45)-(6.47), and the last equation follows from the simple

fact that

$$\left|\frac{1}{\sqrt{1+(x+y)^2}}-\frac{1}{\sqrt{1+x^2}}\right|\leq\left|xy+\frac{1}{2}y^2\right|.$$

It now remains to show (6.40)-(6.43).

Equations (6.41) **and** (6.42)**.**

Write

$$\mathbf{x}_k:=\mathscr{E}_k(\cdot,\mathbf{u}_j^{(2)},\ldots,\mathbf{u}_j^{(p)})\in\mathbb{R}^{d_1}.$$

If we choose an orthogonal basis of $\mathbb{R}^{d_1}$ with $\mathbf{e}_1=\mathbf{u}_j^{(1)}$, then

$$\left\|\frac{1}{2}(\mathbf{y}_j^{(1)[1]}+\mathbf{y}_j^{(1)[2]})\right\|=\left\|\mathcal{P}_{\mathbf{u}_j^{(1)}}^{\perp}\left[\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\left(\sigma_j\theta_{jk}+\mathbf{x}_{1,k}\right)\right]\right\|=\left\|\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_{(-1),k}\left(\sigma_j\theta_{jk}+\mathbf{x}_{1,k}\right)\right\|,$$

in which $\mathbf{x}_{1,k}$ means the first entry of $\mathbf{x}_k$, while $\mathbf{x}_{(-1),k}$ stands for all the other entries (having a dimension of d_1-1).

Observe that $\mathbf{x}_{(-1),k}$ and $\left(\sigma_j\theta_{jk}+\mathbf{x}_{1,k}\right)$ two independent group of i.i.d. random variables, $\mathbf{x}_{(-1),k}$ follows distribution $N(0,\sigma_0^2 I_{d_1-1})$, and $\left(\sigma_j\theta_{jk}+\mathbf{x}_{1,k}\right)$ follows distribution $N(0,\sigma_0^2+\sigma_j^2)$. Thus,

$$\left\|\frac{1}{2}(\mathbf{y}_j^{(1)[1]}+\mathbf{y}_j^{(1)[2]})\right\|=\sqrt{\frac{d_1-1}{n}}\sqrt{\sigma_0^2\sigma_j^2+\sigma_0^4}+O_p\left(\frac{\sqrt{d_1-1}}{n}\right)+O_p\left(\frac{1}{\sqrt{n}}\right).$$

(6.42) can be proven in a similar way.

Equation (6.43)**.**

Recall that

$$
\begin{aligned}
&\tilde{\Sigma}(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}, \grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \\
&= \sum_{l=1}^{r} \sigma_l^2 \left(\frac{2}{n} \sum_{k=1}^{n/2} \theta_{lk}^2 \right) \left(\prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle \right)^2 \langle \mathbf{u}_l^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]} \rangle \mathbf{u}_l^{(1)} \\
&\quad + \sum_{l_1 \neq l_2} \sigma_{l_1} \sigma_{l_2} \left(\frac{2}{n} \sum_{k=1}^{n/2} \theta_{l_1 k} \theta_{l_2 k} \right) \left(\prod_{q=2}^{p} \langle \mathbf{u}_{l_1}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle \right) \left(\prod_{q=2}^{p} \langle \mathbf{u}_{l_2}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle \right) \langle \mathbf{u}_{l_2}^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]} \rangle \mathbf{u}_{l_1}^{(1)} \\
&\quad + \frac{2}{n} \sum_{k=1}^{n/2} \sum_{l=1}^{r} \sigma_l \theta_{lk} \left(\prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle \right) \mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \mathbf{u}_l^{(1)} \\
&\quad + \frac{2}{n} \sum_{k=1}^{n/2} \sum_{l=1}^{r} \mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \sigma_l \theta_{lk} \left(\prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle \right) \langle \mathbf{u}_l^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]} \rangle \\
&\quad + \frac{2}{n} \sum_{k=1}^{n/2} \mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) - \grave{\mathbf{u}}_j^{(1),[1]},
\end{aligned}
$$

so by definition,

$$
\begin{aligned}
\mathbf{z}_j^{(1)[1]} = &\sum_{l\neq j, l\in[r]} \sigma_l^2 \left(\frac{2}{n}\sum_{k=1}^{n/2}\theta_{lk}^2\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)^2 \langle \mathbf{u}_l^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle \mathbf{u}_l^{(1)} \\
&+ \sum_{l_1\neq l_2, l_1\neq j} \sigma_{l_1}\sigma_{l_2}\left(\frac{2}{n}\sum_{k=1}^{n/2}\theta_{l_1k}\theta_{l_2k}\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_{l_1}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_{l_2}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_{l_2}^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle \mathbf{u}_{l_1}^{(1)} \\
&+ \frac{2}{n}\sum_{k=1}^{n/2}\sum_{l\neq j, l\in[r]} \sigma_l\theta_{lk}\left(\prod_{q=2}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\mathbf{u}_l^{(1)} \\
&+ \frac{2}{n}\sum_{k=1}^{n/2}\sum_{l\neq j, l\in[r]} \mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\sigma_l\theta_{lk}\left(\prod_{q=2}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_l^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle \\
&+ \mathcal{P}_{\mathbf{u}_j^{(1)}}^{\perp}\left[\frac{2}{n}\sum_{k=1}^{n/2}\mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\sigma_j\theta_{jk}\left(\prod_{q=2}^{p}\langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_j^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle\right. \\
&\qquad \left. - \frac{2}{n}\sum_{k=1}^{n/2}\mathscr{E}_k(\cdot, \mathbf{u}_j^{(2)}, \ldots, \mathbf{u}_j^{(p)})\sigma_j\theta_{jk}\right] \\
&+ \mathcal{P}_{\mathbf{u}_j^{(1)}}^{\perp}\left[\frac{2}{n}\sum_{k=1}^{n/2}\mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) - \grave{\mathbf{u}}_j^{(1),[1]}\right. \\
&\qquad \left. - \frac{2}{n}\sum_{k=1}^{n/2}\mathscr{E}_k(\cdot, \mathbf{u}_j^{(2)}, \ldots, \mathbf{u}_j^{(p)})\mathscr{E}_k(\mathbf{u}_j^{(1)}, \ldots, \mathbf{u}_j^{(p)}) + \mathbf{u}_j^{(1)}\right], \qquad (6.48)
\end{aligned}
$$

and we will deal with those six terms one by one.

The first term.

$$
\mathbf{z}_1^{[1]} := \sum_{l\neq j, l\in[r]} \sigma_l^2 \left(\frac{2}{n}\sum_{k=1}^{n/2}\theta_{lk}^2\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)^2 \langle \mathbf{u}_l^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle \mathbf{u}_l^{(1)}.
$$

Observe that

$$
\left\|\mathbf{z}_1^{[1]}\right\| \leq \left(\delta^{(1)[2]}\right)^{2p-1} \sum_{l\neq j, l\in[r]} \sigma_l^2 \left|\frac{2}{n}\sum_{k=1}^{n/2}\theta_{lk}^2\right| = O_p\left[\left(\sqrt{\frac{d}{n}}\right)^{2p-1}\right] = O_p\left[\left(\frac{d}{n}\right)^{3/2}\right]. \qquad (6.49)
$$

The last two equality's follow from (6.31), and $p \geq 2$. Because of the assumption $d = o(n)$, it implies

$$\left\|\mathbf{z}_1^{[1]}\right\| \leq O_p\left(\frac{d}{n}\right).$$

The second term.

$$\begin{aligned}
\mathbf{z}_2^{[1]} :=& \sum_{l_1\neq l_2, l_1\neq j} \sigma_{l_1}\sigma_{l_2}\left(\frac{2}{n}\sum_{k=1}^{n/2}\theta_{l_1k}\theta_{l_2k}\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_{l_1}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_{l_2}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_{l_2}^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle \mathbf{u}_{l_1}^{(1)} \\
=& \sum_{l_1\neq l_2, l_1\neq j, l_2\neq j} \sigma_{l_1}\sigma_{l_2}\left(\frac{2}{n}\sum_{k=1}^{n/2}\theta_{l_1k}\theta_{l_2k}\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_{l_1}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_{l_2}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_{l_2}^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle \mathbf{u}_{l_1}^{(1)} \\
&+ \sum_{l_1\neq j} \sigma_{l_1}\sigma_{j}\left(\frac{2}{n}\sum_{k=1}^{n/2}\theta_{l_1k}\theta_{jk}\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_{l_1}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\left(\prod_{q=2}^{p}\langle \mathbf{u}_{j}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_{j}^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle \mathbf{u}_{l_1}^{(1)}.
\end{aligned}$$

Because

$$\left|\frac{2}{n}\sum_{k=1}^{n/2}\theta_{l_1k}\theta_{l_2k}\right| = O_p(\frac{1}{\sqrt{n}}),\ \forall l_1 \neq l_2,$$

we immediately have

$$\begin{aligned}
&\left\|\mathbf{z}_2^{[1]}\right\| \\
\leq& \left(\delta^{(1)[2]}\right)^{2p-1} \sum_{l_1\neq l_2, l_1\neq j, l_2\neq j} \sigma_{l_1}\sigma_{l_2}\left|\frac{2}{n}\sum_{k=1}^{n/2}\theta_{l_1k}\theta_{l_2k}\right| + \left(\delta^{(1)[2]}\right)^{p-1} \sum_{l_1\neq j}\sigma_{l_1}\sigma_j\left|\frac{2}{n}\sum_{k=1}^{n/2}\theta_{l_1k}\theta_{jk}\right| \\
=& O_p(\frac{1}{\sqrt{n}})\cdot O_p\left[\left(\sqrt{\frac{d}{n}}\right)^{2p-1}\right] + O_p(\frac{1}{\sqrt{n}})\cdot\left[\left(\sqrt{\frac{d}{n}}\right)^{p-1}\right] \\
=& O_p\left(\frac{\sqrt{d}}{n}\right). \qquad (6.50)
\end{aligned}$$

The third term.

$$\mathbf{z}_3^{[1]} := \frac{2}{n}\sum_{k=1}^{n/2}\sum_{l\neq j, l\in[r]}\sigma_l\theta_{lk}\left(\prod_{q=2}^{p}\langle \mathbf{u}_{l}^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\mathbf{u}_l^{(1)}.$$

Observe that

$$\begin{aligned}\left\|\mathbf{z}_3^{[1]}\right\| \leq & \sum_{l\neq j, l\in[r]} \sigma_l \left|\prod_{q=2}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right| \cdot \left|\frac{2}{n}\sum_{k=1}^{n/2}\theta_{lk}\mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\right| \\ \leq & r\sigma_1(\delta^{(1)[2]})^{p-1}\cdot \|\widehat{\Sigma}_{\theta,\mathscr{E}}^{[1]}\| \\ = & O_p\left(\sqrt{\frac{d}{n}}\right)^{p-1}\cdot O_p\left(\sqrt{\frac{d}{n}}\right) \\ = & O_p\left(\frac{d}{n}\right), \end{aligned} \tag{6.51}$$

in which the second to last equality follows from (6.31) and (6.30).

The fourth term.

$$\mathbf{z}_4^{[1]} := \frac{2}{n}\sum_{k=1}^{n/2}\sum_{l\neq j, l\in[r]} \mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\sigma_l\theta_{lk}\left(\prod_{q=2}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_l^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle.$$

Similar to $\mathbf{z}_3^{[1]}$,

$$\left\|\mathbf{z}_4^{[1]}\right\| \leq \|\widehat{\Sigma}_{\theta,\mathscr{E}}\| \cdot \left(\delta^{(1)[2]}\right)^p = O_p\left(\frac{d^{3/2}}{n^{3/2}}\right) = O_p\left(\frac{d}{n}\right).$$

The fifth term.

$$\begin{aligned}\mathbf{z}_5^{[1]} := & \mathcal{P}_{\mathbf{u}_j^{(1)}}^{\perp}\left[\frac{2}{n}\sum_{k=1}^{n/2}\mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\sigma_j\theta_{jk}\left(\prod_{q=2}^{p}\langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_j^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle\right. \\ & \left. - \frac{2}{n}\sum_{k=1}^{n/2}\mathscr{E}_k(\cdot, \mathbf{u}_j^{(2)}, \ldots, \mathbf{u}_j^{(p)})\sigma_j\theta_{jk}\right]. \end{aligned}$$

we shall introduce the notation $\biguplus$, as the operation of expanding all the

$$\widehat{\mathbf{u}}_j^{(q)[2]} = \mathbf{u}_j^{(q)} + \Delta_j^{(q)[2]},$$

and

$$\grave{\mathbf{u}}_j^{(1),[1]} = \mathbf{u}_j^{(1)} + \grave{\Delta}_j^{(1)[1]},$$

and then keep all the terms with at least one $\Delta_j^{(\cdot)[2]}$ or $\grave{\Delta}_j^{(1)[1]}$ in it. For instance, here, expanding

$$\frac{2}{n}\sum_{k=1}^{n/2} \mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\sigma_j\theta_{jk}\left(\prod_{q=2}^{p}\langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_j^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle$$

would result in 2^{2p-1} terms, and we keep everything other than the term

$$\frac{2}{n}\sum_{k=1}^{n/2} \mathscr{E}_k(\cdot, \mathbf{u}_j^{(2)}, \ldots, \mathbf{u}_j^{(p)})\sigma_j\theta_{jk}\left(\prod_{q=1}^{p}\langle \mathbf{u}_j^{(q)}, \mathbf{u}_j^{(q)}\rangle\right).$$

Then we have:

$$\mathbf{z}_5^{[1]} = \mathcal{P}^{\perp}_{\mathbf{u}_j^{(1)}}\left\{\biguplus\left[\frac{2}{n}\sum_{k=1}^{n/2} \mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\sigma_j\theta_{jk}\left(\prod_{q=2}^{p}\langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_j^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle\right]\right\}.$$

Then,

$$\left\|\mathbf{z}_5^{[1]}\right\| \le \left\|\biguplus\left[\frac{2}{n}\sum_{k=1}^{n/2} \mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\sigma_j\theta_{jk}\left(\prod_{q=1}^{p}\langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\right]\right\|,$$

and for every term in $\biguplus$, its norm is bounded by

$$\|\widehat{\Sigma}_{\theta,\mathscr{E}}\| \cdot \delta^{(1)[2]} = O_p(\frac{d}{n}),$$

so that

$$\left\|\mathbf{z}_5^{[1]}\right\| = O_p\left(\frac{d}{n}\right).$$

The sixth term. This term can be treated in a similar fashion as the last term.

$$\begin{aligned}
\mathbf{z}_6^{[1]} :=& \mathcal{P}^{\perp}_{\mathbf{u}_j^{(1)}} \Bigg[\frac{2}{n} \sum_{k=1}^{n/2} \mathscr{E}_k(\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) - \grave{\mathbf{u}}_j^{(1),[1]} \\
& \qquad - \frac{2}{n} \sum_{k=1}^{n/2} \mathscr{E}_k(\cdot, \mathbf{u}_j^{(2)}, \ldots, \mathbf{u}_j^{(p)}) \mathscr{E}_k(\mathbf{u}_j^{(1)}, \ldots, \mathbf{u}_j^{(p)}) + \mathbf{u}_j^{(1)} \Bigg] \\
=& \mathcal{P}^{\perp}_{\mathbf{u}_j^{(1)}} \Bigg[\left(\frac{2}{n} \sum_{k=1}^{n/2} \mathscr{E}_k \otimes \mathscr{E}_k - \mathscr{I} \right) (\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}, \grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \\
& \qquad - \left(\frac{2}{n} \sum_{k=1}^{n/2} \mathscr{E}_k \otimes \mathscr{E}_k - \mathscr{I} \right) (\cdot, \mathbf{u}_j^{(2)}, \ldots, \mathbf{u}_j^{(p)}, \mathbf{u}_j^{(1)}, \ldots, \mathbf{u}_j^{(p)}) \Bigg] \\
=& \mathcal{P}^{\perp}_{\mathbf{u}_j^{(1)}} \left\{ \biguplus \left[\left(\frac{2}{n} \sum_{k=1}^{n/2} \mathscr{E}_k \otimes \mathscr{E}_k - \mathscr{I} \right) (\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}, \grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \right] \right\},
\end{aligned}$$

then,

$$\left\| \mathbf{z}_6^{[1]} \right\| \leq \left\| \biguplus \left[\left(\frac{2}{n} \sum_{k=1}^{n/2} \mathscr{E}_k \otimes \mathscr{E}_k - \mathscr{I} \right) (\cdot, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}, \widehat{\mathbf{u}}_j^{(1)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \right] \right\|,$$

and for every term in $\biguplus$, its norm is bounded by

$$\| \widehat{\Sigma}_{\mathscr{E}} - \mathscr{I} \| \cdot \delta^{(1)[2]} = O_p(\frac{d}{n}),$$

so that

$$\left\| \mathbf{z}_6^{[1]} \right\| = O_p\left(\frac{d}{n} \right).$$

Equation (6.40)

We now show that

$$\langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \rangle = \sigma_j^2 + O_p\left(\frac{d}{n} \right).$$

It follows immediately, by symmetry, that

$$\langle \grave{\mathbf{u}}_j^{(1),[2]}, \mathbf{u}_j^{(1)} \rangle = \sigma_j^2 + O_p\left(\frac{d}{n}\right).$$

By definition,

$$\begin{aligned}
&\langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)} \rangle \\
=&\sigma_j^2 \left(\frac{2}{n}\sum_{k=1}^{n/2} \theta_{jk}^2\right)\left(\prod_{q=2}^{p} \langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle\right)^2 \langle \mathbf{u}_j^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]} \rangle \\
&+ \sum_{l \neq j} \sigma_l \left(\frac{2}{n}\sum_{k=1}^{n/2} \theta_{jk}\theta_{lk}\right)\left(\prod_{q=2}^{p} \langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle\right)\left(\prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle\right) \langle \mathbf{u}_l^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]} \rangle \\
&+ \frac{2}{n}\sum_{k=1}^{n/2} \sigma_j \theta_{jk} \left(\prod_{q=2}^{p} \langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle\right) \mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \\
&+ \frac{2}{n}\sum_{k=1}^{n/2}\sum_{l=1}^{r} \mathscr{E}_k(\mathbf{u}_j^{(1)}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}) \sigma_l \theta_{lk} \left(\prod_{q=2}^{p} \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle\right) \langle \mathbf{u}_j^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]} \rangle \\
&+ \frac{2}{n}\sum_{k=1}^{n/2} (\mathscr{E}_k \otimes \mathscr{E}_k - \mathscr{I})(\mathbf{u}_j^{(1)}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}, \grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}). \qquad (6.52)
\end{aligned}$$

All the terms except the first one will be bounded with similar techniques as we bound the six terms in $\mathbf{z}_j^{(1)[1]}$, and we omit some of the details.

The first term. Observe that

$$1 - \langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle \leq \left(\delta^{(1)[2]}\right)^2, \; 1 - \langle \mathbf{u}_j^{(1)}, \grave{\mathbf{u}}_j^{(1)[1]} \rangle \leq \left(\delta^{(1)[2]}\right)^2.$$

Thus,

$$\sigma_j^2 \left(\frac{2}{n}\sum_{k=1}^{n/2} \theta_{jk}^2\right)\left(\prod_{q=2}^{p} \langle \mathbf{u}_j^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle\right)^2 \langle \mathbf{u}_j^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]} \rangle = \sigma_j^2 + O_p\left(\frac{d}{n}\right).$$

The second term.

$$\begin{aligned}
&\sum_{l\neq j}\sigma_l\left(\frac{2}{n}\sum_{k=1}^{n/2}\theta_{jk}\theta_{lk}\right)\left(\prod_{q=2}^{p}\langle\mathbf{u}_j^{(q)},\widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\left(\prod_{q=2}^{p}\langle\mathbf{u}_l^{(q)},\widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle\mathbf{u}_l^{(1)},\grave{\mathbf{u}}_j^{(1),[1]}\rangle\\
=&O_p(\frac{1}{\sqrt{n}})\cdot O_p\left[\left(\sqrt{\frac{d}{n}}\right)^{2p-1}\right]\\
=&o_p(\frac{1}{\sqrt{n}}).
\end{aligned}$$

The third term. Observe that

$$\begin{aligned}
&\frac{2}{n}\sum_{k=1}^{n/2}\sigma_j\theta_{jk}\left(\prod_{q=2}^{p}\langle\mathbf{u}_j^{(q)},\widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]},\widehat{\mathbf{u}}_j^{(2)[2]},\ldots,\widehat{\mathbf{u}}_j^{(p)[2]})\\
=&\frac{2}{n}\sum_{k=1}^{n/2}\sigma_j\theta_{jk}\left(\prod_{q=2}^{p}\langle\mathbf{u}_j^{(q)},\widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\mathscr{E}_k(\mathbf{u}_j^{(1),[1]},\widehat{\mathbf{u}}_j^{(2)[2]},\ldots,\widehat{\mathbf{u}}_j^{(p)[2]})\\
&+\frac{2}{n}\sum_{k=1}^{n/2}\sigma_j\theta_{jk}\left(\prod_{q=2}^{p}\langle\mathbf{u}_j^{(q)},\widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]}-\mathbf{u}_j^{(1),[1]},\widehat{\mathbf{u}}_j^{(2)[2]},\ldots,\widehat{\mathbf{u}}_j^{(p)[2]}),
\end{aligned}$$

in which the first term is $O_p(1/\sqrt{n})$ because $\theta_{jk}\mathscr{E}_k, k\in[n/2]$ are independent with $\widehat{\mathbf{u}}_j^{(2)[2]},\ldots,\widehat{\mathbf{u}}_j^{(p)[2]}$, and the second term is bounded by

$$\|\widehat{\Sigma}_{\theta,\mathscr{E}}\|\cdot\delta^{(1)[2]}=O_p(\frac{d}{n}),$$

so

$$\frac{2}{n}\sum_{k=1}^{n/2}\sigma_j\theta_{jk}\left(\prod_{q=2}^{p}\langle\mathbf{u}_j^{(q)},\widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\mathscr{E}_k(\grave{\mathbf{u}}_j^{(1),[1]},\widehat{\mathbf{u}}_j^{(2)[2]},\ldots,\widehat{\mathbf{u}}_j^{(p)[2]})=O_p\left(\frac{d}{n}\right)+O_p\left(\frac{1}{\sqrt{n}}\right)=O_p\left(\frac{d}{n}\right).$$

The last equality follows from the assumption $d=o(n)$.

The fourth term. Recall that $\widehat{\mathbf{u}}_j^{(q)[2]}$ is a function of the second half of the data, which is independent of the first half of the data, i.e., all the random variables with index $k\leq n/2$, so conditional

on the second half of the data, for any $l \in [r]$,

$$\mathscr{E}_k(\mathbf{u}_j^{(1)}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\theta_{lk}$$

are a product of two independent $N(0, \sigma_0^2)$ and $N(0,1)$ variables, and they are i.i.d across $k \le n/2$. So we have

$$\frac{2}{n}\sum_{k=1}^{n/2}\sum_{l=1}^{r}\mathscr{E}_k(\mathbf{u}_j^{(1)}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]})\sigma_l\theta_{lk}\left(\prod_{q=2}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)\langle \mathbf{u}_j^{(1)}, \grave{\mathbf{u}}_j^{(1),[1]}\rangle = O_p(\frac{1}{\sqrt{n}}).$$

The fifth term. Similar to the sixth term in $\mathbf{z}_j^{(1)[1]}$,

$$\frac{2}{n}\sum_{k=1}^{n/2}(\mathscr{E}_k \otimes \mathscr{E}_k - \mathscr{J})(\mathbf{u}_j^{(1)}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}, \grave{\mathbf{u}}_j^{(1),[1]}, \widehat{\mathbf{u}}_j^{(2)[2]}, \ldots, \widehat{\mathbf{u}}_j^{(p)[2]}).$$

Combining the five bounds above, we get

$$\langle \grave{\mathbf{u}}_j^{(1),[1]}, \mathbf{u}_j^{(1)}\rangle = \sigma_j^2 + O_p\left(\frac{d}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right) = O_p\left(\frac{d}{n}\right).$$

The last equality follows from the assumption $d = o(n)$.

6.4.3 Step 3.

Recall that

$$\widehat{b}_j^{(q)[1]} := \left\langle \widehat{\mathbf{u}}_j^{(q),[1][1]}, \widehat{\mathbf{u}}_j^{(q),[1][2]}\right\rangle - 1,$$

$$\widehat{b}_j^{(q)[2]} := \left\langle \widehat{\mathbf{u}}_j^{(q),[2][1]}, \widehat{\mathbf{u}}_j^{(q),[2][2]}\right\rangle - 1.$$

In this part, we use Lemma 8 to show that

$$\sqrt{n}\left(\widehat{b}_j^{(1)[1]} - b_j^{(1)[1]}\right) \xrightarrow{p} 0, \tag{6.53}$$

in which $b_j^{(1)[1]}$ is defined in Lemma 3. Similarly $\sqrt{n}\left(\widehat{b}_j^{(1)[2]} - b_j^{(1)[2]}\right) \xrightarrow{p} 0$. Then, recall that by definition (6.25),

$$b_j^{(1)} = \frac{\left\|\check{\mathbf{u}}_j^{(1),[1]} + \check{\mathbf{u}}_j^{(1),[2]}\right\|}{\sqrt{1 + b_j^{(1)[1]}} + \sqrt{1 + b_j^{(1)[2]}}} - 1,$$

and by definition (4.2),

$$\widehat{b}_j^{(1)} = \frac{\left\|\check{\mathbf{u}}_j^{(q),[1]} + \check{\mathbf{u}}_j^{(q),[2]}\right\|}{\sqrt{1 + \widehat{b}_j^{(1)[1]}} + \sqrt{1 + \widehat{b}_j^{(1)[2]}}} - 1,$$

we have that (6.29) follows from (6.53) and the fact that $\left\|\check{\mathbf{u}}_j^{(1),[1]} + \check{\mathbf{u}}_j^{(1),[2]}\right\| \leq 2$. Now we turn our attention to (6.53).

By Lemma 4,

$$\begin{aligned}
&\bar{\mathbb{E}}\|\widehat{M}_j^{[1]} - M_j^{[1]}\| \\
\leq& C\|M_j^{[1]}\|\left(\sqrt{\frac{r(M_j^{[1]})}{n}} \bigvee \frac{r(M_j^{[1]})}{n}\right) \\
\leq& \frac{C(\sigma_j^2 + 1)(\sigma_j^2 + d_1 + 1)}{(\frac{1}{2}\sigma_j^2 + 1)n},
\end{aligned}$$

on events $\mathcal{B}_n$. ($M_j^{[1]}$, $\widehat{M}_j^{[1]}$ and events $\mathcal{B}_n$ are defined at the beginning of the proof for Lemma 3.) The last inequality holds because of (6.91) and (6.87). Remember that $\bar{g}_1 \geq \frac{1}{2}\sigma_j^2$ (inequality (6.89)), so with assumption $d = o(n)$, we have that for large enough n, $\bar{\mathbb{E}}\|\widehat{M}_j^{[1]} - M_j^{[1]}\| \leq \frac{1}{4}\bar{g}_1$, i.e., if we let $\gamma = \frac{1}{4}$, then

$$\bar{\mathbb{E}}\|\widehat{M}_j^{[1]} - M_j^{[1]}\| \leq \frac{(1 - 2\gamma)\bar{g}_1}{2}.$$

$1 + b_j^{(1)[1]} \geq 2\gamma$ is also satisfied for large enough n by inequality (6.92).

Let $t = \left(\frac{n}{2r(M_j^{[1]})}\right)^{\frac{1}{3}}$. First note that on the event $\mathcal{B}_n$,

$$t \geq \left(\frac{n(\frac{1}{2}\sigma_j^2 + 1)}{2(\sigma_j^2 + d_1 + 1)}\right)^{\frac{1}{3}},$$

by inequality (6.91), with assumption $d = o(n)$, we have $t \geq 1$ for large enough n.

Let $D := D_{\frac{1}{4}}$ as in Lemma 8. We have

$$D\|M_j^{[1]}\|\left(\sqrt{\frac{t}{n}} \bigvee \sqrt{\frac{\log n}{n}}\right) \leq D(\sigma_j^2 + 1)\left(n^{-\frac{2}{3}} \bigvee \sqrt{\frac{\log n}{n}}\right) \leq \frac{1}{8}\bar{g}_1$$

for large enough n.

So all the conditions in Lemma 8 are satisfied with $\widehat{M}_j^{[1]}$ and $M_j^{[1]}$, conditional on $\{\mathcal{X}_k, k = n/2 + 1, \ldots, n\}$ under events $\mathcal{B}_n$, for large enough n.

Observe that conditional on $\{\mathcal{X}_k, k = n/2 + 1, \ldots, n\}$, $\widehat{b}_j^{(1)[1]}$ and $b_j^{(1)[1]}$ are just defined as the $\widehat{b}_1$ and b_1 [defined in (6.71) and (6.72)] corresponding to $M_j^{[1]}$. Now we can apply Lemma 8:

$$\begin{aligned}
&\sqrt{n}\left|\widehat{b}_j^{(1)[1]} - b_j^{(1)[1]}\right| \\
\leq& D\sqrt{n}\frac{\|M_j^{[1]}\|^2}{\bar{g}_1^2}\left(\sqrt{\frac{r(M_j^{[1]})}{n}} \bigvee \sqrt{\frac{t}{n}} \bigvee \sqrt{\frac{\log n}{n}}\right)\sqrt{\frac{t}{n}} \\
\leq& D\frac{\|M_j^{[1]}\|^2}{\bar{g}_1^2}\left(\left(\frac{r(M_j^{[1]})}{n}\right)^{\frac{1}{3}} \bigvee \frac{1}{r(M_j^{[1]})^{\frac{1}{3}}n^{\frac{1}{6}}} \bigvee \frac{\sqrt{\log n}}{r(M_j^{[1]})^{\frac{1}{6}}n^{\frac{1}{3}}}\right) \\
\leq& D\frac{\|M_j^{[1]}\|^2}{\bar{g}_1^2}\left(\left(\frac{r(M_j^{[1]})}{n}\right)^{\frac{1}{3}} \bigvee \frac{1}{n^{\frac{1}{6}}}\right) \\
\leq& D\left(\frac{\sigma_j^2 + 1}{\frac{1}{2}\sigma_j^2}\right)^2\left[\left(\frac{\sigma_j^2 + d_1 + 1}{n(\frac{1}{2}\sigma_j^2 + 1)}\right)^{\frac{1}{3}} \bigvee \frac{1}{n^{\frac{1}{6}}}\right]
\end{aligned} \tag{6.54}$$

conditional on $\{\mathcal{X}_k, k = n/2 + 1, \ldots, n\}$ under events $\mathcal{B}_n$, for large enough n, with probability at

least $1 - e^{-t}$. The last inequality holds because of (6.87), (6.89) and (6.91).

Since on the event $\mathcal{B}_n$, conditional on $\{\mathcal{X}_{n/2+1}, \ldots, \mathcal{X}_n\}$, (6.54) holds with probability at least $1-e^{-t} \geq 1-\exp\left[-\left(\frac{n(\frac{1}{2}\sigma_j^2+1)}{2(\sigma_j^2+d_1+1)}\right)^{\frac{1}{3}}\right]$, we have that on the event $\mathcal{B}_n$, (6.54) also holds with probability at least $1 - \exp\left[-\left(\frac{n(\frac{1}{2}\sigma_j^2+1)}{2(\sigma_j^2+d_1+1)}\right)^{\frac{1}{3}}\right]$. Combine with assumption $\frac{d}{n} \to 0$, we have that

$$\sqrt{n}\left(\widehat{b}_j^{(1)[1]} - b_j^{(1)[1]}\right) \cdot \mathbb{I}(\mathcal{B}_n) \xrightarrow{p} 0,$$

which leads to

$$\sqrt{n}\left(\widehat{b}_j^{(1)[1]} - b_j^{(1)[1]}\right) \xrightarrow{p} 0.$$

6.5 Proof of Theorem 8

The structure of the proof is exactly the same as that of Theorem 2, so we omit much of the details here. The key technical difficulty is the characterization of initialization, which was shown in Lemma 11.

Outline WLOG, we assume $\sigma_0 = 1$. Just as in the proof of Theorem 2, we can substitute $\widehat{\Sigma}$ with

$$\tilde{\Sigma} := \widehat{\Sigma} - \mathscr{I},$$

as the output of the algorithm will remain the same.

Note that for a given ε, inequality (A.13) in Lemma 11 can be satisfied by letting L be a polynomial of k. The polynomial $\text{poly}(r_0)$ in Theorem 8 is chosen such that $L = \text{poly}(r_0)$ satisfies

$$1 - \frac{\sqrt{2\ln(r_0)} + \sqrt{2\ln(4)}}{\sqrt{3/4}\left(\sqrt{2\ln(L)} - \frac{\ln(\ln(L))+c}{2\sqrt{2\ln(L)}} - \sqrt{2\ln(8)}\right)} \geq \frac{1}{2}. \tag{6.55}$$

For the base case, we first prove that among the "multiple starts" of initialization $\widehat{\mathbf{u}}_1^{(1),ini}(s),\ s \in [M]$ taken from the M columns of $\widehat{\mathbf{U}}\mathbf{A}$, with probability converging to 1 there exists at least one that satisfies

$$\begin{aligned} \left|\langle \widehat{\mathbf{u}}_1^{(1),ini}(s), \mathbf{u}_1^{(1)}\rangle\right| &\geq \frac{1}{2\sqrt{r_0}} \\ \left|\langle \widehat{\mathbf{u}}_1^{(1),ini}(s), \mathbf{u}_1^{(1)}\rangle\right| - \left|\langle \widehat{\mathbf{u}}_1^{(1),ini}(s), \mathbf{u}_l^{(1)}\rangle\right| &\geq \frac{1}{4\sqrt{r_0}},\ \forall l \in [r]. \end{aligned} \tag{6.56}$$

and then we show that this set of initial vectors satisfies

$$\langle \widehat{\mathbf{u}}_1^{(q),0}(s), \mathbf{u}_1^{(q)}\rangle \geq \left(\frac{63}{64}\right)^{1/p},\ q \in [p], \tag{6.57}$$

so

$$\tilde{\Sigma}(\widehat{\mathbf{u}}_1^{(1),0}(s), \ldots, \widehat{\mathbf{u}}_1^{(p),0}(s), \widehat{\mathbf{u}}_1^{(1),0}(s), \ldots, \widehat{\mathbf{u}}_1^{(p),0}(s)) \geq \frac{31}{32}.$$

Remember that at the end of the initialization step, $\{\widehat{\mathbf{u}}_1^{(q),0}\}$ are chosen to be the one that maximise

$$\left\langle \check{\Sigma}, \widehat{\mathscr{U}}_{k,0} \otimes \widehat{\mathscr{U}}_{k,0} \right\rangle,$$

so

$$\tilde{\Sigma}(\widehat{\mathbf{u}}_1^{(1),0}, \ldots, \widehat{\mathbf{u}}_1^{(p),0}, \widehat{\mathbf{u}}_1^{(1),0}, \ldots, \widehat{\mathbf{u}}_1^{(p),0}) \geq \frac{31}{32},$$

which implies

$$\max_{1\leq l\leq r}\left\{\sigma_l^2\left|\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_1^{(q),0})\rangle\right|\right\} \geq \sigma_1^2\left[\frac{31}{32} - \frac{2C\sqrt{d}}{\sqrt{n}}(\frac{1}{\sigma_1} + \frac{1}{\sigma_1^2})\right] \geq \frac{15}{16}, \tag{6.58}$$

under event $\mathcal{E}$.

Define

$$\pi(1) = \text{argmax}_{1\le l\le r}\left\{\sigma_l^2\left|\prod_{q=1}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_1^{(q),0}\rangle\right|\right\}.$$

Recall that in the iterations, $\widehat{\mathbf{u}}_k^{(q),1} \leftarrow$ the leading eigenvector of

$$\check{\Sigma}\big(\widehat{\mathbf{u}}_k^{(1),0},\ldots,\widehat{\mathbf{u}}_k^{(q-1),0},\cdot,\widehat{\mathbf{u}}_k^{(q+1),0},\ldots,\widehat{\mathbf{u}}_k^{(p),0},\widehat{\mathbf{u}}_k^{(1),0},\ldots,\widehat{\mathbf{u}}_k^{(q-1),0},\cdot,\widehat{\mathbf{u}}_k^{(q+1),0},\ldots,\widehat{\mathbf{u}}_k^{(p),0}\big),$$

so with a decomposition in the manner of (6.12), it follows that under the assumptions in Theorem 2,

$$\max_{q\in[p]}\left\{\sin\angle(\mathbf{u}_{\pi(1)}^{(q)}, \widehat{\mathbf{u}}_1^{(q),t})\right\} = \delta \le \frac{C_0\sqrt{d}}{\sqrt{n}}\left(\frac{1}{\sigma_1}+\frac{1}{\sigma_1^2}\right),\ t=1,2.$$

Then, inequality (6.3) can be proven by plugging in $\widehat{\mathbf{u}}_1^{(q),t}$ instead of $\widehat{\mathbf{u}}^{(q)}$ on the right hand side of equation (6.12).

The above arguments can be easily adapted for the induction ($k = 2,\ldots,r$) by noticing in the induction part of the proof for Theorem 2, we only need inequalities (6.2) and (6.3) to hold, yet those are exactly what we proved for the base case in this proof.

In below we will prove (6.56) and (6.57) for the base case.

Initialization By definition,

$$\begin{aligned}
&\text{contr}_1(\tilde{\Sigma})\\
&=\sum_{l=1}^{r}\sigma_l^2\mathbf{u}_l^{(1)}\otimes\mathbf{u}_l^{(1)}+\frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{r}\sigma_l^2(\theta_{lk}^2-1)\mathbf{u}_l^{(1)}\otimes\mathbf{u}_l^{(1)}+\sum_{l=1}^{r}\frac{1}{n}\sum_{k=1}^{n}\sigma_l\theta_{lk}\mathbf{u}_l^{(1)}\otimes\mathscr{E}(\cdot,\mathbf{u}_l^{(2)},\ldots,\mathbf{u}_l^{(p)})\\
&+\sum_{l=1}^{r}\frac{1}{n}\sum_{k=1}^{n}\sigma_l\theta_{lk}\mathscr{E}_k(\cdot,\mathbf{u}_l^{(2)},\ldots,\mathbf{u}_l^{(p)})\otimes\mathbf{u}_l^{(1)}+\left[\frac{1}{n}\sum_{k=1}^{n}\text{mat}_1(\mathscr{E}_k)\otimes\text{mat}_1(\mathscr{E}_k)-d_2\ldots d_p\mathbf{I}\right].
\end{aligned}$$

(I) The first term

$$\sum_{l=1}^{r}\sigma_l^2\mathbf{u}_l^{(1)}\otimes\mathbf{u}_l^{(1)}\in\mathbb{R}^{d_1}$$

serves as the signal part. By our definition of λ^*, there exists $1 \le r^* \le r_0$ such that

$$\sigma_{r^*}^2 - \sigma_{r^*+1}^2 \ge \lambda^*. \tag{6.59}$$

(II) The second term satisfies

$$\left\| \frac{1}{n} \sum_{k=1}^{n} \sum_{l=1}^{r} \sigma_l^2 (\theta_{lk}^2 - 1) \mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)} \right\| \le \frac{1}{64} \sigma_1^2$$

under event $\mathcal{E}$ (those events are defined in the proof of Theorem 2).

(III) The third term satisfies

$$\left\| \sum_{l=1}^{r} \frac{1}{n} \sum_{k=1}^{n} \sigma_l \theta_{lk} \mathbf{u}_l^{(1)} \otimes \mathscr{E}(\cdot, \mathbf{u}_l^{(2)}, \ldots, \mathbf{u}_l^{(p)}) \right\| \le \frac{C\sqrt{d}}{2\sqrt{n}} \sigma_1$$

under event $\mathcal{E}$, and the fourth term is the transpose of the third term.

(IV) Observe that

$$\sum_{k=1}^{n} \mathrm{mat}_1(\mathscr{E}_k) \otimes \mathrm{mat}_1(\mathscr{E}_k) = \sum_{i=1}^{d_2 d_3 \ldots d_p} \eta_i \eta_i^\top,$$

where η_i are i.i.d. $N(0, \mathbf{I}_{d_1})$. So following from Lemma 10,

$$\begin{aligned} \left\| \frac{1}{n} \sum_{k=1}^{n} \mathrm{mat}_1(\mathscr{E}_k) \otimes \mathrm{mat}_1(\mathscr{E}_k) - d_2 d_3 \ldots d_p \mathbf{I} \right\| &\le C \left(\sqrt{d_1} \bigvee \frac{d_1}{\sqrt{n}} \right) \frac{\sqrt{d_2 d_3 \ldots d_p}}{\sqrt{n}}, \\ &\le C \left(\frac{d^{p/2}}{\sqrt{n}} + \frac{d^{(p+1)/2}}{n} \right), \end{aligned}$$

with probability at least

$$1 - \exp(-c d_1).$$

Combine part (I) through (IV), following from Davis-Kahan $\sin(\Theta)$ Theorem, we know that as long as

$$\frac{\sigma_1^2}{\sqrt{n}} + \frac{\sqrt{d}}{\sqrt{n}} \sigma_1 + \left(\frac{d^{p/2}}{\sqrt{n}} + \frac{d^{(p+1)/2}}{n} \right) < \frac{c\lambda_*}{\sqrt{r_0}},$$

there exists $1 \leq r^* \leq r_0$ such that with probability at least

$$1 - \exp\left[-C_1\left(\frac{\min\{n,d\}}{r} - \log(r)\right)\right] - \exp(-cd),$$

$$\left\|\mathbf{U}_{r^*}\mathbf{U}_{r^*}^\top - \widehat{\mathbf{U}}_{r^*}\widehat{\mathbf{U}}_{r^*}^\top\right\| \leq \frac{1}{4\sqrt{r_0}} \leq \frac{1}{4\sqrt{r^*}},$$

in which

$$\mathbf{U}_{r^*} = \left[\mathbf{u}_1^{(1)}, \mathbf{u}_2^{(1)}, \ldots, \mathbf{u}_{r^*}^{(1)}\right] \in \mathbb{R}^{d_1 \times r^*},$$

and $\widehat{\mathbf{U}}_{r^*}$ is the matrix consists of the leading r^* eigenvectors of $\text{contr}_1(\tilde{\Sigma})$.

Then, following from Lemma 11, we have that among the "multiple starts" of initialization $\widehat{\mathbf{u}}_1^{(1),ini}(s),\ s \in [M]$ taken from the M columns of $\widehat{\mathbf{U}}\mathbf{A}$, there exists at least one that satisfies inequality (6.56), with probability at least

$$1 - \left(\frac{1}{2}\right)^{\frac{M}{\text{poly}(r_0)}}.$$

For the second mode $q = 2$,

$$\begin{aligned}
&\text{contr}_1\left[\widehat{\Sigma} \times_1 (\widehat{\mathbf{u}}_1^{(1),ini})^\top \times_{p+1} (\widehat{\mathbf{u}}_1^{(1),ini})^\top\right] \\
&= \sum_{l=1}^{r} \sigma_l^2 \langle \widehat{\mathbf{u}}_1^{(1),ini}, \mathbf{u}_l^{(1)} \rangle^2 \mathbf{u}_l^{(2)} \otimes \mathbf{u}_l^{(2)} \\
&\quad + \frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{r} \sigma_l^2(\theta_{lk}^2 - 1)\langle \widehat{\mathbf{u}}_1^{(1),ini}, \mathbf{u}_l^{(1)} \rangle^2 \mathbf{u}_l^{(2)} \otimes \mathbf{u}_l^{(2)} \\
&\quad + \sum_{l=1}^{r}\frac{1}{n}\sum_{k=1}^{n} \sigma_l\theta_{lk}\langle \widehat{\mathbf{u}}_1^{(1),ini}, \mathbf{u}_l^{(1)} \rangle \mathbf{u}_l^{(2)} \otimes \mathscr{E}(\widehat{\mathbf{u}}_1^{(1),ini}, \cdot, \mathbf{u}_l^{(3)}, \ldots, \mathbf{u}_l^{(p)}) \\
&\quad + \sum_{l=1}^{r}\frac{1}{n}\sum_{k=1}^{n} \sigma_l\theta_{lk}\mathscr{E}_k(\widehat{\mathbf{u}}_1^{(1),ini}, \cdot, \mathbf{u}_l^{(3)}, \ldots, \mathbf{u}_l^{(p)}) \otimes \langle \widehat{\mathbf{u}}_1^{(1),ini}, \mathbf{u}_l^{(1)} \rangle \mathbf{u}_l^{(2)} \\
&\quad + \left[\frac{1}{n}\sum_{k=1}^{n} \text{mat}_1(\mathscr{E}_k \times_1 (\widehat{\mathbf{u}}_1^{(1),ini})^\top) \otimes \text{mat}_1(\mathscr{E}_k \times_1 (\widehat{\mathbf{u}}_1^{(1),ini})^\top) - d_3 \ldots d_p \mathbf{I}\right].
\end{aligned}$$

(I)

$$\sum_{l=1}^{r} \sigma_l^2 \langle \widehat{\mathbf{u}}_1^{(1),ini}, \mathbf{u}_l^{(1)} \rangle^2 \mathbf{u}_l^{(2)} \otimes \mathbf{u}_l^{(2)}$$

serves as the signal part. By inequality (6.56),

$$\sigma_1^2 \langle \widehat{\mathbf{u}}_1^{(1),ini}, \mathbf{u}_1^{(1)} \rangle^2 - \sigma_l^2 \langle \widehat{\mathbf{u}}_1^{(1),ini}, \mathbf{u}_l^{(1)} \rangle^2 \geq \frac{3\sigma_1^2}{16 r_0}, \ \forall l \in [r]/\{1\},$$

which will serve as an eigengap.

(II) The second term has spectral norm no larger than $\frac{1}{64}\sigma_1^2$ under event $\mathcal{E}$.

(III): With the same methods as in the previous part, with probability converging to 1, the sum of the last three terms has spectral norm no larger than

$$C\left(\frac{\sqrt{d}\sigma_1}{\sqrt{n}} + \frac{d^{(p-1)/2}}{\sqrt{n}} + \frac{d^{p/2}}{n}\right),$$

so as long as

$$\sigma_1^2 \left(\frac{\sqrt{d}\sigma_1}{\sqrt{n}} + \frac{d^{(p-1)/2}}{\sqrt{n}} + \frac{d^{p/2}}{n}\right) \leq \frac{c\lambda^*}{r_0} \leq \frac{c\sigma_1^2}{r_0}, \tag{6.60}$$

we have (6.57) holds for $q = 2$, following from Davis-Kahan. (6.60) follows from (5.3) by noticing $r_0 \leq d$. Inequality (6.57) for $q = 3, \ldots, p$ follows in the exact same manner.

6.6 Proof of Lemma 3

The proof of Lemma 3 relies heavily on the techniques and results from Koltchinskii and Lounici 2014 which we will review first.

6.6.1 Preliminaries

Let $\mathbb{H}$ be a Hilbert space and $M : \mathbb{H} \to \mathbb{H}$ be a compact symmetric nonnegative definite operator. It is well known that the following spectral representation holds

$$M = \sum_{r>1} \mu_r \mathcal{P}_r$$

with distinct non-zero eigenvalues μ_r arranged in decreasing order $\mu_1 > \mu_2 > \cdots \geq 0$, and $\mathcal{P}_r$ are the corresponding spectral projectors. The effective rank of M is defined as

$$r(M) := \frac{\mathrm{tr}(M)}{\|M\|}.$$

We will use in particular the results from Koltchinskii and Lounici 2014 for the estimation of $\mathcal{P}_1$, in the case where

$$\mathcal{P}_1 = \mathbf{u}_1 \otimes \mathbf{u}_1,$$

i.e., estimating the leading eigenvector in the case that the leading eigenvalue is an isolated simple eigenvalue. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. centered Gaussian random vectors in $\mathbb{R}^m$ with covariance $M = \mathbb{E}(Y \otimes Y)$. Let

$$\widehat{M} := \frac{1}{n} \sum_{k=1}^{n} Y_k \otimes Y_k$$

be the sample covariance matrix based on the observations $(Y_1, Y_2, \ldots, Y_n)$. The following lemma is a restatement of Theorem 1 from Koltchinskii and Lounici 2014.

Lemma 4.

$$\mathbb{E}\|\widehat{M} - M\| \asymp \|M\| \left(\sqrt{\frac{r(M)}{n}} \bigvee \frac{r(M)}{n} \right),$$

and

$$\mathbb{E}\|\widehat{M} - M\|^2 \asymp \|M\|^2 \left(\sqrt{\frac{r(M)}{n}} \bigvee \frac{r(M)}{n} \right)^2 .$$

Let $\widehat{\mathbf{u}}_1$ be the leading eigenvector of $\widehat{M}$. Without loss of generality, to make the linear form of $\widehat{\mathbf{u}}_1$ well-defined, we always assume that $\langle \widehat{\mathbf{u}}_1, \mathbf{u}_1 \rangle \geq 0$. Denote $\widehat{\mathcal{P}}_1 := \widehat{\mathbf{u}}_1 \otimes \widehat{\mathbf{u}}_1$. Define $\bar{g}_1 := \mu_1 - \mu_2$, the spectral gap of μ_1, and write

$$C_1 = \sum_{s \neq 1} \frac{1}{\mu_1 - \mu_s} \mathcal{P}_s, \tag{6.61}$$

$$\mathcal{L}_1 := C_1(\widehat{M} - M)\mathcal{P}_1 + \mathcal{P}_1(\widehat{M} - M)C_1 = \frac{1}{n} \sum_{j=1}^{n} (C_1 Y_j \otimes \mathcal{P}_1 Y_j + \mathcal{P}_1 Y_j \otimes C_1 Y_j), \tag{6.62}$$

$$\mathcal{S}_1 := \widehat{\mathcal{P}}_1 - \mathcal{P}_1 - \mathcal{L}_1, \tag{6.63}$$

and the remainder in terms of operator

$$\mathcal{R}_1 := \widehat{\mathcal{P}}_1 - \mathbb{E}\widehat{\mathcal{P}}_1 - \mathcal{L}_1. \tag{6.64}$$

Note that $\mathbb{E}\mathcal{L}_1 = 0$, so $\mathcal{R}_1 = \mathcal{S}_1 - \mathbb{E}\mathcal{S}_1$. As in the proof of Theorem 6, we use calligraphic capital letters on $\widehat{\mathcal{P}}_1$, $\mathcal{P}_1$, C_1, $\mathcal{L}_1$, $\mathcal{S}_1$ and $\mathcal{R}_1$ to signify that they are matrices. Rephrasing Lemma 1 in Koltchinskii and Lounici 2014, we have

Lemma 5.

$$\|\mathcal{S}_1\| \leq 14 \left(\frac{\|\widehat{M} - M\|}{\bar{g}_1} \right)^2 \tag{6.65}$$

Combine Lemma 5 and 4, we have

$$\mathbb{E}\|\mathcal{S}_1\| \leq C\frac{\|M\|^2}{\bar{g}_1^2}\left(\sqrt{\frac{r(M)}{n}} \bigvee \frac{r(M)}{n}\right)^2, \tag{6.66}$$

where C is a universal constant. Restating Theorems 3 and 4 of Koltchinskii and Lounici 2014, we get

Lemma 6. *Let $t > 1$ and suppose that, for some $\gamma \in (0,1)$ and a sufficiently large constant $C > 0$,*

$$\mathbb{E}\|\widehat{M} - M\| + C\|M\|\sqrt{\frac{t}{n}} \leq \frac{1-\gamma}{1+\gamma}\frac{\bar{g}_1}{2}. \tag{6.67}$$

Then there exists a constant $D_\gamma > 0$ such that, for all $u, v \in \mathbb{H}$, the following bound holds with probability at least $1 - e^{-t}$:

$$\left|\left\langle(\widehat{\mathcal{P}}_1 - \mathbb{E}\widehat{\mathcal{P}}_1 - \mathcal{L}_1)u, v\right\rangle\right| \leq D_\gamma\frac{\|M\|^2}{\bar{g}_1^2}\left(\sqrt{\frac{r(M)}{n}} \bigvee \sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}}\|u\|\|v\|. \tag{6.68}$$

Lemma 7. *Suppose that for some $\gamma \in (0,1)$ and a sufficiently large constant $C > 0$,*

$$\mathbb{E}\|\widehat{M} - M\| + C\|M\|\frac{\log n}{n} \leq (1-\gamma)\frac{\bar{g}_1}{2}. \tag{6.69}$$

Then, there exists a constant $D_\gamma > 0$ such that

$$\|\mathbb{E}\widehat{\mathcal{P}}_1 - \mathcal{P}_1 - \mathcal{P}_1\mathbb{E}(\mathcal{S}_1)\mathcal{P}_1\| \leq D_\gamma\frac{\|M\|^2}{\bar{g}_1^2}\frac{1}{\sqrt{n}}\left(\sqrt{\frac{r(M)}{n}} \bigvee \sqrt{\frac{\log n}{n}}\right). \tag{6.70}$$

With $\mathcal{S}_1$ defined, we can define a critical quantity that characterizes the bias of $\widehat{\mathcal{P}}_1$:

$$b_1 := \langle\mathbb{E}(\mathcal{S}_1)\mathbf{u}_1, \mathbf{u}_1\rangle. \tag{6.71}$$

Note that $\mathbb{E}\mathcal{L}_1 = 0$ and $\widehat{\mathcal{P}}_1 := \widehat{\mathbf{u}}_1 \otimes \widehat{\mathbf{u}}_1$, in which $\langle \widehat{\mathbf{u}}_1, \mathbf{u}_1 \rangle \geq 0$, we have

$$b_1 = \left\langle \mathbb{E}(\widehat{\mathbf{u}}_1 \otimes \widehat{\mathbf{u}}_1)\mathbf{u}_1, \mathbf{u}_1 \right\rangle - 1,$$

so $-1 \leq b_1 \leq 0$.

There is a way to estimate b_1. Suppose we divide the sample $(Y_1, Y_2, \ldots, Y_n)$ into two subsamples of sample size $\lfloor \frac{n}{2} \rfloor$ each. Let $\check{M}$ be the sample covariance based on the first subsample and $\check{M}'$ be the sample covariance based on the second subsample. Denote by $\check{\mathbf{u}}_1$ the leading eigenvector of $\check{M}$ and by $\check{\mathbf{u}}_1'$ the leading eigenvector of $\check{M}'$. Assume that their signs are chosen in such a way that $\langle \check{\mathbf{u}}_1, \check{\mathbf{u}}_1' \rangle \geq 0$. Define

$$\widehat{b}_1 := \langle \check{\mathbf{u}}_1, \check{\mathbf{u}}_1' \rangle - 1. \tag{6.72}$$

The following lemma, a restatement of Proposition 3 from Koltchinskii and Lounici 2014, provides a concentration inequality of $|\widehat{b}_1 - b_1|$.

Lemma 8. *Let $t \geq 1$ and $\gamma \in (0, 1/2)$. There exists a constant $D_\gamma > 0$ such that, if*

$$\mathbb{E}\|\widehat{M} - M\| \leq \frac{(1-2\gamma)\bar{g}_1}{2},\ 1 + b_1 \geq 2\gamma \tag{6.73}$$

and

$$D_\gamma \|M\| \left(\sqrt{\frac{t}{n}} \bigvee \sqrt{\frac{\log n}{n}} \right) \leq \frac{\gamma \bar{g}_1}{2}, \tag{6.74}$$

then with probability at least $1 - e^{-t}$,

$$|\widehat{b}_1 - b_1| \leq D_\gamma \frac{\|M\|^2}{\bar{g}_1^2} \left(\sqrt{\frac{r(M)}{n}} \bigvee \sqrt{\frac{t}{n}} \bigvee \sqrt{\frac{\log n}{n}} \right) \sqrt{\frac{t}{n}}. \tag{6.75}$$

Proof outline. Without loss of generality, we assume $\|\mathbf{u}\|, \|\mathbf{v}\| = 1$ throughout the proof. Write

$$\mathcal{L}_j := \frac{2}{n} \sum_{k=1}^{n/2} \left(C_j y_{jk} \otimes \mathcal{P}_j y_{jk} + \mathcal{P}_j y_{jk} \otimes C_j y_{jk} \right), \tag{6.76}$$

then

$$\langle \mathcal{L}_j \mathbf{u}, \mathbf{v} \rangle = \frac{2}{n} \sum_{k=1}^{n/2} \left[\langle y_{jk}, \mathcal{P}_j \mathbf{v} \rangle \langle y_{jk}, C_j \mathbf{u} \rangle + \langle y_{jk}, \mathcal{P}_j \mathbf{u} \rangle \langle y_{jk}, C_j \mathbf{v} \rangle \right]. \tag{6.77}$$

Now Lemma 3 is equivalent to: there exists universal constant C such that

$$\mathbb{P}(|b_j^{(q)[1]}| \leq C \frac{d}{n}) \to 1, \tag{6.78}$$

and for $\forall \mathbf{u}, \mathbf{v} \in \mathbb{H}$, $\|\mathbf{u}\|, \|\mathbf{v}\| \leq 1$,

$$\sqrt{\frac{n}{2}} \left\langle \left[\widehat{\mathcal{P}}_j - \mathcal{P}_j - \mathcal{P}_j \bar{\mathbb{E}}(\mathcal{S}^{[1]}) \mathcal{P}_j - \mathcal{L}_j \right] \mathbf{u}, \mathbf{v} \right\rangle \xrightarrow{p} 0. \tag{6.79}$$

We separate the proof for (6.79) into three parts:

$$\sqrt{\frac{n}{2}} \left\langle (\widehat{\mathcal{P}}_j - \bar{\mathbb{E}} \widehat{\mathcal{P}}_j - \mathcal{L}^{[1]}) u, v \right\rangle \xrightarrow{p} 0, \tag{6.80}$$

$$\sqrt{\frac{n}{2}} \left\| \bar{\mathbb{E}} \widehat{\mathcal{P}}_j - \mathcal{P}_j - \mathcal{P}_j \bar{\mathbb{E}}(S^{[1]}) \mathcal{P}_j \right\| \xrightarrow{p} 0, \tag{6.81}$$

$$\sqrt{\frac{n}{2}} \left\langle (\mathcal{L}^{[1]} - \mathcal{L}_j) u, v \right\rangle \xrightarrow{p} 0. \tag{6.82}$$

We will first prove some preliminary bounds and (6.78), and then come back to (6.80), (6.81) and (6.82) to complete the proof.

6.6.2 Proof for (6.78).

Upper bounds (6.2) and (6.3) imply that $\widehat{\mathbf{u}}_j^{(q)[2]}$ satisfies the following conditions:

$$\lim_{n\to\infty} \mathbb{P}\left(\|\widehat{\mathbf{u}}_j^{(q)[2]} - \mathbf{u}_j^{(q)}\| \le a_n,\ \forall q \in [p] \right) = 1, \tag{6.83}$$

where a_n is a numeric sequence such that $\lim_{n\to\infty} a_n = 0$, and

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{l\neq j} \left| \sigma_l^2 \langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]} \rangle \right| \le A,\ \forall q \in [p] \right) = 1, \tag{6.84}$$

where A is a numeric constant. Define events

$$\mathcal{B}_n := \left\{ \|\widehat{\mathbf{u}}_j^{(q)[2]} - \mathbf{u}_j^{(q)}\| \le a_n,\ \|\widehat{\mathbf{u}}_j^{(q)[2]} - \mathbf{u}_j^{(q)}\| \le a_n,\ \forall q \in [p] \right\}.$$

By (6.83) and (6.84), $\lim_{n\to\infty} \mathbb{P}(\mathcal{B}_n) = 1$. Note that event $\mathcal{B}_n$ belongs to the sigma field of $\{\mathcal{X}_{n/2+1}, \mathcal{X}_{n/2+2}, \ldots, \mathcal{X}_n\}$, so it suffices to treat $\bar{\mathbb{P}}(\cdot)$ and $\bar{\mathbb{E}}(\cdot)$ as conditional on the event $\mathcal{B}_n$. Since events $\mathcal{B}_n$ satisfy $\lim_{n\to\infty} \mathbb{P}(\mathcal{B}_n) = 1$, for any sequence of random variables Z_n, to prove $Z_n \xrightarrow{p} 0$, we only need to show $Z_n \cdot \mathbb{I}(\mathcal{B}_n) \xrightarrow{p} 0$, where $\mathbb{I}(\mathcal{B}_n)$ is the indicator function of event $\mathcal{B}_n$. We will use this technique extensively.

We only need to prove for the case $q = 1$. From now on till the end of this proof, for simplicity of notations, we denote $\mathcal{P}_j = \mathcal{P}_j^{(1)}$, $C_j = C_j^{(1)}$, $\widehat{\mathcal{P}}_j = \widehat{\mathcal{P}}_j^{(1)}$, $M_j^{[1]} = M^{(1)[1]}$, $C_j^{[1]} = C_j^{(1)[1]}$, $\mathcal{L}_j^{[1]} = \mathcal{L}_j^{(1)[1]}$, $\mathcal{S}_j^{[1]} = \mathcal{S}_j^{(1)[1]}$. Furthermore, write

$$\mathcal{P}_- = I_{d_q} - \sum_{l=1}^{r} \mathbf{u}_l^{(q)} \otimes \mathbf{u}_l^{(q)} \tag{6.85}$$

Note that $C_j^{[1]}, \mathcal{L}_j^{[1]}, \mathcal{S}_j^{[1]}, R^{[1]}$ are the $C_1, \mathcal{L}_1, \mathcal{S}_1, \mathcal{R}_1$ [defined in (6.61), (6.62), (6.63), (6.64)] corresponding to our covariance matrix $M_j^{[1]}$. Conditional on $\{\mathcal{X}_k, k = n/2+1, \ldots, n\}$,

$$z_{jk} := \mathcal{X}_k \times_2 \widehat{\mathbf{u}}_j^{(1)[2]} \cdots \times_p \widehat{\mathbf{u}}_j^{(p)[2]}, k = 1, \ldots, n/2$$

has covariance matrix $M_j^{[1]}$, and $\widehat{\mathbf{u}}_j^{(q)[1]}$ is the leading eigenvector of the sample covariance matrix

$$\widehat{M}_j^{[1]} = \frac{2}{n}\sum_{k=1}^{n/2} z_{jk} \otimes z_{jk}. \tag{6.86}$$

We first prove some inequalities for $\|M_j^{[1]}\|$, $\bar{g}_1$ the first spectral gap of $M_j^{[1]}$, and $r(M_j^{[1]})$ the effective rank of $M_j^{[1]}$, under event $\mathcal{B}_n$. These inequalities will be used extensively throughout the proof.

Under $\mathcal{B}_n$, since $\|\widehat{\mathbf{u}}_j^{(q)[2]} - \mathbf{u}_j^{(q)}\| \le a_n$, we have $(1-a_n^2)^{\frac{p}{2}}\sigma_j^2 \le \tilde{\sigma}_j^2 \le \sigma_j^2$. Moreover, because $\max_{l\neq j}\left|\sigma_l^2\langle \mathbf{u}_l^{(2)}, \widehat{\mathbf{u}}_j^{(2)[2]}\rangle\right| \le A$ is bounded, and $\left|\langle \mathbf{u}_l^{(2)}, \widehat{\mathbf{u}}_j^{(2)[2]}\rangle\right| \le \sqrt{1-\langle \widehat{\mathbf{u}}_j^{(2)[2]}, \mathbf{u}_j^{(2)}\rangle^2} \le a_n$, we have $\max_{l\neq j}\{\tilde{\sigma}_l^2\} \le Aa_n$. Since $\lim_{n\to\infty} a_n = 0$, for large enough n, we have $(1-a_n^2)^{\frac{p}{2}}\sigma_j^2 > Aa_n$. So the leading eigenvector of $M_j^{[1]}$ is $\mathbf{u}_j^{(1)}$, with corresponding eigenvalue $\tilde{\sigma}_j^2 + 1$. So for large enough n,

$$\frac{1}{2}\sigma_j^2 + 1 \le (1-a_n^2)^{\frac{p}{2}}\sigma_j^2 + 1 \le \|M_j^{[1]}\| \le \sigma_j^2 + 1, \tag{6.87}$$

and the spectral gap for the leading eigenvalue of $M_j^{[1]}$

$$\bar{g}_1 \ge (1-a_n^2)^{\frac{p}{2}}\sigma_j^2 - Aa_n. \tag{6.88}$$

Since $\lim_{n\to\infty} a_n = 0$, we have for large enough n,

$$\bar{g}_1 \ge \frac{1}{2}\sigma_j^2. \tag{6.89}$$

By definition,

$$r(M_j^{[1]}) = \frac{\mathrm{tr}(M_j^{[1]})}{\|M_j^{[1]}\|}. \tag{6.90}$$

Combine with (6.87), and

$$\begin{aligned}
&\operatorname{tr}(M_j^{[1]}) \\
=&\sum_{l\neq j}\sigma_l^2\left(\prod_{q=2}^{p}\langle \mathbf{u}_l^{(q)}, \widehat{\mathbf{u}}_j^{(q)[2]}\rangle\right)^2 + \tilde{\sigma}_j^2 + d_1 \\
\leq& \max_{l\neq j}\left|\sigma_l^2\langle \mathbf{u}_l^{(2)}, \widehat{\mathbf{u}}_j^{(2)[2]}\rangle\right| + \sigma_j^2 + d_1 \\
\leq& A a_n + \sigma_j^2 + d_1
\end{aligned}$$

where the first inequality is by Cauchy inequality. So for large enough n,

$$\frac{d_1}{\sigma_j^2+1} \leq r(M_j^{[1]}) \leq \frac{\sigma_j^2 + d_1 + 1}{\frac{1}{2}\sigma_j^2 + 1}. \tag{6.91}$$

Combine inequalities (6.66) from Lemma 6 with bounds (6.87), (6.89) and (6.91), we have that under event $\mathcal{B}_n$, for large enough n,

$$\left|b_j^{(1)[1]}\right| \leq \left\|\bar{\mathbb{E}}(\mathcal{S}_j^{[1]})\right\| \leq \bar{\mathbb{E}}\left\|\mathcal{S}_j^{[1]}\right\| \leq C\frac{(\sigma_j^2+1)^2}{\frac{1}{4}\sigma_j^4}\left(\sqrt{\frac{\sigma_j^2+d_1+1}{n(\frac{1}{2}\sigma_j^2+1)}} \bigvee \frac{\sigma_j^2+d_1+1}{n(\frac{1}{2}\sigma_j^2+1)}\right)^2, \tag{6.92}$$

since we treat σ_j as fixed, the right hand side is bounded by $C\left(\sqrt{\frac{d}{n}} \vee \frac{d}{n}\right)^2 \leq C\frac{d}{n}$. Recall that $\lim_{n\to\infty}\mathbb{P}(\mathcal{B}_n) = 1$, we have $\mathbb{P}(|b_j^{(1)[1]}| \leq C\frac{d}{n}) \to 1$ as $n \to \infty$.

6.6.3 Proof for (6.80).

In this part, we will use Lemma 6 to show that

$$\sqrt{\frac{n}{2}}\left\langle (\widehat{\mathcal{P}}_j - \bar{\mathbb{E}}\widehat{\mathcal{P}}_j - \mathcal{L}_j^{[1]})\mathbf{u}, \mathbf{v}\right\rangle \xrightarrow{p} 0.$$

Let $t = \left(\frac{n}{2r(M_j^{[1]})}\right)^{\frac{1}{3}}$. First note that on the event $\mathcal{B}_n$, by inequality (6.91),

$$t \geq \left(\frac{n(\frac{1}{2}\sigma_j^2 + 1)}{2(\sigma_j^2 + d_1 + 1)}\right)^{\frac{1}{3}} \geq 1,$$

for large enough n. By Lemma 4,

$$\begin{aligned}
\delta_n(t) &= \bar{\mathbb{E}}\|\widehat{M}_j^{[1]} - M_j^{[1]}\| + C\|M_j^{[1]}\|\sqrt{\frac{t}{n}} \\
&\leq C\|M_j^{[1]}\|\left(\sqrt{\frac{r(M_j^{[1]})}{n}} \bigvee \frac{r(M_j^{[1]})}{n}\right) + C\|M_j^{[1]}\|\sqrt{\frac{t}{n}} \\
&\leq C\|M_j^{[1]}\|\left[\frac{4(\sigma_j^2 + d_1 + 1)}{(\frac{1}{2}\sigma_j^2 + 1)n} + \sqrt{\frac{2\left(\frac{n}{2r(M_j^{[1]})}\right)^{\frac{1}{3}}}{n}}\right] \\
&\leq C(\sigma_j^2 + 1)\left[\frac{\sigma_j^2 + d_1 + 1}{(\frac{1}{2}\sigma_j^2 + 1)n} + n^{-\frac{1}{3}}\right],
\end{aligned}$$

on events $\mathcal{B}_n$. The last inequality holds because of (6.91) and (6.87). Recall that $\bar{g}_1 \geq \frac{1}{2}\sigma_j^2$. We have that for large enough n, $\delta_n(t) \leq \frac{1}{6}\bar{g}_1$.

Now, by Lemma 6 with $t = \left(\frac{n}{2r(M_j^{[1]})}\right)^{\frac{1}{3}}$ and $\gamma = \frac{1}{2}$, there exists a constant $D := D_{\frac{1}{2}}$ such that on the event $\mathcal{B}_n$, conditional on $\{\mathcal{X}_{n/2+1}, \mathcal{X}_{n/2+2}, \ldots, \mathcal{X}_n\}$, with probability at least $1 - e^{-t} \geq$

$1-\exp\left[-\left(\frac{n(\frac{1}{2}\sigma_j^2+1)}{2(\sigma_j^2+d_1+1)}\right)^{\frac{1}{3}}\right]$:

$$
\begin{aligned}
&\sqrt{\frac{n}{2}}\left|\left\langle(\widehat{\mathcal{P}}_j-\bar{\mathbb{E}}\widehat{\mathcal{P}}_j-\mathcal{L}_j^{[1]})\mathbf{u},\mathbf{v}\right\rangle\right| \\
\leq & D\sqrt{\frac{n}{2}}\frac{\|M_j^{[1]}\|^2}{\bar{g}_1^2}\left(\sqrt{\frac{r(M_j^{[1]})}{n}}\bigvee\sqrt{\frac{t}{n}}\right)\sqrt{\frac{t}{n}} \\
= & D\frac{\|M_j^{[1]}\|^2}{\bar{g}_1^2}\left[\left(\frac{r(M_j^{[1]})}{n}\right)^{\frac{1}{3}}\bigvee\frac{1}{r(M_j^{[1]})^{\frac{1}{3}}(n/2)^{\frac{1}{6}}}\right] \\
\leq & 4D\left(\frac{\sigma_j^2+1}{\frac{1}{2}\sigma_j^2}\right)^2\left[\left(\frac{\sigma_j^2+d_1+1}{n(\frac{1}{2}\sigma_j^2+1)}\right)^{\frac{1}{3}}\bigvee\frac{1}{n^{\frac{1}{6}}}\right],
\end{aligned}
\tag{6.93}
$$

conditional on $\{\mathcal{X}_k, k=n/2+1,\ldots,n\}$ under events $\mathcal{B}_n$, for large enough n, with probability at least $1-e^{-t}$. The last inequality holds because of (6.87), (6.89) and (6.91).

Since on the event $\mathcal{B}_n$, conditional on $\{\mathcal{X}_{n/2+1},\mathcal{X}_{n/2+2},\ldots,\mathcal{X}_n\}$, (6.93) holds with probability at least $1-\exp\left[-\left(\frac{n(\frac{1}{2}\sigma_j^2+1)}{2(\sigma_j^2+d_1+1)}\right)^{\frac{1}{3}}\right]$, we have that on the event $\mathcal{B}_n$, (6.93) also holds with probability at least $1-\exp\left[-\left(\frac{n(\frac{1}{2}\sigma_j^2+1)}{2(\sigma_j^2+d_1+1)}\right)^{\frac{1}{3}}\right]$. Combining with the assumption that $d/n\to 0$, we have

$$
\sqrt{\frac{n}{2}}\left|\left\langle(\widehat{\mathcal{P}}_j-\bar{\mathbb{E}}\widehat{\mathcal{P}}_j-\mathcal{L}_j^{[1]})\mathbf{u},\mathbf{v}\right\rangle\right|\cdot\mathbb{I}(\mathcal{B}_n)\xrightarrow{p}0.
$$

6.6.4 Proof for (6.81).

In this part, we will use Lemma 7 to show that

$$
\sqrt{\frac{n}{2}}\left\|\bar{\mathbb{E}}\widehat{\mathcal{P}}_j-\mathcal{P}_j-\mathcal{P}_j\bar{\mathbb{E}}(\mathcal{S}_j^{[1]})\mathcal{P}_j\right\|\xrightarrow{p}0.
$$

By Lemma 4,

$$\begin{aligned}
&\bar{\mathbb{E}}\|\widehat{M}_j^{[1]} - M_j^{[1]}\| + C\|M_j^{[1]}\|\frac{\log n}{n} \\
\leq& C\|M_j^{[1]}\|\left(\sqrt{\frac{r(M_j^{[1]})}{n}} \bigvee \frac{r(M_j^{[1]})}{n}\right) + C\|M_j^{[1]}\|\frac{\log n}{n} \\
\leq& C(\sigma_j^2+1)\left[\frac{\sigma_j^2+d_1+1}{(\frac{1}{2}\sigma_j^2+1)n} + \frac{\log n}{n}\right],
\end{aligned}$$

on events $\mathcal{B}_n$. The last inequality holds because of (6.91) and (6.87). Remember that $\bar{g}_1 \geq \frac{1}{2}\sigma_j^2$ (inequality 6.89), since $d = o(n)$, we have that for large enough n, $\delta_n(t) \leq \frac{1}{4}\bar{g}_1$.

Now, by Lemma 7, with $\gamma = \frac{1}{2}$, there exists a constant $D := D_{\frac{1}{2}}$ such that on the event $\mathcal{B}_n$:

$$\begin{aligned}
&\sqrt{\frac{n}{2}}\left\|\bar{\mathbb{E}}\widehat{\mathcal{P}}_j - \mathcal{P}_j - \mathcal{P}_j\bar{\mathbb{E}}(\mathcal{S}_j^{[1]})\mathcal{P}_j\right\| \\
=& D\sqrt{n/2}\frac{\|M_j^{[1]}\|^2}{\bar{g}_1^2}\frac{1}{\sqrt{n/2}}\left(\sqrt{\frac{r(M_j^{[1]})}{n}} \bigvee \sqrt{\frac{\log n}{n}}\right) \\
\leq& 4D\left(\frac{\sigma_j^2+1}{\frac{1}{2}\sigma_j^2}\right)^2\left[\sqrt{\frac{\sigma_j^2+d_1+1}{n(\frac{1}{2}\sigma_j^2+1)}} \bigvee \sqrt{\frac{\log n}{n}}\right],
\end{aligned} \tag{6.94}$$

for large enough n. The last inequality holds because of (6.87), (6.89) and (6.91). Equation (6.81) then follows.

6.6.5 Proof for (6.82).

Note that for $\mathbf{u}, \mathbf{v} \in \mathbb{H}$,

$$\langle \mathcal{L}_j^{[1]}\mathbf{u}, \mathbf{v}\rangle = \frac{2}{n}\sum_{k=1}^{n/2}\left[\langle z_{jk}, \mathcal{P}_j\mathbf{v}\rangle\langle z_{jk}, C_j^{[1]}\mathbf{u}\rangle + \langle z_{jk}, \mathcal{P}_j\mathbf{u}\rangle\langle z_{jk}, C_j^{[1]}\mathbf{v}\rangle\right], \tag{6.95}$$

and remember that $\langle L_j\mathbf{u},\mathbf{v}\rangle = \frac{2}{n}\sum_{k=1}^{n/2}\left[\langle y_{jk},\mathcal{P}_j\mathbf{v}\rangle\langle y_{jk},C_j\mathbf{u}\rangle + \langle y_{jk},\mathcal{P}_j\mathbf{u}\rangle\langle y_{jk},C_j\mathbf{v}\rangle\right]$, so to prove (6.82), we only need to show: for $\forall\mathbf{u},\mathbf{v}\in\mathbb{H}$, $\|\mathbf{u}\|,\|\mathbf{v}\|\le 1$,

$$\sqrt{\frac{2}{n}}\sum_{k=1}^{n/2}\left[\langle z_{jk},\mathcal{P}_j\mathbf{v}\rangle\langle z_{jk},C_j^{[1]}\mathbf{u}\rangle - \langle y_{jk},\mathcal{P}_j\mathbf{v}\rangle\langle y_{jk},C_j\mathbf{u}\rangle\right]\xrightarrow{p}0. \tag{6.96}$$

We further break the proof for (6.96) into three steps.

Step 1.

$$\begin{aligned}&\langle z_{jk},\mathcal{P}_j\mathbf{v}\rangle\langle z_{jk},C_j^{[1]}\mathbf{u}\rangle - \langle y_{jk},\mathcal{P}_j\mathbf{v}\rangle\langle z_{jk},C_j^{[1]}\mathbf{u}\rangle\\ =&\langle z_{jk}-y_{jk},\mathcal{P}_j\mathbf{v}\rangle\langle z_{jk},C_j^{[1]}\mathbf{u}\rangle\end{aligned} \tag{6.97}$$

but conditional on $\{\mathcal{X}_{n/2+1},\mathcal{X}_{n/2+2},\ldots,\mathcal{X}_n\}$, $\langle z_{jk}-y_{jk},\mathcal{P}_j\mathbf{v}\rangle$ and $\langle z_{jk},C_j^{[1]}\mathbf{u}\rangle$, $k=1,\ldots,n$ are mean-0 Gaussian random variables and uncorrelated hence independent. So $\bar{\mathbb{E}}(\langle z_{jk}-y_{jk},\mathcal{P}_j\mathbf{v}\rangle\langle z_{jk},C_j^{[1]}\mathbf{u}\rangle) = 0$. Moreover, direct calculation gives us

$$\begin{aligned}&\bar{\mathbb{E}}\langle z_{jk}-y_{jk},\mathcal{P}_j\mathbf{v}\rangle^2\\ =&\langle\mathbf{u}_j^{(1)},\mathbf{v}\rangle^2\bar{\mathbb{E}}\left[\mathcal{X}_k\times_1\mathbf{u}_j^{(1)}\times_2\widehat{\mathbf{u}}_j^{(2)[2]}\cdots\times_p\widehat{\mathbf{u}}_j^{(p)[2]} - \mathcal{X}_k\times_1\mathbf{u}_j^{(1)}\times_2\mathbf{u}_j^{(2)}\cdots\times_p\mathbf{u}_j^{(p)}\right]^2\\ =&\langle\mathbf{u}_j^{(1)},\mathbf{v}\rangle^2\sigma_j^2\prod_{q=2}^{r}\left(1-\langle\widehat{\mathbf{u}}_j^{(q)[2]},\mathbf{u}_j^{(q)}\rangle\right)^2 + \langle\mathbf{u}_j^{(1)},\mathbf{v}\rangle^2\left\|\widehat{\mathbf{u}}_j^{(2)[2]}\otimes\cdots\otimes\widehat{\mathbf{u}}_j^{(p)[2]} - \mathbf{u}_j^{(2)}\otimes\cdots\otimes\mathbf{u}_j^{(p)}\right\|^2\\ =&\langle\mathbf{u}_j^{(1)},\mathbf{v}\rangle^2\sigma_j^2\prod_{q=2}^{r}\left(1-\langle\widehat{\mathbf{u}}_j^{(q)[2]},\mathbf{u}_j^{(q)}\rangle\right)^2 + \langle\mathbf{u}_j^{(1)},\mathbf{v}\rangle^2\left(2-2\prod_{q=2}^{r}\langle\widehat{\mathbf{u}}_j^{(q)[2]},\mathbf{u}_j^{(q)}\rangle\right)\\ \le&\sigma_j^2\left(\frac{1}{4}a_n^4\right)^{p-1} + 2 - 2\left(1-\frac{1}{2}a_n^2\right)^{p-1}\end{aligned} \tag{6.98}$$

since $\|\widehat{\mathbf{u}}_j^{(q)[2]}-\mathbf{u}_j^{(q)}\|\le a_n$ which implies $1-\langle\widehat{\mathbf{u}}_j^{(q)[2]},\mathbf{u}_j^{(q)}\rangle\le\frac{1}{2}a_n^2$. Also we have

$$\bar{\mathbb{E}}\left[\langle z_{jk},C_j^{[1]}\mathbf{u}\rangle\right]^2 = \langle C_j^{[1]}M_j^{[1]}C_j^{[1]}\mathbf{u},\mathbf{u}\rangle, \tag{6.99}$$

where

$$C_j^{[1]} M_j^{[1]} C_j^{[1]} := \sum_{l \neq j} \frac{\tilde{\sigma}_l^2}{\left(\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2\right)^2} \mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)} + \frac{1}{\tilde{\sigma}_j^4} \mathcal{P}_{-}. \tag{6.100}$$

As already proven, $(1 - a_n^2)^{\frac{p}{2}} \sigma_j^2 \leq \tilde{\sigma}_j^2$, and $\max_{l \neq j}\{\tilde{\sigma}_l^2\} \leq A a_n \to 0$, and for large enough n, $\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2 \geq \frac{1}{2}\sigma_j^2$, so for large enough n, we have

$$\left\| C_j^{[1]} M_j^{[1]} C_j^{[1]} \right\| \leq \frac{1}{\left[(1 - a_n^2)^{\frac{p}{2}} \sigma_j^2\right]^4}. \tag{6.101}$$

Thus,

$$\begin{aligned} &\bar{\mathbb{E}} \left[\langle z_{jk} - y_{jk}, \mathcal{P}_j \mathbf{v} \rangle \langle z_{jk}, C_j^{[1]} \mathbf{u} \rangle \right]^2 \\ \leq &\frac{1}{\left[(1 - a_n^2)^{\frac{p}{2}} \sigma_j^2\right]^4} \left[\sigma_j^2 \left(\frac{1}{4} a_n^4 \right)^{p-1} + 2 - 2 \left(1 - \frac{1}{2} a_n^2 \right)^{p-1} \right]. \end{aligned} \tag{6.102}$$

Step 2.

$$\begin{aligned} &\langle y_{jk}, \mathcal{P}_j \mathbf{v} \rangle \langle z_{jk}, C_j^{[1]} \mathbf{u} \rangle - \langle y_{jk}, \mathcal{P}_j \mathbf{v} \rangle \langle y_{jk}, C_j^{[1]} \mathbf{u} \rangle \\ =&\langle y_{jk}, \mathcal{P}_j \mathbf{v} \rangle \langle z_{jk} - y_{jk}, C_j^{[1]} \mathbf{u} \rangle \end{aligned} \tag{6.103}$$

but conditional on $\{\mathcal{X}_{n/2+1}, \mathcal{X}_{n/2+2}, \ldots, \mathcal{X}_n\}$, $\langle y_{jk}, \mathcal{P}_j \mathbf{v} \rangle$ and $\langle z_{jk} - y_{jk}, C_j^{[1]} \mathbf{u} \rangle$, $k = 1, \ldots, n$ are mean-0 Gaussian random variables and uncorrelated hence independent. So $\bar{\mathbb{E}}(\langle z_{jk} - y_{jk}, \mathcal{P}_j \mathbf{v} \rangle \langle z_{jk}, C_j^{[1]} \mathbf{u} \rangle) = 0$.

Moreover,

$$\begin{aligned}
&\bar{\mathbb{E}}\langle z_{jk} - y_{jk}, C_j^{[1]}\mathbf{u}\rangle^2 \\
&= \sum_{l\neq j} \sigma_l^2 \langle C_j^{[1]}\mathbf{u}, \mathbf{u}_l^{(1)}\rangle^2 \prod_{q=2}^{r} \left(\langle \widehat{\mathbf{u}}_j^{(q)[2]}, \mathbf{u}_l^{(q)}\rangle\right)^2 \\
&\qquad + \left\|C_j^{[1]}\mathbf{u}\right\|^2 \cdot \left\|\widehat{\mathbf{u}}_j^{(2)[2]} \otimes \cdots \otimes \widehat{\mathbf{u}}_j^{(p)[2]} - \mathbf{u}_j^{(2)} \otimes \cdots \otimes \mathbf{u}_j^{(p)}\right\|^2 \\
&\leq \left\|C_j^{[1]}\right\|^2 \max_{l\neq j} \left|\sigma_l^2 \langle \mathbf{u}_l^{(2)}, \widehat{\mathbf{u}}_j^{(2)[2]}\rangle\right| \sqrt{1 - \langle \mathbf{u}_j^{(2)}, \widehat{\mathbf{u}}_j^{(2)[2]}\rangle^2} + \left\|C_j^{[1]}\right\|^2 \left(2 - 2\left(1 - \frac{1}{2}a_n^2\right)^{p-1}\right) \qquad (6.104)
\end{aligned}$$

Since

$$C_j^{[1]} = \sum_{l\neq j} \frac{1}{\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2} \mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)} + \frac{1}{\tilde{\sigma}_j^2} \mathcal{P}_{-}, \qquad (6.105)$$

As already proven, for large enough n, $\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2 \geq \frac{1}{2}\sigma_j^2$, so for large enough n, we have

$$\left\|C_j^{[1]}\right\| \leq \frac{2}{\sigma_j^2}. \qquad (6.106)$$

Because $\|\widehat{\mathbf{u}}_j^{(q)[2]} - \mathbf{u}_j^{(q)}\| \leq a_n$, and $\max_{l\neq j} \left|\sigma_l^2 \langle \mathbf{u}_l^{(2)}, \widehat{\mathbf{u}}_j^{(2)[2]}\rangle\right| \leq A$,

$$\begin{aligned}
&\bar{\mathbb{E}}\langle z_{jk} - y_{jk}, C_j^{[1]}\mathbf{u}\rangle^2 \\
&\leq \frac{2}{\sigma_j^2}\left[Aa_n + 2 - 2\left(1 - \frac{1}{2}a_n^2\right)^{p-1}\right] \qquad (6.107)
\end{aligned}$$

and $E\langle y_{jk}, \mathcal{P}_j\mathbf{v}\rangle^2 = (\sigma_j^2 + 1)\langle \mathbf{u}_j^{(1)}, v\rangle^2$, so

$$\bar{\mathbb{E}}\left[\langle y_{jk}, \mathcal{P}_j\mathbf{v}\rangle\langle z_{jk} - y_{jk}, C_j^{[1]}\mathbf{u}\rangle\right]^2 \leq (\sigma_j^2 + 1)\frac{2}{\sigma_j^2}\left[Aa_n + 2 - 2\left(1 - \frac{1}{2}a_n^2\right)^{p-1}\right] \qquad (6.108)$$

Step 3.

$$\begin{aligned}&\langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle\langle y_{jk}, C_j^{[1]}\mathbf{u}\rangle - \langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle\langle y_{jk}, C_j\mathbf{u}\rangle\\ =&\langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle\langle y_{jk}, (C_j^{[1]} - C_j)\mathbf{u}\rangle\end{aligned} \tag{6.109}$$

but conditional on $\{\mathcal{X}_{n/2+1}, \mathcal{X}_{n/2+2}, \ldots, \mathcal{X}_n\}$, $\langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle$ and $\langle y_{jk}, (C_j^{[1]} - C_j)\mathbf{u}\rangle$, $k = 1, \ldots, n$ are mean-0 Gaussian random variables and uncorrelated hence independent. So $\bar{\mathbb{E}}(\langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle\langle y_{jk}, (C_j^{[1]} - C_j)\mathbf{u}\rangle) = 0$.

Moreover,

$$C_j^{[1]} - C_j = \sum_{l\neq j}\left(\frac{1}{\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2} - \frac{1}{\sigma_j^2}\right)\mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)} + \left(\frac{1}{\tilde{\sigma}_j^2} - \frac{1}{\sigma_j^2}\right)\mathcal{P}_- - \frac{1}{\sigma_j^2}\left(I_{\mathbb{H}} - I_{d_1}\right) \tag{6.110}$$

$$\begin{aligned}&\bar{\mathbb{E}}\langle y_{jk}, (C_j^{[1]} - C_j)\mathbf{u}\rangle^2\\ =&\left\|\left[\sum_{l\neq j}\left(\frac{1}{\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2} - \frac{1}{\sigma_j^2}\right)\mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)} + \left(\frac{1}{\tilde{\sigma}_j^2} - \frac{1}{\sigma_j^2}\right)\mathcal{P}_-\right]\mathbf{u}\right\|^2\end{aligned} \tag{6.111}$$

but as already proven, $(1 - a_n^2)^{\frac{p}{2}}\sigma_j^2 \leq \tilde{\sigma}_j^2$, and $\max_{l\neq j}\{\tilde{\sigma}_l^2\} \leq Aa_n \to 0$, and for large enough n, $\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2 \geq \frac{1}{2}\sigma_j^2$, so for large enough n, we have

$$\begin{aligned}&\left\|\sum_{l\neq j}\left(\frac{1}{\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2} - \frac{1}{\sigma_j^2}\right)\mathbf{u}_l^{(1)} \otimes \mathbf{u}_l^{(1)} + \left(\frac{1}{\tilde{\sigma}_j^2} - \frac{1}{\sigma_j^2}\right)\mathcal{P}_-\right\|\\ =&\max_{l\neq j}\left(\frac{1}{\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2} - \frac{1}{\sigma_j^2}\right)\bigvee\left(\frac{1}{\tilde{\sigma}_j^2} - \frac{1}{\sigma_j^2}\right)\\ =&\max_{l\neq j}\left(\frac{1}{\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2} - \frac{1}{\sigma_j^2}\right)\\ \leq&\frac{1}{(1 - a_n^2)^{\frac{p}{2}}\sigma_j^2 - Aa_n} - \frac{1}{\sigma_j^2}\end{aligned} \tag{6.112}$$

and $E\langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle^2 = (\sigma_j^2 + 1)\langle \mathbf{u}_j^{(1)}, v\rangle^2$, so

$$\bar{\mathbb{E}}\left[\langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle \langle y_{jk}, (C_j^{[1]} - C_j)\mathbf{u}\rangle\right]^2 \leq (\sigma_j^2 + 1)\left(\frac{1}{(1-a_n^2)^{\frac{p}{2}}\sigma_j^2 - Aa_n} - \frac{1}{\sigma_j^2}\right). \tag{6.113}$$

Combining (6.102), (6.108), (6.113), we have that on events $\mathcal{B}_n$, for large enough n,

$$\begin{aligned}
&\bar{\mathbb{E}}\left(\langle z_{jk}, \mathcal{P}_j \mathbf{v}\rangle \langle z_{jk}, C_j^{[1]}\mathbf{u}\rangle - \langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle \langle y_{jk}, C_j \mathbf{u}\rangle\right)^2 \\
\leq &\frac{1}{\left[(1-a_n^2)^{\frac{p}{2}}\sigma_j^2\right]^4}\left[\sigma_j^2\left(\frac{1}{4}a_n^4\right)^{p-1} + 2 - 2\left(1 - \frac{1}{2}a_n^2\right)^{p-1}\right] \\
&+ (\sigma_j^2 + 1)\frac{2}{\sigma_j^2}\left[Aa_n + 2 - 2\left(1 - \frac{1}{2}a_n^2\right)^{p-1}\right] + (\sigma_j^2 + 1)\left(\frac{1}{(1-a_n^2)^{\frac{p}{2}}\sigma_j^2 - Aa_n} - \frac{1}{\sigma_j^2}\right).
\end{aligned}$$

Note that conditional on $\{\mathcal{X}_{n/2+1}, \mathcal{X}_{n/2+2}, \ldots, \mathcal{X}_n\}$,

$$\langle z_{jk}, \mathcal{P}_j \mathbf{v}\rangle \langle z_{jk}, C_j^{[1]}\mathbf{u}\rangle - \langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle \langle y_{jk}, C_j \mathbf{u}\rangle$$

are independent with each other, so

$$\begin{aligned}
&\bar{\mathbb{E}}\left(\sqrt{\frac{2}{n}}\sum_{k=1}^{n/2}\left[\langle z_{jk}, \mathcal{P}_j \mathbf{v}\rangle \langle z_{jk}, C_j^{[1]}\mathbf{u}\rangle - \langle y_{jk}, \mathcal{P}_j \mathbf{v}\rangle \langle y_{jk}, C_j \mathbf{u}\rangle\right]\right)^2 \\
\leq &\frac{1}{\left[(1-a_n^2)^{\frac{p}{2}}\sigma_j^2\right]^4}\left[\sigma_j^2\left(\frac{1}{4}a_n^4\right)^{p-1} + 2 - 2\left(1 - \frac{1}{2}a_n^2\right)^{p-1}\right] \\
&+ (\sigma_j^2 + 1)\frac{2}{\sigma_j^2}\left[Aa_n + 2 - 2\left(1 - \frac{1}{2}a_n^2\right)^{p-1}\right] + (\sigma_j^2 + 1)\left(\frac{1}{(1-a_n^2)^{\frac{p}{2}}\sigma_j^2 - Aa_n} - \frac{1}{\sigma_j^2}\right),
\end{aligned}$$

on events $\mathcal{B}_n$, for large enough n. Combine with $\lim_n a_n = 0$, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left(\sqrt{\frac{2}{n}}\sum_{k=1}^{n/2}\left[\langle z_{jk}, \mathcal{P}_j\mathbf{v}\rangle\langle z_{jk}, C_j^{[1]}\mathbf{u}\rangle - \langle y_{jk}, \mathcal{P}_j\mathbf{v}\rangle\langle y_{jk}, C_j\mathbf{u}\rangle\right]\right)\cdot\mathbb{I}(\mathcal{B}_n)\right\}^2 \\
=&\mathbb{E}\left\{\bar{\mathbb{E}}\left(\sqrt{\frac{2}{n}}\sum_{k=1}^{n/2}\left[\langle z_{jk}, \mathcal{P}_j\mathbf{v}\rangle\langle z_{jk}, C_j^{[1]}\mathbf{u}\rangle - \langle y_{jk}, \mathcal{P}_j\mathbf{v}\rangle\langle y_{jk}, C_j\mathbf{u}\rangle\right]\right)^2\cdot\mathbb{I}(\mathcal{B}_n)\right\} \\
\to& 0.
\end{aligned}$$

So

$$\left(\sqrt{\frac{2}{n}}\sum_{k=1}^{n/2}\left[\langle z_{jk}, \mathcal{P}_j\mathbf{v}\rangle\langle z_{jk}, C_j^{[1]}\mathbf{u}\rangle - \langle y_{jk}, \mathcal{P}_j\mathbf{v}\rangle\langle y_{jk}, C_j\mathbf{u}\rangle\right]\right)\cdot\mathbb{I}(\mathcal{B}_n) \xrightarrow{p} 0,$$

thus proving (6.96), and (6.82) follows.

Conclusion

In this thesis, we study PCA under the settings that each observation is a matrix or more generally a multiway array. We investigate how to extract multiway PCs and study their statistical properties. In addition to the obvious advantages of increased efficiency and enhanced interpretability, our analysis provides a number of new insights into the operating characteristics of multiway PCA and their methodological implications.

First, we show that multiway PCs can be estimated without the eigengap requirement. Specifically, under a spike covariance model, we establish rates of convergence for the sample multiway PCs. In particular, they are consistent whenever the signal-to-noise ratio $\frac{\sigma_k}{\sigma_0} \gg \max\left\{\frac{d}{n}, (\frac{d}{n})^{1/4}\right\}$, where d is the dimension of one mode. Perhaps more interestingly, we prove that the sample multiway PCs are asymptotically independent of each other. In higher dimensions, the sample PCs can be biased and the bias can be corrected via sample-splitting to lead to asymptotically normal estimates of the multiway PCs, which enables us to construct confidence intervals or conduct hypothesis testing for linear forms of the PCs. We further propose a computational feasible algorithm for sample PCs, which has an initialization procedure combining matricization and random starts, followed by two-step iterations. It is shown that consistency of estimation is also achieved by the algorithm under further signal-to-noise ratio assumptions.

Our theoretical developments are complemented by numerical experiments, both simulated and real. In particular, meaningful findings can be inferred when applying our methods to two real-world datasets, further demonstrating the merits of our methodology.

References

Anandkumar, Animashree et al. (2014). "Tensor decompositions for learning latent variable models". In: *Journal of machine learning research* 15, pp. 2773–2832.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. second. New York, NY: Wiley.

Auddy, Arnab and Ming Yuan (2020). "Perturbation Bounds for (Nearly) Orthogonally Decomposable Tensors". In: *arXiv preprint arXiv:2007.09024*.

Bai, Zhidong and Jack W Silverstein (2010). *Spectral analysis of large dimensional random matrices*. Vol. 20. Springer.

Bai, Zhidong and Jianfeng Yao (2012). "On sample eigenvalues in a generalized spiked population model". In: *Journal of Multivariate Analysis* 106, pp. 167–177.

Baik, Jinho and Jack W Silverstein (2006). "Eigenvalues of large sample covariance matrices of spiked population models". In: *Journal of multivariate analysis* 97.6, pp. 1382–1408.

Benaych-Georges, Florent and Raj Rao Nadakuditi (2011). "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices". In: *Advances in Mathematics* 227.1, pp. 494–521.

Bi, Xuan et al. (2021). "Tensors in statistics". In: *Annual review of statistics and its application* 8, pp. 345–368.

Birnbaum, Aharon et al. (2013). "Minimax bounds for sparse PCA with noisy high-dimensional data". In: *Annals of statistics* 41.3, p. 1055.

Chang, Jinyuan et al. (2021). "Modelling matrix time series via a tensor CP-decomposition". In: *arXiv preprint arXiv:2112.15423*.

Chen, Elynn Y, Jianqing Fan, and Ellen Li (2020). "Statistical inference for high-dimensional matrix-variate factor model". In: *arXiv preprint arXiv:2001.01890*.

Chen, Elynn Y, Dong Xia, et al. (2020). "Semiparametric tensor factor analysis by iteratively projected svd". In: *arXiv preprint arXiv:2007.02404*.

Chen, Rong, Dan Yang, and Cun-Hui Zhang (2021). "Factor models for high-dimensional tensor time series". In: *Journal of the American Statistical Association*, pp. 1–23.

Chetty, Raj et al. (2016). “The association between income and life expectancy in the United States, 2001-2014”. In: *Jama* 315.16, pp. 1750–1766.

Cichocki, Andrzej et al. (2015). “Tensor decompositions for signal processing applications: From two-way to multiway component analysis”. In: *IEEE signal processing magazine* 32.2, pp. 145–163.

De Lathauwer, Lieven, Bart De Moor, and Joos Vandewalle (2000). “A multilinear singular value decomposition”. In: *SIAM journal on Matrix Analysis and Applications* 21.4, pp. 1253–1278.

Friedland, Shmuel (2013). “Best rank one approximation of real symmetric tensors can be chosen symmetric”. In: *Frontiers of Mathematics in China* 8.1, pp. 19–40.

Hackbusch, Wolfgang (2012). *Tensor spaces and numerical tensor calculus*. Vol. 42. Springer.

Han, Rungang, Rebecca Willett, and Anru R Zhang (2022). “An optimal statistical and computational framework for generalized tensor estimation”. In: *The Annals of Statistics* 50.1, pp. 1–29.

Han, Yuefeng, Rong Chen, et al. (2020). “Tensor factor model estimation by iterative projection”. In: *arXiv preprint arXiv:2006.02611*.

Han, Yuefeng, Cun-Hui Zhang, and Rong Chen (2021). “CP Factor Model for Dynamic Tensors”. In: *arXiv preprint arXiv:2110.15517*.

Harshman, Richard A and Margaret E Lundy (1984). “The PARAFAC model for three-way factor analysis and multidimensional scaling”. In: *Research methods for multimode data analysis* 46, pp. 122–215.

Hillar, Christopher J and Lek-Heng Lim (2013). “Most tensor problems are NP-hard”. In: *Journal of the ACM (JACM)* 60.6, pp. 1–39.

Hopkins, Samuel B, Jonathan Shi, and David Steurer (2015). “Tensor principal component analysis via sum-of-square proofs”. In: *Conference on Learning Theory*, pp. 956–1006.

Janzamin, Majid et al. (2019). “Spectral learning on matrices and tensors”. In: *Foundations and Trends® in Machine Learning* 12.5-6, pp. 393–536.

Johnstone, Iain M (2001). “On the distribution of the largest eigenvalue in principal components analysis”. In: *Annals of statistics*, pp. 295–327.

Johnstone, Iain M and Arthur Yu Lu (2009). “On consistency and sparsity for principal components analysis in high dimensions”. In: *Journal of the American Statistical Association* 104.486, pp. 682–693.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer.

Jung, Sungkyu and J Stephen Marron (2009). "PCA consistency in high dimension, low sample size context". In: *The Annals of Statistics* 37.6B, pp. 4104–4130.

Kolda, Tamara G (2001). "Orthogonal tensor decompositions". In: *SIAM Journal on Matrix Analysis and Applications* 23.1, pp. 243–255.

Koltchinskii, Vladimir, Matthias Löffler, and Richard Nickl (2020). "Efficient estimation of linear functionals of principal components". In: *The Annals of Statistics* 48.1, pp. 464–490.

Koltchinskii, Vladimir and Karim Lounici (2014). "Asymptotics and concentration bounds for spectral projectors of sample covariance". In: *arXiv preprint arXiv:1408.4643*.

Koltchinskii, Vladimir, Karim Lounici, et al. (2017). "Concentration inequalities and moment bounds for sample covariance operators". In: *Bernoulli* 23.1, pp. 110–133.

Kong, Hui et al. (2005). "Generalized 2D principal component analysis for face image representation and recognition". In: *Neural Networks* 18.5-6, pp. 585–594.

Kroonenberg, Pieter M (2008). *Applied multiway data analysis*. John Wiley & Sons.

Kroonenberg, Pieter M and Jan De Leeuw (1980). "Principal component analysis of three-mode data by means of alternating least squares algorithms". In: *Psychometrika* 45.1, pp. 69–97.

Lee, Seunggeun, Fei Zou, and Fred A Wright (2010). "Convergence and prediction of principal component scores in high-dimensional settings". In: *Annals of statistics* 38.6, p. 3605.

Li, Xuelong, Yanwei Pang, and Yuan Yuan (2010). "L1-norm-based 2DPCA". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40.4, pp. 1170–1175.

Liu, Tianqi, Ming Yuan, and Hongyu Zhao (2017). "Characterizing spatiotemporal transcriptome of human brain via low rank tensor decomposition". In: *arXiv preprint arXiv:1702.07449*.

Lu, Haiping, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos (2006). "Multilinear principal component analysis of tensor objects for recognition". In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 2. IEEE, pp. 776–779.

— (2008). "MPCA: Multilinear principal component analysis of tensor objects". In: *IEEE transactions on Neural Networks* 19.1, pp. 18–39.

— (2011). "A survey of multilinear subspace learning for tensor data". In: *Pattern Recognition* 44.7, pp. 1540–1551.

Montecino, Juan and Gerald Epstein (2015). "Did Quantitative Easing increase income inequality?" In: *Institute for New Economic Thinking working paper series* 28.

Nadler, Boaz (2008). "Finite sample approximation results for principal component analysis: A matrix perturbation approach". In: *The Annals of Statistics* 36.6, pp. 2791–2817.

Paul, Debashis (2007). "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model". In: *Statistica Sinica*, pp. 1617–1642.

Richard, Emile and Andrea Montanari (2014). "A statistical model for tensor PCA". In: *Advances in Neural Information Processing Systems*, pp. 2897–2905.

Shen, Dan et al. (2013). "Surprising asymptotic conical structure in critical sample eigen-directions". In: *arXiv preprint arXiv:1303.6171*.

Taguchi, Y-H (2018). "Tensor decomposition-based and principal-component-analysis-based unsupervised feature extraction applied to the gene expression and methylation profiles in the brains of social insects with multiple castes". In: *BMC bioinformatics* 19.4, p. 99.

Vasilescu, M Alex O and Demetri Terzopoulos (2002). "Multilinear analysis of image ensembles: Tensorfaces". In: *European conference on computer vision*. Springer, pp. 447–460.

Vershynin, Roman (2010). "Introduction to the non-asymptotic analysis of random matrices". In: *arXiv preprint arXiv:1011.3027*.

Wang, Weichen and Jianqing Fan (2017). "Asymptotics of empirical eigenstructure for high dimensional spiked covariance". In: *Annals of statistics* 45.3, p. 1342.

Xia, Dong, Anru R Zhang, and Yuchen Zhou (2020). "Inference for Low-rank Tensors–No Need to Debias". In: *arXiv preprint arXiv:2012.14844*.

Yang, Jian et al. (2004). "Two-dimensional PCA: a new approach to appearance-based face representation and recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 26.1, pp. 131–137.

Zhang, Anru and Dong Xia (2018). "Tensor svd: Statistical and computational limits". In: *IEEE Transactions on Information Theory* 64.11, pp. 7311–7338.

Zhang, Daoqiang and Zhi-Hua Zhou (2005). "(2D) 2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition". In: *Neurocomputing* 69.1-3, pp. 224–231.

Zhang, Tong and Gene H Golub (2001). "Rank-one approximation to high order tensors". In: *SIAM Journal on Matrix Analysis and Applications* 23.2, pp. 534–550.

Appendix A: Technical Lemmas

Lemma 9. *Suppose* $0 < \varepsilon < 1/p$. $(Y, \|\cdot\|_Y)$ *and* $(X_q, \|\cdot\|_q)$, $q = 1, 2, \ldots, p$ *are Banach spaces. For any mapping from* $X_1 \times X_2 \times \cdots \times X_p$ *to* Y*:*

$$(x_1, x_2, \ldots, x_p) \mapsto y(x_1, x_2, \ldots, x_p)$$

that satisfies

$$(i) \quad y(x_1, \ldots, x_{q-1}, 0, x_{q+1}, \ldots, x_p) = 0, \ \forall q \in [p], \tag{A.1}$$

$$\begin{aligned}(ii) \quad & y(x_1, \ldots, x_{q-1}, ax_q + bx'_q, x_{q+1}, \ldots, x_p) \\ =& ay(x_1, \ldots, x_{q-1}, x_q, x_{q+1}, \ldots, x_p) + by(x_1, \ldots, x_{q-1}, x'_q, x_{q+1}, \ldots, x_p), \\ & \forall q \in [p], \ \forall a, b \in \mathbb{R},\end{aligned} \tag{A.2}$$

i.e., multi-linear, denote $\mathcal{B}_q = \{x \in X_q : \|x\|_q \leq 1\}$, $\forall q \in [p]$, *and* $\mathcal{N}_q$ *are any* ε*-nets of* $\mathcal{B}_q$*, respectively, then:*

$$\begin{aligned}\sup_{x_q \in \mathcal{N}_q, \forall q \in [p]} \|y(x_1, x_2, \ldots, x_p)\|_Y &\leq \sup_{\|x_q\|_q \leq 1, \forall q \in [p]} \|y(x_1, x_2, \ldots, x_p)\|_Y \\ &\leq \frac{1}{1 - p\varepsilon} \sup_{x_q \in \mathcal{N}_q, \forall q \in [p]} \|y(x_1, x_2, \ldots, x_p)\|_Y\end{aligned} \tag{A.3}$$

Proof of Lemma 9. Denote $S := \sup_{\|x_q\|_q \leq 1, \forall q \in [p]} \|y(x_1, x_2, \ldots, x_p)\|_Y$ and $S_0 := \sup_{x_q \in \mathcal{N}_q, \forall q \in [p]} \|y(x_1, x_2, \ldots, x_p)\|_Y$.

For any $x_q \in X_q$, $\|x_q\|_q \leq 1$, we can find $x'_q \in \mathcal{N}_q$ such that $\|x_q - x'_q\|_q \leq \varepsilon$ for all $q \in [p]$,

since $\mathcal{N}_q$ are ε-nets of $\mathcal{B}_q$. Then

$$\begin{aligned} & y(x_1, x_2, \ldots, x_p) \\ =& y(x'_1, x'_2, \ldots, x'_p) + \sum_{q=1}^{p} y(x'_1, \ldots, x'_{q-1}, x_q - x'_q, x_{q+1}, \ldots, x_p). \end{aligned} \tag{A.4}$$

With conditions (A.1) and (A.2), we have

$$y(x_1, \ldots, x_{q-1}, ax_q, x_{q+1}, \ldots, x_p) = ay(x_1, \ldots, x_{q-1}, x_q, x_{q+1}, \ldots, x_p).$$

Note that $x'_1, \ldots, x'_{q-1}$ and $x_{q+1}, \ldots, x_p$ all have norms no larger than 1, so

$$\left\| y(x'_1, \ldots, x'_{q-1}, x_q - x'_q, x_{q+1}, \ldots, x_p) \right\|_Y \leq \varepsilon S.$$

Put it back into (A.4), we get

$$\|y(x_1, x_2, \ldots, x_p)\|_Y \leq \|y(x'_1, x'_2, \ldots, x'_p)\|_Y + p\varepsilon S \leq S_0 + p\varepsilon S, \tag{A.5}$$

but this holds for any $x_q \in X_q$, $\|x_q\|_q \leq 1$, so

$$S = \sup_{\|x_q\|_q \leq 1, \forall q \in [p]} \|y(x_1, x_2, \ldots, x_p)\|_Y \leq S_0 + p\varepsilon S, \tag{A.6}$$

and the upper bound in (A.3) follows. The lower bound is trivial.

□

Lemma 10. *Suppose $(X, Y) \in \mathbb{R}^{r_1+r_2}$ is a mean 0 gaussian random vector where $X \in \mathbb{R}^{r_1}$, $Y \in \mathbb{R}^{r_2}$, with covariance matrix*

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}, \tag{A.7}$$

and $(X_1, Y_1), (X_2, Y_2)$, *s are i.i.d random samples of* $(X, Y) \in \mathbb{R}^{r_1+r_2}$. *Then for any* $t > 0$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i Y_i^T - \Sigma_{xy} \right\| \leq \sqrt{\|\Sigma_{xx}\| \cdot \|\Sigma_{yy}\|} \frac{t}{\sqrt{n}} \tag{A.8}$$

with probability at least $1 - \exp\left(-c_0(\min(t^2, \sqrt{n}t) - r_1 - r_2)\right)$, *where* C *and* c_0 *are universal constants.*

Proof of Lemma 10. For $\forall u \in \mathbb{R}^{r_1}, v \in \mathbb{R}^{r_2}, \|u\|_2 = \|v\|_2 = 1$, we have that

$$u^T \left(X_i Y_i^T - \Sigma_{xy} \right) v \tag{A.9}$$

is mean 0 sub-exponential random variable with norm no larger than $C\sqrt{\|\Sigma_{xx}\| \cdot \|\Sigma_{yy}\|}$. Then by Bernstein's inequality,

$$P\left(\left| u^T \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i^T - \Sigma_{xy} \right) v \right| \geq \sqrt{\|\Sigma_{xx}\| \cdot \|\Sigma_{yy}\|} \frac{t}{\sqrt{n}} \right) \leq 2\exp(-c\min(t^2, \sqrt{n}t)). \tag{A.10}$$

Now, we bound $\left\| \frac{1}{n} \sum_{i=1}^{n} X_i Y_i^T - \Sigma_{xy} \right\|$ using lemma 9 with union bound. Here, the parameters as in lemma 9 are: $p = 2$, $(Y, \|\cdot\|_Y)$ is just $\mathbb{R}$ with absolute value, and the Banach spaces X_1 and X_2 are $\mathbb{R}^{r_1}$ and $\mathbb{R}^{r_2}$ with 2-norm. $\varepsilon = \frac{1}{4}$, $|\mathcal{N}_1| \leq 8^{r_1}$, $|\mathcal{N}_2| \leq 8^{r_2}$. The mapping is

$$(u, v) \mapsto u^T \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i^T - \Sigma_{xy} \right) v,$$

which obviously satisfies the conditions there in. Then

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}X_iY_i^T-\Sigma_{xy}\right\|\geq\left(\frac{1}{1-2\varepsilon}\right)\sqrt{\|\Sigma_{xx}\|\cdot\|\Sigma_{yy}\|}\frac{t}{\sqrt{n}}\right)\\
=&\mathbb{P}\left(\sup_{\|u\|\leq1,\|v\|\leq1}\left|u^T\left(\frac{1}{n}\sum_{i=1}^{n}X_iY_i^T-\Sigma_{xy}\right)v\right|\geq\left(\frac{1}{1-2\varepsilon}\right)\sqrt{\|\Sigma_{xx}\|\cdot\|\Sigma_{yy}\|}\frac{t}{\sqrt{n}}\right)\\
\leq&\mathbb{P}\left(\sup_{\|u\|\in\mathcal{N}_1,\|v\|\in\mathcal{N}_2}\left|u^T\left(\frac{1}{n}\sum_{i=1}^{n}X_iY_i^T-\Sigma_{xy}\right)v\right|\geq\sqrt{\|\Sigma_{xx}\|\cdot\|\Sigma_{yy}\|}\frac{t}{\sqrt{n}}\right)\\
\leq&8^{r_1+r_2}\cdot2\exp(-c\min(t^2,\sqrt{n}t))
\end{aligned}$$

i.e.

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}X_iY_i^T-\Sigma_{xy}\right\|\geq\sqrt{\|\Sigma_{xx}\|\cdot\|\Sigma_{yy}\|}\frac{t}{\sqrt{n}}\right)\leq\exp\left(-c_0(\min(t^2,\sqrt{n}t)-r_1-r_2)\right). \tag{A.11}$$

□

Lemma 11. *Suppose* $\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}$ *is a set of orthonormal vectors in* $\mathbb{R}^d$*. Let* r^* *be a integer and* $1\leq r^*\leq k$*. Denote* $\mathbf{V}=[\mathbf{v}_1,\ldots,\mathbf{v}_{r^*}]\in\mathbb{R}^{d\times r^*}$*. Suppose* $\widehat{\mathbf{V}}$ *is a* $d\times r^*$ *matrix with orthonormal columns that satisfies*

$$\operatorname{dist}(\mathbf{V},\widehat{\mathbf{V}}):=\left\|\mathbf{V}\mathbf{V}^\top-\widehat{\mathbf{V}}\widehat{\mathbf{V}}^\top\right\|\leq\frac{1}{4\sqrt{r^*}}. \tag{A.12}$$

Suppose L *is an integer that satisfies*

$$1-\frac{\sqrt{2\ln(r^*)}+\sqrt{2\ln(4)}}{\sqrt{1-\frac{1}{4\sqrt{r^*}}}\left(\sqrt{2\ln(L)}-\frac{\ln(\ln(L))+c}{2\sqrt{2\ln(L)}}-\sqrt{2\ln(8)}\right)}\geq\frac{1}{2}. \tag{A.13}$$

Then, with probability at least $1/2$ *over the choice of* L *i.i.d. random vectors drawn uniformly distributed over the unit sphere* S^{r^*-1} *in* $\mathbb{R}^{r^*}$*, at least one of the vector* θ *satisfies:*

$$\left|\left\langle\widehat{\mathbf{V}}\theta,\mathbf{v}_1\right\rangle\right|\geq\frac{1}{2\sqrt{r^*}}$$

and

$$\left|\left\langle \widehat{\mathbf{V}}\theta, \mathbf{v}_1 \right\rangle\right| - \left|\left\langle \widehat{\mathbf{V}}\theta, \mathbf{v}_i \right\rangle\right| \geq \frac{1}{4\sqrt{r^*}}, \quad \textit{for all } i \in [k]/\{1\}.$$

Proof of Lemma 11. Denote $\delta := 1/(4\sqrt{r^*})$. Consider a random matrix $\mathbf{Z} \in \mathbb{R}^{r^* \times L}$ whose entries are independent $N(0,1)$ random variables; we take the j-th column of $\mathbf{Z}$ to be comprised of the random variables used for the j-th random vector $\theta^{[j]}$ (before normalization). Let

$$\mathbf{W} := \mathbf{V}^T \widehat{\mathbf{V}} \mathbf{Z} \in \mathbb{R}^{r^* \times L},$$

so that

$$\left\langle \widehat{\mathbf{V}}\theta^{[j]}, \mathbf{v}_i \right\rangle = \left[\mathbf{V}^T \widehat{\mathbf{V}} \mathbf{Z}_{:,j} / \|\mathbf{Z}_{:,j}\|\right]_i.$$

$\mathbf{W}$ has i.i.d. mean-0 Gaussian columns with covariance

$$\Sigma = \mathbf{V}^T \widehat{\mathbf{V}} \widehat{\mathbf{V}}^T \mathbf{V}.$$

Following from condition (A.12), we have $\Sigma_{11} \geq 1 - \delta$. Observes that $\mathbf{W}_{1,1}, \mathbf{W}_{1,2}, \ldots, \mathbf{W}_{1,L}$ are i.i.d. mean-0 Gaussian random variables with variance Σ_{11}, so following from the same argument as in the proof of Lemma B.1 of Animashree Anandkumar et al. 2014, with probability at least $3/4$, there exists an index $j^* \in [l]$ such that

$$\begin{aligned} |\mathbf{W}_{1,j^*}| &\geq \sqrt{\Sigma_{11}} \left(\sqrt{2\ln(L)} - \frac{\ln(\ln(L)) + c}{2\sqrt{2\ln(L)}} - \sqrt{2\ln(8)} \right) \\ &\geq \sqrt{1-\delta} \left(\sqrt{2\ln(L)} - \frac{\ln(\ln(L)) + c}{2\sqrt{2\ln(L)}} - \sqrt{2\ln(8)} \right). \end{aligned}$$

Now for each $j \in [L]$, let

$$\max |\mathbf{W}_{2:r^*,j}| := \max\{|\mathbf{W}_{2,j}|, |\mathbf{W}_{3,j}|, \ldots, |\mathbf{W}_{r^*,j}|\}.$$

Note that $\mathbf{W}_{:,j} = \mathbf{V}^T\widehat{\mathbf{V}}\mathbf{Z}_{:,j}$, so $\max|\mathbf{W}_{2:r^*,j}|$ is a 1-Lipschitz function of $\mathbf{Z}_{:,j}$, with entries coming from r^* independent $N(0,1)$ random variables. It follows that

$$\mathbb{P}\left[\max|\mathbf{W}_{2:r^*,j}| > \mathbb{E}\left(\max|\mathbf{W}_{2:r^*,j}|\right) + \sqrt{2\ln(4)}\right] \leq \frac{1}{4}.$$

Observe that $\mathbf{W}_{2:r^*,j}$ has covariance $\Sigma_{2:r^*,2:r^*}$ which satisfies $\Sigma_{2:r^*,2:r^*} \preccurlyeq \mathbf{I}$, so by Lemma 12,

$$\mathbb{E}\left(\max|\mathbf{W}_{2:r^*,j}|\right) \leq \sqrt{2\ln(r^*)},$$

so for each $j \in [L]$,

$$\max|\mathbf{W}_{2:r^*,j}| \leq \sqrt{2\ln(r^*)} + \sqrt{2\ln(4)}.$$

with probability at least $3/4$. Therefore we conclude with a union bound that with probability at least $1/2$,

$$\begin{aligned} |\mathbf{W}_{1,j^*}| &\geq \sqrt{1-\delta}\left(\sqrt{2\ln(L)} - \frac{\ln(\ln(L)) + c}{2\sqrt{2\ln(L)}} - \sqrt{2\ln(8)}\right), \\ \max|\mathbf{W}_{2:r^*,j}| &\leq \sqrt{2\ln(r^*)} + \sqrt{2\ln(4)}. \end{aligned} \tag{A.14}$$

Following from condition (A.13), we immediately see that (A.14) means

$$|\mathbf{W}_{1,j^*}| \geq 2\max|\mathbf{W}_{2:r^*,j}|.$$

Yet $\left[\mathbf{V}^T\widehat{\mathbf{V}}\mathbf{Z}_{:,j}\,/\,\|\mathbf{Z}_{:,j}\|\right]$ has norm at least $1-\delta \geq 3/4$, and

$$\left\langle\widehat{\mathbf{V}}\theta^{[j^*]}, \mathbf{v}_i\right\rangle = \left[\mathbf{V}^T\widehat{\mathbf{V}}\mathbf{Z}_{:,j^*}\,/\,\|\mathbf{Z}_{:,j^*}\|\right]_i, \text{ where } \mathbf{V}^T\widehat{\mathbf{V}}\mathbf{Z}_{:,j^*} = \mathbf{W}_{:,j^*}, i \in [r^*],$$

so

$$\left|\left\langle \widehat{\mathbf{V}}\theta, \mathbf{v}_1 \right\rangle\right| \geq \frac{3}{4\sqrt{r^*}}$$

and

$$\left|\left\langle \widehat{\mathbf{V}}\theta, \mathbf{v}_1 \right\rangle\right| - \left|\left\langle \widehat{\mathbf{V}}\theta, \mathbf{v}_i \right\rangle\right| \geq \frac{1}{2}\left|\left\langle \widehat{\mathbf{V}}\theta, \mathbf{v}_1 \right\rangle\right| \geq \frac{1}{4\sqrt{r^*}}, \quad \text{for all } i \in [r^*]/\{1\}.$$

Observes that

$$\left|\left\langle \widehat{\mathbf{V}}\theta, \mathbf{v}_i \right\rangle\right| \leq \frac{1}{4\sqrt{r^*}} \text{ for all } i \in [k]/[r^*]$$

follows directly from (A.12), the proof is thus finished.

□

Lemma 12. *Suppose random vector $x \in \mathbb{R}^k$ has distribution $N(0, \Sigma)$ where $\Sigma \preccurlyeq \mathbf{I}$, then*

$$\mathbb{E}\left[\max_{i\in[k]} |x_i|\right] \leq \sqrt{2\ln(k)}.$$

Proof. Let $y \in \mathbb{R}^k$ be a random vector that is independent with x and has distribution $N(0, I - \Sigma)$, then

$$z := x + y$$

has distribution $N(0, I)$. By a standard argument,

$$\mathbb{E}\left[\max_{i\in[k]} |z_i|\right] \leq \sqrt{2\ln(k)},$$

so we only have to show that

$$\mathbb{E}\left[\max_{i\in[k]}|x_i|\right] \leq \mathbb{E}\left[\max_{i\in[k]}|z_i|\right].$$

Given x, let

$$i^* := \text{argmax}_{i\in[k]}\,|x_i|,$$

then observe the conditional expectation of $\max_{i\in[k]}|z_i|$ given x satisfies:

$$\mathbb{E}\left[\max_{i\in[k]}|z_i|\middle|x\right] \geq \mathbb{E}\left(|x_{i^*}+y_{i^*}|\middle|x\right) \geq \mathbb{E}\left(|x_{i^*}|+y_{i^*}\middle|x\right) = |x_{i^*}|,$$

where the last equality holds because y is mean-0 and independent with x. We finish the proof by considering unconditional expectations on both sides.

□