THE TUCKER LOW-RANK CLASSIFICATION MODEL FOR TENSOR DATA

Supplementary Materials

The Supplementary Materials provide additional estimation details, numerical studies, and theoretical proofs for the paper "The Tucker Low-Rank Classification Model For Tensor Data". Specifically, Section S1 presents detailed algorithms for the penalized estimator and the maximum likelihood estimator (MLE). Section S2 includes simulation studies on covariance MLE, matrix-variates where coefficients are randomly generated, model mis-specification, and rank selection. We provide more real data analysis in Section S3 and the comparison between TLC and tensor factor analysis in Section S4. Theoretical proofs are given in Section S5. Section S6 reviews tensor decomposition methods. At last, technical lemmas used in the proofs are summarized in Section S7.

The estimation process of TLC is constructed based on the connection between the low-rank structure in mean differences and discriminant coefficients. Therefore, two sets of Tucker decomposition expressions are used and summarized in Table S1. Note that the two sets depend on each other and could be transformed interchangeably by $\mathbf{\Phi}_k = \boldsymbol{\mathcal{G}}_k - \boldsymbol{\mathcal{G}}_1$ and $\mathbf{D}_m = \mathbf{\Sigma}_m^{-1}\mathbf{A}_m$ because $\mathbf{B}_k$ is defined as $\mathbf{B}_k = [\![\boldsymbol{\mu}_k - \boldsymbol{\mu}_1; \mathbf{\Sigma}_1^{-1}, ..., \mathbf{\Sigma}_M^{-1}]\!]$.

Tensor	Core tensor	Factor matrix
$\boldsymbol{\mu}_k$	$\boldsymbol{\mathcal{G}}_k$	$\mathbf{A}_m$
$\mathbf{B}_k$	$\mathbf{\Phi}_k$	$\mathbf{D}_m$

Table S1: Tucker decomposition expressions

In addition to notations introduced in Section 2, we further introduce following no-

tations to facilitate the readability. For two numbers a, b, define $a \vee b = \max\{a, b\}$. For two sequences of numbers a_n, b_n, we say $a_n \lesssim b_n$ if there exists a constant c such that $a_n \leq cb_n$. We say $a_n \asymp b_n$ if there exists a constant c such that $a_n \leq cb_n$ and $b_n \leq ca_n$. For two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times q}$, we say $\mathbf{A} \propto \mathbf{B}$ if there exists a constant $c \in \mathbb{R}$ such that $\mathbf{A} = c\mathbf{B}$. For a positive integer $1 \leq i \leq p$, $\mathbf{A}[i,:]$ denotes the i-th row of $\mathbf{A}$. Further define $\|\mathbf{A}\|_\infty = \max_{1\leq i\leq p} \sum_{j=1}^q |a_{ij}|$, $\|\mathbf{A}\|_1 = \max_{1\leq j\leq q} \sum_{i=1}^p |a_{ij}|$, $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$ and $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$. When $p = q$ and $\mathbf{A}$ and $\mathbf{B}$ are symmetric, we say that $\mathbf{A} \leq \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is positive semi-definite. We also define $\lambda_{\max}(\mathbf{A}), \lambda_{\min}(\mathbf{A})$ to be the largest and the smallest eigenvalue of $\mathbf{A}$, respectively. For tensor $\mathbf{A} \in \mathbb{R}^{p_1\times\cdots p_M}$, its subtensor $\mathbf{B} = \mathbf{A}_{[:,\ldots,:,k]}$ is in $\mathbb{R}^{p_1\times\cdots p_{M-1}}$ with $\mathbf{B}_{[i_1,\ldots,i_{M-1}]} = \mathbf{A}_{[i_1,\ldots,i_{M-1},k]}$. The inner product of two tensors, $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p_1\times\cdots p_M}$, is defined to be $\langle \mathbf{A}, \mathbf{B}\rangle = \sum_{i_1\cdots i_M} A_{i_1\cdots i_M} B_{i_1\cdots i_M}$.

S1 Algorithms and Rank Selection

In this section, we summarize the TLC algorithm in Algorithm S1 in Section S1.1, present the algorithm to solve (4.14) in Section S1.2, and introduce the estimation process for the maximum likelihood estimator of $\{\mathbf{B}_k\}_{k=2}^K$ in Section S1.3.

S1.1 The TLC Algorithm

Algorithm S1 Algorithm for TLC

1. Input $\widehat{\boldsymbol{\mu}}$, $\{\widehat{\boldsymbol{\Sigma}}_m\}_{m=1}^M$, and rank $\boldsymbol{r} = (r_1, ..., r_M)$.
2. Apply HOOI on $\widehat{\boldsymbol{\mu}}$ to obtain $\{\widehat{\boldsymbol{\Phi}}_k\}_{k=2}^K$ and $\{\widehat{\mathbf{A}}_m\}_{m=1}^M$.
3. For $m = 1, ..., M$, minimize the following objective function with Algorithm S1 and let $\widehat{\mathbf{D}}_m$ denote the solution
$$\min_{\mathbf{D}\in\mathbb{R}^{p_m\times r_m}} \left\{ \mathrm{tr}\left(\frac{1}{2}\mathbf{D}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_m\mathbf{D} - \widehat{\mathbf{A}}_m^{\mathrm{T}}\mathbf{D}\right) + \lambda \sum_{l=1}^{p_m} \sqrt{\sum_{j=1}^{r_m} \mathbf{D}_{lj}^2} \right\}.$$
4. Compute and output $\widehat{\mathbf{B}}_k = [\![\widehat{\boldsymbol{\Phi}}_k; \widehat{\mathbf{D}}_1, ..., \widehat{\mathbf{D}}_M]\!], k = 2, ..., K$.

S1.2 The Block Coordinate Descent Algorithm to Solve (4.14)

The algorithm to solve (4.14) relies on the following result.

Proposition S1. Let $\mathbf{D}_{l\cdot} = (d_{l1}, ..., d_{lr_m})$ denote the l-th row of the matrix $\mathbf{D} \in \mathbb{R}^{p_m \times r_m}$. With $\widehat{\mathbf{A}}_m$, $\widehat{\mathbf{\Sigma}}_m$, and $\{\mathbf{D}_{l'\cdot}, l' \neq l\}$ being known, the solution of $\mathbf{D}_{l\cdot}$ to (4.14) is

$$\widehat{\mathbf{D}}_{l\cdot} = \widetilde{\mathbf{D}}_{l\cdot} \left(1 - \frac{\lambda}{\|\widetilde{\mathbf{D}}_{l\cdot}\|_2}\right)_+, \tag{S1.1}$$

where $\widetilde{\mathbf{D}}_{l\cdot} = (\tilde{d}_{l1}, ..., \tilde{d}_{lr_m})$ with $\tilde{d}_{lj} = \frac{\hat{a}_{lj} - \sum_{t\neq l}^{p_m} d_{tj}\hat{\sigma}_{m,lt}}{\hat{\sigma}_{m,ll}}$.

Proof of Proposition S1. The conclusion can be proved in the same way as Lemma 1 in Mai et al. (2019a) and hence is omitted. □

By Proposition S1, we can solve (4.14) by iteratively updating d_{lj} while keeping all other elements fixed. Such an algorithm is summarized in Algorithm S2.

Algorithm S2 Algorithm to solve (4.14)

1. Input $\widehat{\mathbf{\Sigma}}_m$, $\widehat{\mathbf{A}}_m$. Initialize $\widehat{\mathbf{D}}_m^{(0)} = 0$.
2. For $w = 1, 2, ...$, do the following steps until convergence:
   For $l = 1, ..., p_m$, do
   (a) Based on $\widehat{\mathbf{D}}_{m,l'\cdot}^{(w-1)}$, $l' \neq l$, update $\widetilde{\mathbf{D}}_{m,l\cdot}^{(w)}$ with

$$\tilde{d}_{m,lj}^{(w)} = \frac{\hat{a}_{m,lj} - \sum_{t\neq l}^{p_m} \hat{d}_{m,tj}^{(w-1)}\hat{\sigma}_{m,lt}}{\hat{\sigma}_{m,ll}}, \quad j = 1, ..., r_m;$$

   (b) Compute

$$\widehat{\mathbf{D}}_{m,l\cdot}^{(w)} = \widetilde{\mathbf{D}}_{m,l\cdot}^{(w)} \left(1 - \frac{\lambda}{\|\widetilde{\mathbf{D}}_{m,l\cdot}^{(w)}\|_2}\right)_+.$$

3. Output $\widehat{\mathbf{D}}_m = \widehat{\mathbf{D}}_m^{(w)}$ at convergence.

S1.3 The Algorithm to Obtain Maximum Likelihood Estimators

In this section, we present the initialization, estimation algorithm, and the stopping rule to solve for maximum likelihood estimators (MLEs) of TLC.

Initialization To implement the iterative method to solve for MLEs, $\widehat{\boldsymbol{\mathcal{G}}}_k$, $\widehat{\mathbf{A}}_m$, and $\widehat{\boldsymbol{\Sigma}}_m$, we first need to determine initial values as follows.

$$\widehat{\boldsymbol{\mu}}_k^{(0)} = \frac{1}{n_k}\sum_{Y_i=k}\mathbf{X}_i, \quad n_k = \sum_{i=1}^{n} 1_{Y_i=k}, \quad \widehat{\boldsymbol{\mathcal{G}}}_k^{(0)} = \frac{1}{n_k}\sum_{Y_i=k}[\![\mathbf{X}_i; \widehat{\mathbf{J}}_1^{(0)}, ..., \widehat{\mathbf{J}}_M^{(0)}]\!], \tag{S1.2}$$

$$\widehat{\boldsymbol{\Sigma}}_m^{(0)} = \frac{1}{nq_m}\sum_{i=1}^{n}\left(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i}^{(0)}\right)_{(m)}\left(\otimes_{m'\neq m}\widehat{\boldsymbol{\Sigma}}_{m'}^{(0)}\right)^{-1}\left(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i}^{(0)}\right)_{(m)}^{\mathrm{T}}, \tag{S1.3}$$

$$\widetilde{\mathbf{A}}_m^{(0)} = \begin{bmatrix}\mathbf{I}_{r_m} \\ \mathbf{0}_{(p_m-r_m)\times r_m}\end{bmatrix}, \quad \widehat{\mathbf{A}}_m^{(0)} = \underset{\mathbf{A}_m^{\mathrm{T}}\mathbf{A}_m=\mathbf{I}_{r_m}}{\arg\max}\ \mathrm{tr}\left(\widehat{\mathbf{H}}_{1m}^{(0)}\mathbf{A}_m\right) - \frac{1}{2}\mathrm{tr}\left(\widehat{\mathbf{H}}_{2m}^{(0)}\mathbf{A}_m^{\mathrm{T}}\left(\widehat{\boldsymbol{\Sigma}}_m^{(0)}\right)^{-1}\mathbf{A}_m\right), \tag{S1.4}$$

where

$$\widehat{\mathbf{J}}_m^{(0)} = \left((\widetilde{\mathbf{A}}_m^{(0)})^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}_m^{(0)})^{-1}\widetilde{\mathbf{A}}_m^{(0)}\right)^{-1}(\widetilde{\mathbf{A}}_m^{(0)})^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}_m^{(0)})^{-1}, \quad \widehat{\mathbf{H}}_{1m}^{(0)} = \sum_{i=1}^{n}\widehat{\boldsymbol{\mathcal{G}}}_{Y_i(m)}^{(0)}\left(\otimes_{m'\neq m}(\widetilde{\mathbf{A}}_{m'}^{(0)})^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}_{m'}^{(0)})^{-1}\right)\mathbf{X}_{i(m)}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}_m^{(0)})^{-1},$$

$$\widehat{\mathbf{H}}_{2m}^{(0)} = \sum_{i=1}^{n}\widehat{\boldsymbol{\mathcal{G}}}_{Y_i(m)}^{(0)}\left(\otimes_{m'\neq m}(\widetilde{\mathbf{A}}_{m'}^{(0)})^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}_{m'}^{(0)})^{-1}\widetilde{\mathbf{A}}_{m'}^{(0)}\right)(\widehat{\boldsymbol{\mathcal{G}}}_{Y_i(m)}^{(0)})^{\mathrm{T}}.$$

The objective function in (S1.4) could be converted into a minimization problem under the constraint $\mathbf{A}_m^{\mathrm{T}}\mathbf{A}_m = \mathbf{I}_{r_m}$ and solved using the function `OptStiefelGBB` in the R-package `TRES`.

Estimation The estimation process is summarized in Algorithm S3 and implemented by applying `OptStiefelGBB` iteratively.

Algorithm S3 Algorithm for TLC MLEs

1. Input: $\boldsymbol{r}, \widehat{\boldsymbol{\mu}}_k^{(0)}, \widehat{\boldsymbol{\mathcal{G}}}_k^{(0)}, \widehat{\mathbf{A}}_m^{(0)}, \widehat{\boldsymbol{\Sigma}}_m^{(0)}, k = 1, ..., K, m = 1, ..., M.$

2. For $t = 0, 1, 2, ...,$ do the following steps until convergence or $t = t_{\max}$:

   (a) $\widehat{\boldsymbol{\mu}}_k^{(t+1)} = [\![\widehat{\boldsymbol{\mathcal{G}}}_k^{(t)}; \widehat{\mathbf{A}}_1^{(t)}, ..., \widehat{\mathbf{A}}_M^{(t)}]\!], \quad k = 1, ..., K.$

   (b) $\widehat{\boldsymbol{\Sigma}}_m^{(t+1)} = \frac{1}{nq_m} \sum_{i=1}^n \left(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i}^{(t+1)}\right)_{(m)} \left(\otimes_{m' \neq m} \widehat{\boldsymbol{\Sigma}}_{m'}^{(t+1)}\right)^{-1} \left(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i}^{(t+1)}\right)_{(m)}^{\mathrm{T}}, m = 1, ..., M.$

   (c) $\widehat{\boldsymbol{\mathcal{G}}}_k^{(t+1)} = \frac{1}{n_k} \sum_{Y_i = k} [\![\mathbf{X}_i; \widehat{\mathbf{J}}_1^{(t+1)}, ..., \widehat{\mathbf{J}}_M^{(t+1)}]\!]$, where

   $\widehat{\mathbf{J}}_m^{(t+1)} = \left((\widehat{\mathbf{A}}_m^{(t)})^{\mathrm{T}} (\widehat{\boldsymbol{\Sigma}}_m^{(t+1)})^{-1} \widehat{\mathbf{A}}_m^{(t)}\right)^{-1} (\widehat{\mathbf{A}}_m^{(t)})^{\mathrm{T}} (\widehat{\boldsymbol{\Sigma}}_m^{(t+1)})^{-1}, m = 1, ..., M, \, k = 1, ..., K.$

   (d) Solve

   $$\max_{\mathbf{A}_m^{\mathrm{T}} \mathbf{A}_m = \mathbf{I}_{r_m}} \mathrm{tr}\left(\widehat{\mathbf{H}}_{1m}^{(t+1)} \mathbf{A}_m\right) - \frac{1}{2} \mathrm{tr}\left(\widehat{\mathbf{H}}_{2m}^{(t+1)} \mathbf{A}_m^{\mathrm{T}} \left(\widehat{\boldsymbol{\Sigma}}_m^{(t+1)}\right)^{-1} \mathbf{A}_m\right), \tag{S1.5}$$

   for $\widehat{\mathbf{A}}_m^{(t+1)}$, $m = 1, ..., M$ with

   $\widehat{\mathbf{H}}_{1m}^{(t+1)} = \sum_{i=1}^n \widehat{\boldsymbol{\mathcal{G}}}_{Y_i(m)}^{(t+1)} \left(\otimes_{m' \neq m} (\widehat{\mathbf{A}}_{m'}^{(t)})^{\mathrm{T}} (\widehat{\boldsymbol{\Sigma}}_{m'}^{(t+1)})^{-1}\right) \mathbf{X}_{i(m)}^{\mathrm{T}} (\widehat{\boldsymbol{\Sigma}}_m^{(t+1)})^{-1},$

   $\widehat{\mathbf{H}}_{2m}^{(t+1)} = \sum_{i=1}^n \widehat{\boldsymbol{\mathcal{G}}}_{Y_i(m)}^{(t+1)} \left(\otimes_{m' \neq m} (\widehat{\mathbf{A}}_{m'}^{(t)})^{\mathrm{T}} (\widehat{\boldsymbol{\Sigma}}_{m'}^{(t+1)})^{-1} \widehat{\mathbf{A}}_{m'}^{(t)}\right) (\widehat{\boldsymbol{\mathcal{G}}}_{Y_i(m)}^{(t+1)})^{\mathrm{T}}.$

   (Note that $\widehat{\mathbf{H}}_{1m}^{(t+1)}$ and $\widehat{\mathbf{H}}_{2m}^{(t+1)}$ are calculated using $\widehat{\mathbf{A}}_m^{(t)}$ rather than $\widehat{\mathbf{A}}_m^{(t+1)}$, which avoids introducing another iterative sub-process.)

   end for

3. Output: $\widehat{\boldsymbol{\mathcal{G}}}_k, \widehat{\mathbf{A}}_m, \widehat{\boldsymbol{\Sigma}}_m$, and

$$\widehat{\boldsymbol{\mu}}_k = [\![\widehat{\boldsymbol{\mathcal{G}}}_k; \widehat{\mathbf{A}}_1, ..., \widehat{\mathbf{A}}_M]\!], \quad \widehat{\mathbf{B}}_k = [\![\widehat{\boldsymbol{\mathcal{G}}}_k - \widehat{\boldsymbol{\mathcal{G}}}_1; \widehat{\boldsymbol{\Sigma}}_1^{-1} \widehat{\mathbf{A}}_1, ..., \widehat{\boldsymbol{\Sigma}}_M^{-1} \widehat{\mathbf{A}}_M]\!], \quad k = 1, ..., K, m = 1, ..., M.$$

Stopping rule We use $\sum_{k=1}^K \|\widehat{\boldsymbol{\mu}}_k^{(t+1)} - \widehat{\boldsymbol{\mu}}_k^{(t)}\|_F \leq \epsilon$ as the criterion to determine if the Algorithm has achieved convergence. In simulation studies, we set $\epsilon = 10^{-4}$ and $t_{\max} = 101$.

S1.4 Rank Selection

In practice, true ranks of mean differences are rarely known and hence need to be estimated. One solution is to treat the rank as a tuning parameter and then apply cross validation to select the rank. However, it is usually computationally expensive and time consuming to tune the rank in this way, especially when it comes to multiway data.

Therefore, we propose to use the following BIC criterion to conduct rank selection,

$$\mathrm{BIC}(\boldsymbol{r}) = -2l(\widehat{\boldsymbol{\theta}}) + p_e(\boldsymbol{r}) \log(n) \tag{S1.6}$$

where $l(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} (\log f_{Y_i}(\mathbf{X}_i; \widehat{\boldsymbol{\theta}}) + \log \widehat{\pi}_{Y_i})$ is the log-likelihood function of the sample with $\boldsymbol{\theta}$ representing all parameters in TLC, and $p_e(\boldsymbol{r}) = (K-1) \prod_{m=1}^{M} r_m + \sum_{m=1}^{M} r_m(p_m - r_m)$ is the number of free parameters in $\{\mathbf{B}_k\}_{k=2}^{K}$. Like classical BIC, (S1.6) balances the goodness-of-fit for the data and the degree of freedom of the model. We search over all candidate ranks with λ set to be 0 and select the one that minimizes the proposed BIC. After we select the rank, we further find the penalized estimator by using cross validation to determine λ.

S2 Additional Simulation Studies

To implement the competing methods, we use R packages `sparsediscrim` for DLDA, `glmnet` for l_1-GLM, `penalizedLDA` for l_1-FDA, `catch` for CATCH, and the MATLAB toolbox `TensorReg` for TuckerReg. Authors of Li and Schonfeld (2014) kindly provided the code for CMDA and DGTDA. Since DLDA, l_1-GLM, and l_1-FDA are designed for vector predictors, we vectorize tensor data before applying these two methods.

S2.1 Variable Selection

To quantify the variable selection performance of the proposed penalized estimator, we use the practical true positive rate (PTPR) and the practical false positive rate (PFPR) which are defined as,

$$\text{PTPR} = \frac{|\widetilde{\mathcal{D}} \cap \mathcal{D}|}{|\mathcal{D}|}, \quad \text{PFPR} = \frac{|\widetilde{\mathcal{D}} \cap \mathcal{D}^c|}{|\mathcal{D}^c|},$$

where $\mathcal{D}$ is the set of nonzero entries in discriminant coefficients and $\widetilde{\mathcal{D}} = \{(k, i_1, ..., i_M) : |\hat{b}_{k,i_1 \cdots i_M}| \geq c, c \text{ is the 5th percentile of the absolute value of } \{\widehat{\mathbf{B}}_k\}_{k=2}^K \text{ entries in } \widehat{\mathcal{D}} \cap \mathcal{D}\}$. PTPR and PFPR of Models M1-M4 are presented in Table S2. We adopt PTPR and PFPR because we find that TLC may result in small entries that are very close to zero and most of them are false positives. As shown by the visualization of $\widehat{\mathbf{B}}_2$ for Model M1 (see Figure S1 in Supplementary Materials) and numerical results for Models M1-M4 in Table 2, these false positives have little effect on signal recovery and classification. Specifically, from Table S2 we can see that TLC preserves a high accuracy in selecting important features in all considered scenarios and bears a reasonable amount of false positives. The method l_1-FDA has similar performance with matrix-variates, but the performance deteriorates much when the method is applied on higher-order predictors. CATCH and l_1-GLM significantly under-select important variables in the matrix case and the under-selection becomes more severe when we consider 3-way tensors. The performance of TuckerReg is not reported in Table S2 because its penalty term is applied on the core tensor and hence has much worse variable selection result compared to all other methods. Results in Table S2 illustrate the advantage of TLC in variable selection, especially when data are high-order and high-dimensional.

		M1		M2			M3			M4			S.E. $\leq$
		(a)	(b)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	
TLC	PTPR	94.91	88.55	94.45	90.95	79.62	94.21	86.43	83.70	94.12	78.19	81.49	(1.71)
	PFPR	0.03	4.13	2.13	9.93	15.33	1.13	6.17	3.93	1.09	8.60	4.24	(1.07)
CATCH	PTPR	71.77	53.13	13.07	12.28	8.06	21.82	15.91	14.92	4.68	6.98	6.53	(0.50)
	PFPR	6.32	1.91	0.57	0.67	0.52	0.86	0.37	0.11	0.14	0.22	0.03	(0.15)
l_1-GLM	PTPR	46.47	37.22	8.53	7.44	6.27	3.49	2.67	1.71	1.07	0.85	0.46	(0.53)
	PFPR	2.80	0.99	0.27	0.32	0.52	0.02	0.04	0.01	0.02	0.01	0.00	(0.12)
l_1-FDA	PTPR	86.25	94.15	38.23	46.84	26.50	35.02	47.11	50.57	47.16	55.51	58.12	(2.56)
	PFPR	17.17	1.45	14.11	15.96	12.30	12.78	20.19	23.88	34.74	28.28	26.01	(1.59)

Table S2: Variable selection comparison. Mean and standard error of practical true positive rates and practical false positive rates in M1-M4.

S2.2 Covariance MLE

In parallel to the MOM covariance estimate (4.10) proposed in Section 4.1, another widely used estimator is the MLE. Under (3.1), the covariance MLE can be estimated as follows,

$$\widetilde{\pi}_k = \frac{n_k}{n}, \quad \widetilde{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{Y_i=k}\mathbf{X}_i, \quad n_k = \sum_{i=1}^{n} 1_{Y_i=k}, \quad k = 1, ..., K, \tag{S2.7}$$

$$\widetilde{\boldsymbol{\Sigma}}_m = \frac{1}{nq_m}\sum_{i=1}(\mathbf{X}_i - \widetilde{\boldsymbol{\mu}}_{Y_i})_{(m)}(\otimes_{m'\neq m}\widehat{\boldsymbol{\Sigma}}_m)^{-1}(\mathbf{X}_i - \widetilde{\boldsymbol{\mu}}_{Y_i})_{(m)}, m = 1, ..., M. \tag{S2.8}$$

The calculation of (S2.8) involves an iterative process and hence is more time-consuming than the MOM estimator. Under the same settings considered in Section 6, we report the prediction and computational cost comparison of covariance MOM estimator and MLE in Tables S3 and S4. It is clear that there is no significant difference between error rates obtained using MOM estimator and MLE of covariances. However, covariance MLE is much more time-consuming, and the time cost margin becomes larger as the dimension of data grows.

Error(%)	M1		M2			M3			M4			S.E.≤
	(a)	(b)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	
Bayes	7.11	6.65	6.34	5.90	6.41	6.54	6.72	7.14	7.79	5.23	5.15	(0.05)
TLC-Oracle (covMOM)	8.47	7.93	7.41	9.67	13.07	7.31	9.14	9.06	11.21	13.59	10.08	(0.23)
TLC-Oracle (covMLE)	8.48	7.90	7.41	9.68	13.04	7.31	9.10	9.07	11.21	13.57	10.10	(0.23)

Table S3: Prediction comparison. Mean and standard error of classification error rates in **M1**-**M4** in 100 replicates.

Time(s)	Method	M1		M2			M3			M4			S.E.≤
		(a)	(b)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	
Covariance calculation	TLC-Oracle (covMOM)	0.99	1.01	4.30	4.30	4.51	4.17	4.22	4.36	19.42	18.91	18.57	(0.17)
	TLC-Oracle (covMLE)	3.80	4.69	29.97	36.00	38.32	28.86	34.86	37.16	170.55	200.48	198.23	(1.61)
Iteration times	TLC-Oracle (covMLE)	4	5	4	4	4	4	4	4	4	4	4	(0)

Table S4: Computational cost comparison. Mean and standard error of covariance calculation time (in s) in **M1**-**M4** in 100 replicates. For TLC-Oracle (covMOM), we stop the iteration when $\|\widetilde{\boldsymbol{\Sigma}}_m^{(t+1)} - \widetilde{\boldsymbol{\Sigma}}_m^{(t)}\|_F < 10^{-6}$, $\forall m \in \{1, ..., M\}$.

S2.3 Classification on Matrix Data

Recovery of the 2-D signal in M1

For Model M1, we present the visualization of $\widehat{\mathbf{B}}_2$ in Figure S1. It can be seen that TLC recovers the cross shape well, although it induces some small false positives in the recovered image signal. However, as suggested by Table 2, Table S2, and Figure S1, these practical zeros barely have an impact on classification and signal recovery.

Randomly generated discriminant coefficients

In parallel with Models M2-M4 in Section 6, we compare the performance of TLC with alternative methods in the case where $\{\mathbf{B}_k\}_{k=2}^K$ are randomly generated matrices. Core tensors $\{\boldsymbol{\mathcal{G}}_k\}_{k=1}^K$ and factor matrices $\{\mathbf{D}_k\}_{m=1}^M$ are generated using the same way as in

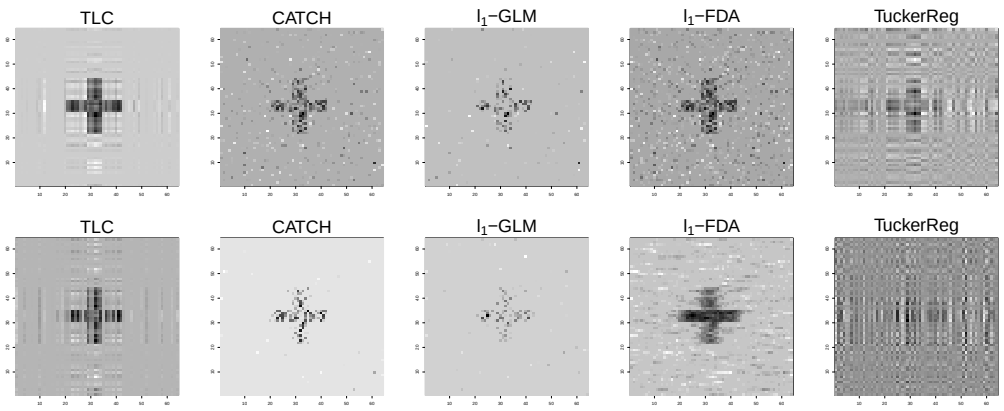

Figure S1: Snapshots of $\widehat{\mathbf{B}}_2$ in Model **M1** in one replicate.

Section 6. In particular, $\boldsymbol{\mathcal{G}}_k$'s are diagonal matrices here. Let $\pi_k = 1/K$, $\boldsymbol{p} = (64, 64)$, $\boldsymbol{r} = (3, 3)$, and $\boldsymbol{s} = (10, 10)$. We examine the performance of these methods in binary and multiclass ($K = 4$) classification with Models M5 and M6. For each model, we consider the following three covariance structures,

(a) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_{64}$,

(b) $\boldsymbol{\Sigma}_1 = AR(0.3), \boldsymbol{\Sigma}_2 = AR(0.7)$,

(c) $\boldsymbol{\Sigma}_1 = CS(0.3)$, $\boldsymbol{\Sigma}_2 = AR(0.7)$.

(a) mimics the case where all entries of a predictor are independent to each other, while (b) and (c) are cases where entries are correlated along both modes with different covariances. We report classification results in Table S5 and variable selection results in Table S6. Again, TLC has significantly outperformed alternative methods across different covariance structures and the number of classes with the lowest error rate and the highest variable selection accuracy.

Error(%)	M5			M6			S.E.≤
	(a)	(b)	(c)	(a)	(b)	(c)	
Bayes	5.72	6.86	6.39	6.47	7.38	6.40	(0.05)
TLC-Oracle (sparse, covMOM)	**6.54**	**8.30**	**8.49**	**6.85**	**8.02**	**7.22**	(0.08)
TLC-Oracle (sparse, covMLE)	**6.53**	**8.34**	**8.45**	**6.83**	**8.03**	**7.17**	(0.08)
TLC-Oracle (MLE)	9.08	11.04	10.35	**7.01**	**8.10**	**6.98**	(0.13)
TLC-BIC (sparse)	8.06	9.77	10.24	7.35	**8.02**	**7.32**	(0.19)
CATCH	8.62	10.09	9.24	8.87	9.80	8.90	(0.08)
CMDA	13.81	16.19	13.95	12.47	14.07	11.91	(0.12)
DGTDA	50.10	50.03	43.32	74.92	74.86	51.45	(0.22)
TukerReg	30.26	32.45	28.68	-	-	-	(0.23)
DLDA	20.92	18.59	18.03	23.61	20.89	20.17	(0.10)
l_1-GLM	10.25	12.00	10.09	14.82	15.75	12.18	(0.10)
l_1-FDA	8.16	13.76	17.47	9.06	13.49	17.95	(0.09)

Table S5: Prediction comparison. Mean and standard error of classification error rates in M5 and M6 in 100 replicates.

		M5			M6			S.E.≤
		(a)	(b)	(c)	(a)	(b)	(c)	
TLC	PTPR	93.20	85.52	87.29	94.61	89.88	94.14	(1.06)
	PFPR	2.48	2.54	11.19	1.19	3.13	6.52	(0.96)
CATCH	PTPR	40.50	40.95	23.53	55.89	54.15	29.69	(0.62)
	PFPR	0.35	0.95	0.25	0.01	0.20	0.07	(0.05)
l_1-GLM	PTPR	31.97	31.17	19.93	21.60	21.25	17.01	(0.57)
	PFPR	0.92	1.18	0.63	0.28	0.43	0.60	(0.08)
l_1-FDA	PTPR	48.94	62.20	76.35	39.77	34.40	66.28	(1.38)
	PFPR	1.78	8.48	36.46	0.70	2.61	36.06	(1.91)

Table S6: Variable selection comparison. Mean and standard error of practical true positive rates and practical false positive rates for M5-M6.

S2.4 Rank Selection Performance

To examine the performance of the proposed BIC, we apply it on Models M1-M6 and choose the rank that minimizes BIC among all candidate ranks. In particular, for models with matrix predictors, we search over 5 candidate ranks, $\{(1,1),(2,2),...,(5,5)\}$ so that

the selected Tucker rank is in line with the definition of matrix rank. For models with 3-way tensor predictors, we conduct a grid search over $\{1, ..., 5\} \times \{1, ..., 5\} \times \{1, ..., 5\}$ for Models M2-M3 and $\{3, ..., 7\} \times \{3, ..., 7\} \times \{3, ..., 7\}$ for Model M4. As in Wang and Zeng (2019) and Lu et al. (2020), we report the mean and standard error of selected ranks in Table S7

	M1				M2				M3			
	true $\boldsymbol{r}$	(a)	(b)	(c)	true $\boldsymbol{r}$	(a)	(b)	(c)	true $\boldsymbol{r}$	(a)	(b)	(c)
mode-1	2	1.78 (0.04)	1.00 (0)	2.02 (0.01)	3	3.00 (0)	3.16 (0.04)	4.04 (0.08)	2	2.01 (0.01)	2.00 (0)	2.29 (0.05)
mode-2	2	1.78 (0.04)	1.00 (0)	2.02 (0.01)	3	2.99 (0.01)	2.99 (0.01)	3.62 (0.07)	3	3.01 (0.01)	3.00 (0)	3.38 (0.06)
mode-3	-	-	-	-	3	3.00 (0)	3.12 (0.04)	3.98 (0.08)	4	4.00 (0)	4.09 (0.05)	4.48 (0.05)
	M4				M5				M6			
	true $\boldsymbol{r}$	(a)	(b)	(c)	true $\boldsymbol{r}$	(a)	(b)	(c)	true $\boldsymbol{r}$	(a)	(b)	(c)
mode-1	5	3.01 (0.01)	2.17 (0.06)	3.57 (0.09)	3	2.04 (0.04)	2.48 (0.06)	1.00 (0)	3	4.65 (0.06)	3.00 (0)	3.12 (0.04)
mode-2	5	3.01 (0.01)	2.53 (0.05)	3.38 (0.06)	3	2.04 (0.04)	2.48 (0.06)	1.00 (0)	3	4.65 (0.06)	3.00 (0)	3.12 (0.04)
mode-3	5	3.03 (0.02)	2.14 (0.06)	3.29 (0.08)	-	-	-	-	-	-	-	-

Table S7: Rank selection result. Mean and standard error of the selected ranks in 100 replicates using the proposed BIC.

As we can see from Table S7, the true rank is within 3 standard errors of the selected rank under most settings, especially when data are of high-order and when there are multiple classes. This agrees with intuition because high-order variates tend to contain more structure information than matrix ones when they have similar numbers of elements. Similarly, the structure information could be better collected and analyzed when there are multiple classes due to constant loading matrices. On the other hand, when data are high-dimensional along all modes and have high-rank, e.g., Model **M4**, their intrinsic structures become less parsimonious, which makes the rank selection more challenging. However,

even in this case, the difference between error rates obtained by using the true rank and the selected rank is not significant, which supports the application of the proposed BIC criterion.

S2.5 Model mis-specification

In this section, we test the robustness of the sparse TLC estimator under the following settings where TLC assumptions are violated.

- **TuckerReg**: Assume that data follow the TuckerReg model

$$\Pr(Y = 1|\mathbf{X}) = p, \quad \log \frac{p}{1-p} = \beta_0 + \langle \mathbf{B}, \mathbf{X} \rangle \tag{S2.9}$$

  where the regression coefficient $\mathbf{B}$ is the one used in **M1** without scaling.

  (a) Each element in $\mathbf{X}$ follows $N(0, 1)$ independently and identically

  (b) Each element in $\mathbf{X}$ follows t_4 independently and identically

- **Separable covariance assumption being violated**: Assume that

$$\Pr(Y = k) = \pi_k, \quad \text{vec}(\mathbf{X})|(Y = k) \sim \text{N}(\text{vec}(\boldsymbol{\mu}), \boldsymbol{\Sigma}) \tag{S2.10}$$

$$\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} = [\![\boldsymbol{\mathcal{G}}; \mathbf{A}_1, ..., \mathbf{A}_M]\!], \quad k = 1, ..., K, \tag{S2.11}$$

  where $\boldsymbol{\Sigma} \in \mathbb{R}^{\prod_{m=1}^{M} p_m \times \prod_{m=1}^{M} p_m}$. It follows that $\text{vec}(\mathbf{B}_k) = \boldsymbol{\Sigma}^{-1}\text{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$. Set $K = 2$, $\pi_1 = \pi_2 = 1/2$, $\boldsymbol{p} = (10, 10, 10)$, $\boldsymbol{r} = (3, 3, 3)$, and consider following covariance settings.

  (a) $\boldsymbol{\Sigma} = \text{AR}(0.5)$

  (b) $\boldsymbol{\Sigma}$ = Erdős-Rényi random graph, ER: Let $\widetilde{\boldsymbol{\Omega}} = (\widetilde{\omega}_{ij})$ where $\widetilde{\omega}_{ij} = u_{ij}\delta_{ij}$ with $\delta_{ij} \sim \text{Ber}(1, 0.05)$ and $u_{ij} \sim \text{Unif}[0.5, 1] \cup [-1, -0.5]$. Symmetrize the matrix by

setting $\check{\mathbf{\Omega}} = (\widetilde{\mathbf{\Omega}} + \widetilde{\mathbf{\Omega}}^T)/2$. Let $\mathbf{\Omega} = \check{\mathbf{\Omega}} + \{\max(-\lambda_{\min}(\check{\mathbf{\Omega}}), 0) + 0.05\}\mathbf{I}_p$ and then rescale it so that $\mathbf{\Omega}$ is positive definite with unit diagonals. Let $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$.

Note that $\mathbf{B}_k$ does not share the same rank with $\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}$.

- **Low-rankness assumption being violated**: Set $K = 2$, $\pi_1 = \pi_2 = 1/2$, and $\boldsymbol{p} = (64, 64)$. Generate $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$ from $N(0, 1)$ and scale it so that the Bayes error is controlled to be 5-10%. Note that $\mathbf{B}_2$ is dense in this case. Consider following covariances settings.

  (a) $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_{64}$.

  (b) $\mathbf{\Sigma}_1 = AR(0.3), \mathbf{\Sigma}_2 = AR(0.7)$.

  (c) $\mathbf{\Sigma}_1 = CS(0.3)$, $\mathbf{\Sigma}_2 = AR(0.7)$.

Model	TuckerReg		Non-seperable covariance		Full-rankness		
	(a)	(b)	(a)	(b)	(a)	(b)	(c)
Bayes	-	-	10.53 (0.06)	10.51 (0.06)	5.31 (0.04)	6.82 (0.04)	5.30 (0.04)
TLC-Oracle (sparse)	**26.21** (0.26)	**25.79** (0.26)	**14.17** (0.14)	**16.15** (0.34)	**19.98** (0.09)	**23.40** (0.10)	**20.22** (0.09)
TLC-BIC (sparse)	**25.42** (0.15)	**25.08** (0.16)	**13.92** (0.11)	20.33 (0.11)	42.18 (0.15)	48.44 (0.15)	49.22 (0.12)
CATCH	41.22 (0.14)	41.54 (0.14)	18.84 (0.10)	19.68 (0.29)	29.72 (0.11)	43.26 (0.16)	41.53 (0.13)
CMDA	37.92 (0.47)	36.46 (0.47)	23.62 (0.14)	24.48 (0.15)	38.60 (0.12)	39.23 (0.12)	37.50 (0.13)
DGTDA	49.84 (0.08)	49.95 (0.10)	50.03 (0.10)	50.11 (0.09)	49.95 (0.09)	50.11 (0.09)	50.06 (0.08)
TuckerReg	29.88 (0.41)	29.88 (0.37)	18.04 (0.13)	18.07 (0.14)	43.49 (0.12)	44.46 (0.11)	43.00 (0.12)
DLDA	40.66 (0.10)	40.71 (0.10)	31.17 (0.11)	26.25 (0.19)	**20.00** (0.09)	44.69 (0.09)	47.63 (0.10)
l_1-GLM	41.81 (0.14)	42.12 (0.15)	20.93 (0.13)	**16.48** (0.10)	34.11 (0.14)	45.15 (0.14)	44.24 (0.15)
l_1-FDA	40.17 (0.10)	40.44 (0.11)	22.98 (0.14)	**15.99** (0.11)	**20.15** (0.09)	44.75 (0.09)	47.79 (0.12)

Table S8: Prediction comparison. Mean and standard error of classification error rates in 100 replicates.

Overall, TLC-Oracle (sparse) performs the best or very close to the best under all settings in this section, and TLC-BIC (sparse) is fairly competitive under most settings. When data are generated following the TuckerReg model, TLC outperforms compet-

ing methods when $\mathbf{X}$ follows the standard normal distribution and t_4-distribution. This suggests the stableness of TLC when data follow non-normal heavy-tailed distributions. When the Kronecker covariance assumption in (3.1) is violated, the performance of TLC-Oracle (sparse) is among the best and TLC-BIC (sparse) also has comparable performance to other methods. This is because the low-rank structure in mean contrasts is still informative in projecting data onto a low-dimensional subspace and conducting classification. When the low-rankness assumption is violated, the setting reduces to that of the TDA model without sparsity. Not surprisingly, TLC-BIC (sparse) is outperformed by other methods because TLC-BIC (sparse) under-selects the rank and pursues for a low-dimensional subspace ineffectively with information loss. However, it is encouraging to note that TLC-Oracle (sparse) performs well or even the best under this setting. This indicates the effectiveness of estimation process of TLC when the rank is selected correctly. In general, TLC is robust when there exists mild model mis-specification and can achieve promising performance under various settings.

S3 Additional Real Data Analysis

In this section, we present additional analysis on the Gene Time Course dataset and provide the application of our method on another three datasets. Moreover, as recommended by the AE, we conduct and report the results of the covariance separability test (Aston et al., 2017) for these datasets to assess the validity of the Kronecker product assumption.

S3.1 Gene Time Course Data

To examine the relationship between gene expression profiles and patients' responses to rIFNβ. We examine the dataset with the $\widehat{\mathbf{B}}_2$ obtained by the TLC model. Specifically, we apply TLC on the whole dataset and tune the parameter using 10-fold cross-validation. The discriminant coefficient estimate is plotted in Figure S2. It can be seen that only 42 out of 76 genes have nonzero coefficients, which suggests that only part of the measured genes may have an impact on responses to the treatment. Factor matrix estimates are plotted in Figure S3. For $\widehat{\mathbf{D}}_1$, coefficients are all less than 0.26 and show little variability across time after baseline. Moreover, coefficients corresponding to 3 month and 9 month are close to zero, suggesting that we can ignore corresponding gene expressions when performing classification. Considering $\widehat{\mathbf{D}}_1 \in \mathbb{R}^{7\times 1}$, this implies that even a partial average of gene expressions does not lead to any information loss. In other words, the response to the treatment is invariant to changes of gene expressions across time. For $\widehat{\mathbf{D}}_2$, out of the 42 genes that may contribute to the binary classification, the top 4 genes that have the largest negative and positive coefficients are Jak2, STAT-6, Caspase 10, CD80, and IRF8, FLIP, RIP, STAT3, respectively.

To investigate the effect of time and individual genes in more details, we present side-by-side boxplots from the two classes after reducing $\mathbf{X}_i$ to $\mathbf{X}_i^{\mathsf{T}}\widehat{\mathbf{D}}_1 \in \mathbb{R}^{76\times 1}$ and $\mathbf{X}_i\widehat{\mathbf{D}}_2 \in \mathbb{R}^{7\times 1}$. Figures S6 and S7 suggest that expression values combined across time for each single gene are not discriminative enough, even for Jak2 and IRF8 which have the largest negative and positive coefficient values in $\widehat{\mathbf{D}}_2$. However, if we consider a linear combination of all gene values instead, as shown in Figure S8, the separation is perfect and presents a similar pattern at different time points. This also implies that gene expressions may be

associated with patients' responses. Furthermore, such an association is not influenced by time.

To visualize the classification performance of TLC, we present the scatter plot of $\langle \mathbf{X}_i, \widehat{\mathbf{B}}_2 \rangle$, which is equivalent to $\langle \widetilde{\mathbf{X}}_i, \widehat{\mathbf{\Phi}}_2 \rangle$ with $\widetilde{\mathbf{X}}_i = [\![\mathbf{X}_i; \widehat{\mathbf{D}}_1, \widehat{\mathbf{D}}_2]\!]$, in Figure S4. It is clear that the two classes are well separated. Together, the performance of TLC and the estimated parameters suggest that there may exist an association between the baseline gene expressions and patients' responses to the treatment, which is in agreement with the conclusions in Baranzini et al. (2004), Lyu et al. (2017), and Molstad and Rothman (2019).

S3.2 Finger-Tapping fMRI Data

The Finger-Tapping fMRI dataset was collected by Maitra et al. (2002). This dataset provides 12 functional magnetic resonance imaging (fMRI) scans of the brain of a right-hand dominant male subject when he was performing a right-hand finger-thumb opposition task. Another 12 fMRI scans of the same subject when he used the left hand were recorded as well. Each pair of scans includes 12 sessions within a two-month period. As suggested by Thompson et al. (2020), we focus on the analysis of the 20th slice of the image volume as it is adequate to distinguish the activation between the left- and right-hand finger-tapping (Maitra, 2010). The 20th slice of the image has 128×128 pixels. Considering the limited sample size (24 observations in total), we follow the procedure in Thompson et al. (2020) and select a 20×20 section of the 20th slice which has the left-topmost pixel located at (33, 67). This section displays the highest average activation in the left-hand activation images as determined by the FAST-fMRI algorithm in Almodóvar-Rivera and Maitra (2019). We consider the binary classification problem which distinguishes right-

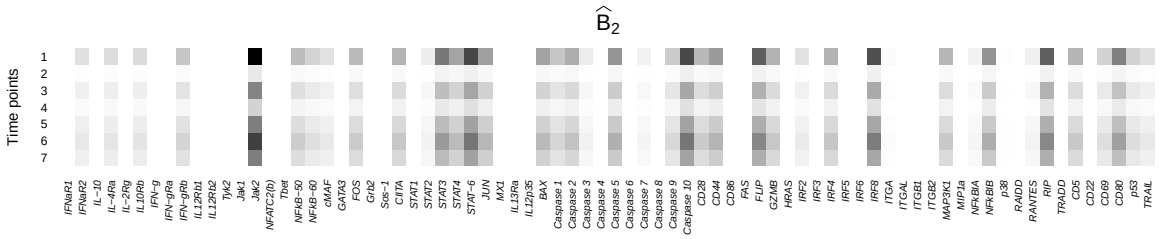

Figure S2: The absolute value of $\widehat{\mathbf{B}}_2$. White entries have zero values, while dark entries correspond to values of large magnitude.

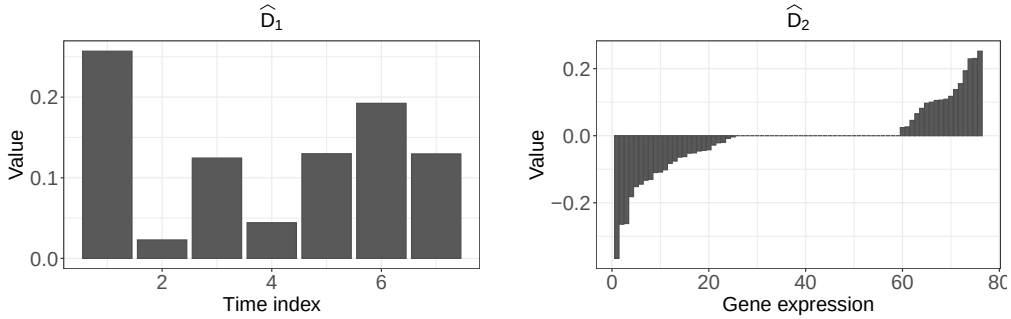

Figure S3: Coefficients of $\widehat{\mathbf{D}}_1$ and $\widehat{\mathbf{D}}_2$.

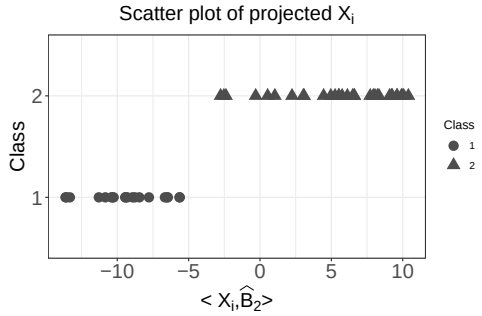

Figure S4: The scatter plot of $\langle \mathbf{X}_i, \widehat{\mathbf{B}}_2 \rangle$.

Figure S5: Gene Time Course Data. Estimated coefficients and the classification result.

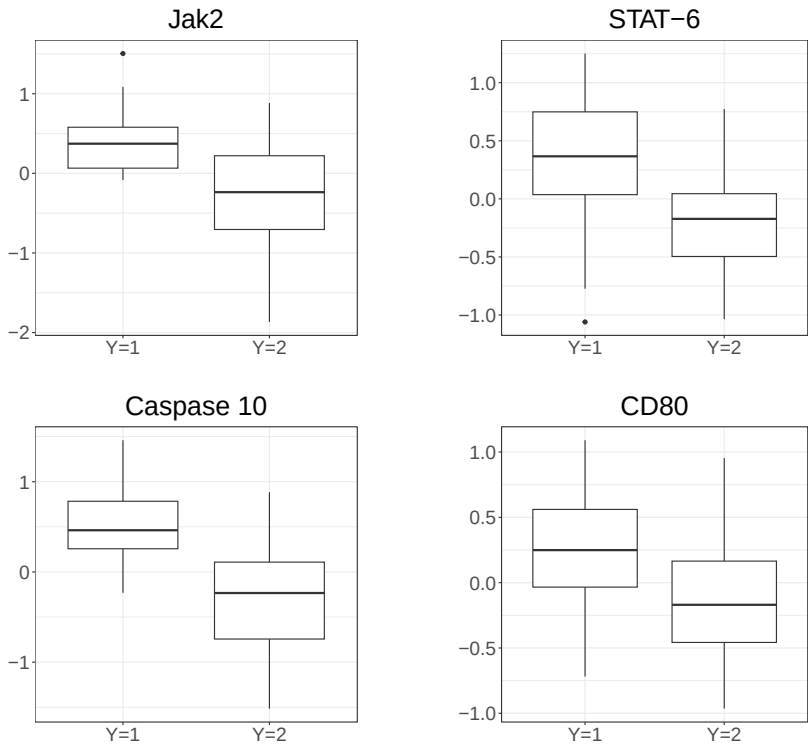

Figure S6: Side-by-side boxplots of gene expression combinations across time for STAT-6, Caspase10, and CD80 from the two groups of patients.

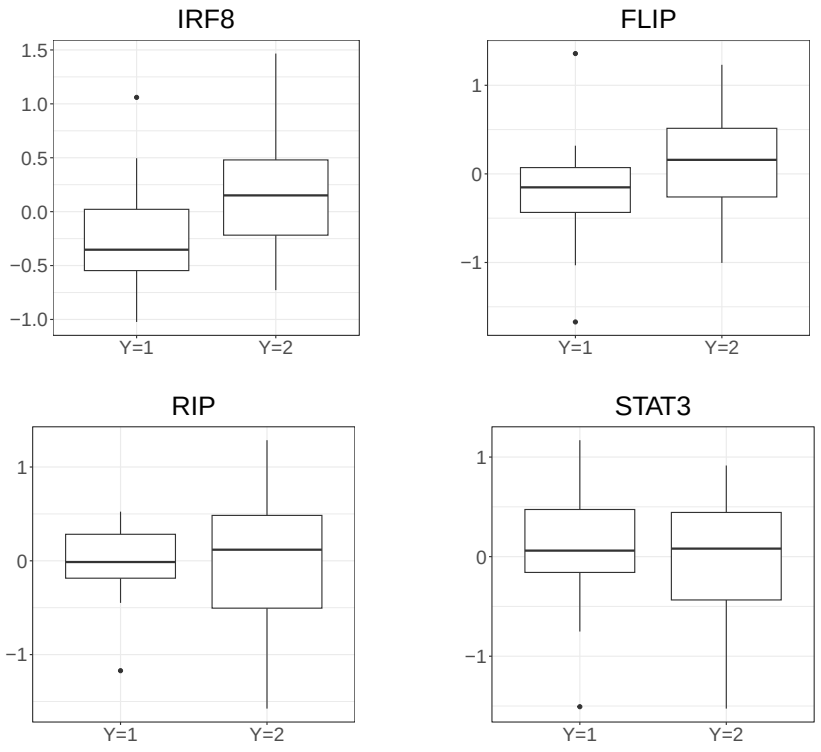

Figure S7: Side-by-side boxplots of gene expression combinations across time for FLIP, RIP, and STAT3 from the two groups of patients.

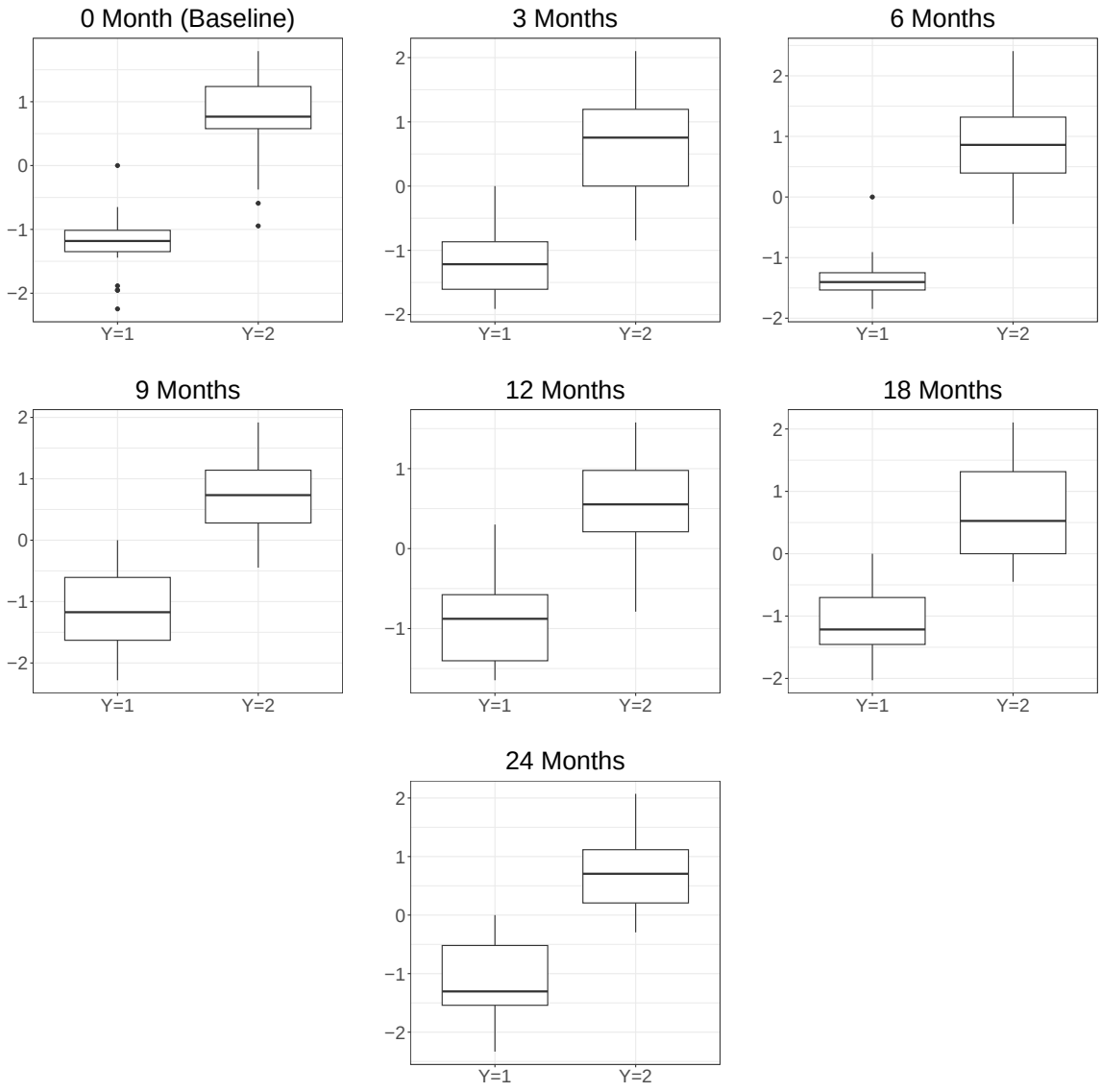

Figure S8: Side-by-side boxplots of expressions across genes at time points 3 months, 6 months, 9 months, 12 months, and 24 months.

and left-hand tasks. We randomly split the sample into a training set of size 20 and a test size of 4 and compare the performance of TLC with competing methods. For TLC, we use $\hat{\boldsymbol{r}} = (1, 1)$, which is suggested by singular values of $\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1$. Other tuning parameters of are selected based on 5-fold cross validation on the training set. Error rates based on 100 replicates are reported in Table S9. It can be seen that TLC has the best performance with the error rate being significantly smaller than others.

S3.3 Primary Biliary Cirrhosis Data

Primary Biliary Cirrhosis (PBC) is an autoimmune disease which leads to cirrhosis and liver decompensation slowly. To study this disease, the Mayo Clinic collected clinical, biochemical, serologic, and histologic measurements of 312 patients during their regular visits between 1974 and 1984 (Fleming and Harrington, 2011). This dataset is available in the R package `survival`. We apply TLC and established classification methods on the measurements of 187 patients who had complete records and study the relationship between biomarkers and the survival time. Specifically, we consider the measurement of bilirubin, albumin levels, and the prothrombin time at 4 time points: 6 month, 1 year, 2 years, and 3 years, which results in 3×4 matrix predictors. We categorize the survival time into two classes depending on whether the patient had survived within 5 years after treatment or not. For the TLC model, BIC suggests that the selected rank is $\widehat{\boldsymbol{r}} = (3, 3)$. Error rates based on 100 replicates are reported in Table S9. We can see that TLC has the smallest error rate among all methods.

S3.4 The Extended Yale Face Database B

The Extended Yale Face Database B (Georghiades et al., 2001) contains 16128 greyscale images of 28 human subjects. For each subject, there are 576 images under 9 poses and 64 illumination conditions, respectively. To facilitate the computation, we crop these images so that only the area containing a face is kept. Then, we downsize each image to 30×24 pixels. Furthermore, for each person, we stack images of the same pose with odd and even illumination condition indices separately and form tensor observations of dimension $30 \times 24 \times 32$. Finally, we obtain $28 \times 9 \times 2 = 504$ observations which are evenly from 28 classes. We randomly split the sample into a training set of size 454 and a test set of size 50. Tuning parameters of the methods are selected by 5-fold cross validation on the training set. We report the classification error rates in 100 replicates in Table S9. It is clear that TLC outperforms other methods with the lowest error rate.

Datasets	TLC-BIC	CATCH	CMDA	DGTDA	DLDA	l_1-GLM	l_1-FDA
FTF	8.00 (1.17)	13.75 (1.60)	12.00 (1.40)	29.00 (2.30)	11.75 (1.44)	13.75 (1.72)	11.75 (1.44)
PBC	14.74 (0.54)	14.50 (0.54)	20.32 (0.70)	26.37 (0.70)	17.16 (0.54)	16.08 (0.54)	15.34 (0.54)
EYF	0.48 (0.10)	8.28 (0.51)	0.48 (0.11)	3.36 (0.35)	3.16 (0.28)	2.08 (0.23)	3.16 (0.28)

Table S9: Mean and standard error of classification error rates based on 100 replicates for the Finger-Tapping fMRI (FTF), Primary Biliary Cirrhosis (PBC), and the Extended Yale Face Database B (EYF) datasets.

S3.5 Covariance separability test

We assume that the covariance matrix has a Kronecker product structure. To check if the data conforms with this structure, Aston et al. (2017) developed the computationally efficient and theoretically guaranteed nonparametric bootstrap test. We perform this test in our real data analysis after preprocessing and centering data in each class using the corresponding sample mean. Specifically, we use the function *empirical_bootstrap_test*

in the R package `covsep`. The Extended Yale Face Database B is not tested because observations in this dataset are 3-way tensors while the test was designed to work on matrix-variate data only. The null hypothesis is that the covariance admits a Kronecker product structure. The obtained p-values of the three matrix-variate datasets are presented in Table S10.

Datasets	GTC	FTF	PBC
P-values	0.197	0.003	0.518

Table S10: P-values of the Gene Time Course (GTC), Finger-Tapping fMRI (FTF), and Primary Biliary Cirrhosis (PBC) datasets.

For both the GTC and the PBC data, we cannot reject the null hypothesis, which supports the separable covariance assumption and the application of our model. For the FTF data, we have a reason to doubt the Kronecker product covariance assumption. However, considering the limited sample size of this dataset ($n = 24$) and the relatively high dimension ($p_1 \times p_2 = 20 \times 20$), we need some assumptions to facilitate efficient estimation and computation. Our model serves this purpose by drastically reducing the number of parameters in the covariance matrix. Moreover, our method leads to higher accuracy than many state-of-art methods. In this sense, we believe that it is reasonable to apply our model to improve the efficiency of classification.

S4 Connection to tensor factor analysis

In this section, we discuss the connection between the TLC model and the tensor factor analysis model. Without loss of generality, assume that $\bar{\boldsymbol{\mu}} = 0$. Then conditional on Class $Y = k$, we have that

$$\mathbf{X} = [\![\boldsymbol{\mathcal{G}}_k; \mathbf{A}_1, ..., \mathbf{A}_M]\!] + \mathbf{E}, \tag{S4.12}$$

where $\mathbf{E} \sim \mathrm{TN}(0, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_M)$. Equivalently, we can write a marginal model for $\mathbf{X}$ that absorbs Y into the mean effect. Define $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1 \times \cdots \times r_M \times K}$, where $\boldsymbol{\mathcal{G}}[:, \ldots, :, k] = \boldsymbol{\mathcal{G}}_k$, and a dummy variable $Y^{\mathrm{dm}} \in \mathbb{R}^{K \times 1}$, where $Y^{\mathrm{dm}}_{k1} = 1(Y = k)$. Then we have

$$\mathbf{X} = [\![\boldsymbol{\mathcal{G}}; \mathbf{A}_1, ..., \mathbf{A}_M, Y^{\mathrm{dm}}]\!] + \mathbf{E}. \tag{S4.13}$$

Equation (S4.13) appears similar to the tensor factor analysis model in Equation (4) of Chen et al. (2022). However, there are noticeable differences in the interpretation. For example, our core tensor $\boldsymbol{\mathcal{G}}$ is fixed, while in the tensor factor analysis model the core tensor is random; one of the factor loading in our model, Y^{dm} is random but observed, while in the tensor factor analysis model all factor loadings are constants and unobservable. Moreover, in tensor factor analysis model, the primary interest is to estimate the factor loadings, while the variance structure of $\mathbf{E}$ is nuisance. For this reason, in tensor factor analysis usually few assumptions are made on the variance of $\mathbf{E}$. In contrast, in TLC our main goal is accurate classification that relies on the discriminant coefficients $\mathbf{B}_k$ defined in (3.6). In order for $\mathbf{B}_k$ to have a Tucker low-rank decomposition, we have to assume the separable covariance on $\mathbf{E}$.

Nevertheless, results in tensor factor analysis shed light on other possible methods to estimate the TLC model, especially when the dimensions $p_m, m = 1, \ldots, M$ diverge. Straightforward calculation shows that

$$\mathrm{Var}(\mathrm{vec}(\mathbf{X})) = (\bigotimes_{m=M}^{m=1} \mathbf{A}_m)\mathrm{vec}(\boldsymbol{\mathcal{G}})\mathrm{vec}^{\mathrm{T}}(\boldsymbol{\mathcal{G}})(\bigotimes_{m=M}^{m=1} \mathbf{A}_m^{\mathrm{T}}) + \bigotimes_{m=M}^{m=1} \boldsymbol{\Sigma}_m \tag{S4.14}$$

Under the pervasive-type condition that the factors $\mathbf{A}_m$ have nonnegligible contributions

(Stock and Watson, 2002; Chen and Fan, 2021, e.g), the factors are closely connected to the top eigenvectors of Var(vec($\mathbf{X}$)) (Wang et al., 2019; Han et al., 2022; Yu et al., 2022). In this case, we can develop methods in the same line as the tensor factor analysis. However, we leave such a topic as future research, as it has many challenges beyond the TLC model. For example, it is worth investigating how the pervasive-type condition would restrict the difficulty of the classification problem. If we require $\mathbf{A}_m$ to have sufficiently large contributions, we are essentially requiring the Bayes error to have an upper bound. The interplay between the factors and the Bayes error will be a question specific to classification problems.

S5 Proofs

In this section, we present proofs for the Lemmas, Proposition, and Theorems given in the paper.

Proof of Lemma 1

Under the TDA model in (3.1), the probability density function of $\mathbf{X}$ given $Y = k$ is

$$f_k(\mathbf{X}) = (2\pi)^{-\frac{p}{2}} \left(\prod_{m=1}^{M} \det(\mathbf{\Sigma}_m)^{-\frac{q_m}{2}} \right) \exp\left\{ -\frac{1}{2} \mathrm{vec}^{\mathrm{T}}(\mathbf{X} - \boldsymbol{\mu}_k) \mathbf{\Sigma}^{-1} \mathrm{vec}(\mathbf{X} - \boldsymbol{\mu}_k) \right\}$$

where $p = \prod_{m=1}^{M} p_m$, $q_m = p/p_m$, and $\mathbf{\Sigma} = \otimes_{m=M}^{1} \mathbf{\Sigma}_m$. Thus, (3.4) is equivalent to

$$\begin{aligned}
\widehat{Y} &= \underset{k=2,...,K}{\arg\max} \left\{ \log \frac{\pi_k}{\pi_1} + \log \frac{f_k(\mathbf{X})}{f_1(\mathbf{X})} \right\} \\
&= \underset{k=2,...,K}{\arg\max} \left\{ \log \frac{\pi_k}{\pi_1} - \frac{1}{2} \mathrm{vec}^{\mathrm{T}}(\mathbf{X} - \boldsymbol{\mu}_k) \mathbf{\Sigma}^{-1} \mathrm{vec}(\mathbf{X} - \boldsymbol{\mu}_k) + \frac{1}{2} \mathrm{vec}^{\mathrm{T}}(\mathbf{X} - \boldsymbol{\mu}_1) \mathbf{\Sigma}^{-1} \mathrm{vec}(\mathbf{X} - \boldsymbol{\mu}_1) \right\} \\
&= \underset{k=2,...,K}{\arg\max} \left\{ \log \frac{\pi_k}{\pi_1} - \frac{1}{2} \mathrm{vec}^{\mathrm{T}}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) \mathbf{\Sigma}^{-1} \mathrm{vec}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_1) + \mathrm{vec}^{\mathrm{T}}(\mathbf{X}) \mathbf{\Sigma}^{-1} \mathrm{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) \right\} \\
&= \underset{k=2,...,K}{\arg\max} \left\{ \log \frac{\pi_k}{\pi_1} - \frac{1}{2} \langle \mathbf{B}_k, \boldsymbol{\mu}_k + \boldsymbol{\mu}_1 \rangle + \langle \mathbf{B}_k, \mathbf{X} \rangle \right\}
\end{aligned}$$

where $\mathbf{B}_k = [\![\boldsymbol{\mu}_k - \boldsymbol{\mu}_1; \mathbf{\Sigma}_1^{-1}, ..., \mathbf{\Sigma}_M^{-1}]\!], k = 2, ..., K$.

Proof of Proposition 1

To prove Proposition 1, we need the following result.

Proposition S2 (Pan et al. (2019a), Lemma 2)**.** If $\mathbf{W} \sim TN(\mathbf{0}, \mathbf{\Omega}_1, ..., \mathbf{\Omega}_M)$, then $\mathrm{E}[\mathbf{W}_{(j)} \mathbf{W}_{(j)}^{\mathrm{T}}] = \mathbf{\Omega}_j \cdot \prod_{l \neq j} \mathrm{tr}(\mathbf{\Omega}_l)$, where $\mathbf{W}_{(j)}$ is the mode-j matricization of $\mathbf{W}$.

Now, we proceed to the proof of Proposition 1.

For any $i \in \{1, ..., n\}$, $(\mathbf{X}_i - \boldsymbol{\mu}_{Y_i}) \overset{i.i.d.}{\sim} TN(\mathbf{0}, \mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_M)$. Then,

$$\begin{aligned}
\mathrm{E}[\mathbf{S}_m] &= \frac{1}{(n-K)p_{-m}} \mathrm{E} \left[\sum_{i=1}^{n} (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i})_{(m)} (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i})_{(m)}^{\mathrm{T}} \right] \\
&= \frac{1}{(n-K)p_{-m}} \sum_{k=1}^{K} \sum_{l, Y_l = k} \mathrm{E} \left[(\mathbf{X}_l - \boldsymbol{\mu}_k + \boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k)_{(m)} (\mathbf{X}_l - \boldsymbol{\mu}_k + \boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k)_{(m)}^{\mathrm{T}} \right].
\end{aligned}$$

In particular,

$$\begin{aligned}
\mathrm{E} \left[(\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k)_{(m)} (\mathbf{X}_l - \boldsymbol{\mu}_k)_{(m)}^{\mathrm{T}} \right] &= \mathrm{E} \left[(\boldsymbol{\mu}_k - \frac{1}{n_k} \sum_{j, Y_j = k} \mathbf{X}_j)_{(m)} (\mathbf{X}_l - \boldsymbol{\mu}_k)_{(m)}^{\mathrm{T}} \right] \\
&= -\frac{1}{n_k} \mathrm{E} \left[(\mathbf{X}_l - \boldsymbol{\mu}_k)_{(m)} (\mathbf{X}_l - \boldsymbol{\mu}_k)_{(m)}^{\mathrm{T}} \right],
\end{aligned}$$

$$\mathrm{E}\left[(\boldsymbol{\mu}_k-\widehat{\boldsymbol{\mu}}_k)_{(m)}(\boldsymbol{\mu}_k-\widehat{\boldsymbol{\mu}}_k)_{(m)}^{\mathrm{T}}\right]=\mathrm{E}\left[(\boldsymbol{\mu}_k-\frac{1}{n_k}\sum_{l,Y_l=k}\mathbf{X}_l)_{(m)}(\boldsymbol{\mu}_k-\frac{1}{n_k}\sum_{l,Y_l=k}\mathbf{X}_l)_{(m)}^{\mathrm{T}}\right]$$

$$=\frac{1}{n_k^2}\sum_{l,Y_l=k}\mathrm{E}\left[(\mathbf{X}_l-\boldsymbol{\mu}_k)_{(m)}(\mathbf{X}_l-\boldsymbol{\mu}_k)_{(m)}^{\mathrm{T}}\right]=\frac{1}{n_k}\mathrm{E}\left[(\mathbf{X}_l-\boldsymbol{\mu}_k)_{(m)}(\mathbf{X}_l-\boldsymbol{\mu}_k)_{(m)}^{\mathrm{T}}\right].$$

According to Proposition S.1,

$$\mathrm{E}[\mathbf{S}_m]=\frac{1}{(n-K)p_{-m}}\sum_{k=1}^{K}(n_k-1)\mathrm{E}\left[(\mathbf{X}_l-\boldsymbol{\mu}_k)_{(m)}(\mathbf{X}_l-\boldsymbol{\mu}_k)_{(m)}^{\mathrm{T}}\right]$$

$$=\frac{1}{p_{-m}}\boldsymbol{\Sigma}_m\cdot\prod_{h\neq m}\mathrm{tr}(\boldsymbol{\Sigma}_h)\propto\boldsymbol{\Sigma}_m.$$

The second part of the conclusion can be proved in a similar fashion as the Lemma 3 in Pan et al. (2019a) and hence is omitted.

Proof of Lemma 2

For observations $\{Y_i,\mathbf{X}_i\}_{i=1}^n$ that are independently and identically generated under the TLC model, the log-likelihood function is

$$l(\boldsymbol{\theta}|\{Y_i,\mathbf{X}_i\}_{i=1}^n)=\sum_{i=1}^{n}\log\left[f(\mathbf{X}_i|Y_i)f(Y_i=k)\right]$$

$$=-\sum_{i=1}^{n}\sum_{m=1}^{M}\frac{q_m}{2}\log|\boldsymbol{\Sigma}_m|-\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\left\{\boldsymbol{\Sigma}_m^{-1}\left(\mathbf{X}_i-\boldsymbol{\mu}_{Y_i}\right)_{(m)}\left(\otimes_{m'\neq m}\boldsymbol{\Sigma}_{m'}^{-1}\right)\left(\mathbf{X}_i-\boldsymbol{\mu}_{Y_i}\right)_{(m)}^{\mathrm{T}}\right\}+c \tag{S5.15}$$

where $c=\sum_{i=1}^n\log\pi_{Y_i}-(np\log 2\pi)/2$. Denote

$$h(\boldsymbol{\theta}|\mathbf{X}_i)=\langle[\![\mathbf{X}_i-\boldsymbol{\mu}_{Y_i};\boldsymbol{\Sigma}_1^{-1},...\boldsymbol{\Sigma}_M^{-1}]\!],\mathbf{X}_i-\boldsymbol{\mu}_{Y_i}\rangle=\mathrm{tr}\left\{\boldsymbol{\Sigma}_m^{-1}\left(\mathbf{X}_i-\boldsymbol{\mu}_{Y_i}\right)_{(m)}\left(\otimes_{m'\neq m}\boldsymbol{\Sigma}_{m'}^{-1}\right)\left(\mathbf{X}_i-\boldsymbol{\mu}_{Y_i}\right)_{(m)}^{\mathrm{T}}\right\}. \tag{S5.16}$$

Then (S5.15) reduces to

$$l(\boldsymbol{\theta})=\sum_{i=1}^{n}\left(\log\pi_{Y_i}-\sum_{m=1}^{M}\frac{q_m}{2}\log|\boldsymbol{\Sigma}_m|-\frac{1}{2}h(\boldsymbol{\theta}|\mathbf{X}_i)\right)+c. \tag{S5.17}$$

Since

$$\frac{\partial h}{\partial \boldsymbol{\Sigma}_m} = -\boldsymbol{\Sigma}_m^{-1} \left(\mathbf{X}_i - \boldsymbol{\mu}_{Y_i}\right)_{(m)} \left(\otimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1}\right) \left(\mathbf{X}_i - \boldsymbol{\mu}_{Y_i}\right)_{(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1}, \tag{S5.18}$$

it follows that

$$\frac{\partial l}{\partial \boldsymbol{\Sigma}_m} = \sum_{i=1}^{n} \left[-\frac{q_m}{2} \boldsymbol{\Sigma}_m^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_m^{-1} \left(\mathbf{X}_i - \boldsymbol{\mu}_{Y_i}\right)_{(m)} \left(\otimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1}\right) \left(\mathbf{X}_i - \boldsymbol{\mu}_{Y_i}\right)_{(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} \right]. \tag{S5.19}$$

Note that $l(\boldsymbol{\theta})$ is concave. With other parameters being fixed, setting $\partial l / \partial \boldsymbol{\Sigma}_m$ to 0 gives

$$\widehat{\boldsymbol{\Sigma}}_m = \frac{1}{n q_m} \sum_{i=1}^{n} \left(\mathbf{X}_i - \boldsymbol{\mu}_{Y_i}\right)_{(m)} \left(\otimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1}\right) \left(\mathbf{X}_i - \boldsymbol{\mu}_{Y_i}\right)_{(m)}^{\mathrm{T}}, \quad m = 1, ..., M. \tag{S5.20}$$

Without loss of generality, we assume that $\bar{\boldsymbol{\mu}} = 0$ and hence the Tucker low-rank assumption reduces to $\boldsymbol{\mu}_k = [\![\boldsymbol{\mathcal{G}}_k; \mathbf{A}_1, ..., \mathbf{A}_M]\!]$, $k = 1, ..., K$. Then, $h(\boldsymbol{\theta}|\mathbf{X}_i)$ has the form

$$\begin{aligned} h(\boldsymbol{\theta}|\mathbf{X}_i) =& \mathrm{tr}\left\{ \left(\otimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1}\right) \left(\mathbf{X}_i - \boldsymbol{\mu}_{Y_i}\right)_{(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} \left(\mathbf{X}_i - \boldsymbol{\mu}_{Y_i}\right)_{(m)} \right\} \\ =& \mathrm{tr}\left\{ \left(\otimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1}\right) \mathbf{X}_{i(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} \mathbf{X}_{i(m)} \right\} - 2 \underbrace{\mathrm{tr}\left\{ \left(\otimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1}\right) \mathbf{X}_{i(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\mu}_{Y_i(m)} \right\}}_{\mathrm{I}_1} \\ & + \underbrace{\mathrm{tr}\left\{ \left(\otimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1}\right) \boldsymbol{\mu}_{Y_i(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\mu}_{Y_i(m)} \right\}}_{\mathrm{I}_2}, \end{aligned} \tag{S5.21}$$

where

$$\begin{aligned} \mathrm{I}_1 &= \mathrm{tr}\left\{ \left(\otimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1}\right) \mathbf{X}_{i(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} \mathbf{A}_m \boldsymbol{\mathcal{G}}_{Y_i(m)} \left(\otimes_{m' \neq m} \mathbf{A}_{m'}\right)^{\mathrm{T}} \right\} \\ &= \mathrm{tr}\left\{ \boldsymbol{\mathcal{G}}_{Y_i(m)} \left(\otimes_{m' \neq m} \mathbf{A}_{m'}^{\mathrm{T}} \boldsymbol{\Sigma}_{m'}^{-1}\right) \mathbf{X}_{i(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} \mathbf{A}_m \right\} \\ &= \mathrm{tr}\left\{ \left(\otimes_{m' \neq m} \mathbf{A}_{m'}^{\mathrm{T}} \boldsymbol{\Sigma}_{m'}^{-1}\right) \mathbf{X}_{i(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} \mathbf{A}_m \boldsymbol{\mathcal{G}}_{Y_i(m)} \right\}, \end{aligned} \tag{S5.22}$$

and

$$\begin{aligned}
\mathrm{I}_2 &= \mathrm{tr}\left\{\left(\otimes_{m'\neq m}\boldsymbol{\Sigma}_{m'}^{-1}\right)\left(\otimes_{m'\neq m}\mathbf{A}_{m'}\right)\boldsymbol{\mathcal{G}}_{Y_i(m)}^{\mathrm{T}}\mathbf{A}_m^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\mathbf{A}_m\boldsymbol{\mathcal{G}}_{Y_i(m)}\left(\otimes_{m'\neq m}\mathbf{A}_{m'}\right)^{\mathrm{T}}\right\} \\
&= \mathrm{tr}\left\{\boldsymbol{\mathcal{G}}_{Y_i(m)}\left(\otimes_{m'\neq m}\mathbf{A}_{m'}^{\mathrm{T}}\boldsymbol{\Sigma}_{m'}^{-1}\mathbf{A}_{m'}\right)\boldsymbol{\mathcal{G}}_{Y_i(m)}^{\mathrm{T}}\mathbf{A}_m^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\mathbf{A}_m\right\} \qquad \text{(S5.23)} \\
&= \mathrm{tr}\left\{\left(\otimes_{m'\neq m}\mathbf{A}_{m'}^{\mathrm{T}}\boldsymbol{\Sigma}_{m'}^{-1}\mathbf{A}_{m'}\right)\boldsymbol{\mathcal{G}}_{Y_i(m)}^{\mathrm{T}}\mathbf{A}_m^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\mathbf{A}_m\boldsymbol{\mathcal{G}}_{Y_i(m)}\right\}.
\end{aligned}$$

It follows that

$$\frac{\partial l}{\partial \boldsymbol{\mathcal{G}}_{k(m)}} = \sum_{Y_i=k}\left[\left(\otimes_{m'\neq m}\mathbf{A}_{m'}^{\mathrm{T}}\boldsymbol{\Sigma}_{m'}^{-1}\right)\mathbf{X}_{i(m)}^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\mathbf{A}_m - \left(\otimes_{m'\neq m}\mathbf{A}_{m'}^{\mathrm{T}}\boldsymbol{\Sigma}_{m'}^{-1}\mathbf{A}_{m'}\right)\boldsymbol{\mathcal{G}}_{k(m)}^{\mathrm{T}}\mathbf{A}_m^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\mathbf{A}_m\right]. \qquad \text{(S5.24)}$$

With other parameters being fixed, setting $\partial l/\partial \boldsymbol{\mathcal{G}}_{k(m)}$ to 0 gives

$$\widehat{\boldsymbol{\mathcal{G}}}_{k(m)} = \frac{1}{n_k}\sum_{Y_i=k}\left\{\left(\left(\mathbf{A}_m^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\mathbf{A}_m\right)^{-1}\mathbf{A}_m^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\right)\mathbf{X}_{i(m)}\left[\otimes_{m'\neq m}\left(\left(\mathbf{A}_{m'}^{\mathrm{T}}\boldsymbol{\Sigma}_{m'}^{-1}\mathbf{A}_{m'}\right)^{-1}\mathbf{A}_{m'}^{\mathrm{T}}\boldsymbol{\Sigma}_{m'}^{-1}\right)^{\mathrm{T}}\right]\right\},$$

or equivalently,

$$\widehat{\boldsymbol{\mathcal{G}}}_k = \frac{1}{n_k}\sum_{Y_i=k}[\![\mathbf{X}_i; (\mathbf{A}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\mathbf{A}_1)^{-1}\mathbf{A}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}, ..., (\mathbf{A}_M^{\mathrm{T}}\boldsymbol{\Sigma}_M^{-1}\mathbf{A}_M)^{-1}\mathbf{A}_M^{\mathrm{T}}\boldsymbol{\Sigma}_M^{-1}]\!], \quad k = 1, ..., K. \qquad \text{(S5.25)}$$

When $\boldsymbol{\Sigma}_m = \mathbf{I}_{p_m}$ for all $m = 1, ..., M$, the maximizer in (S5.25) reduces to

$$\widehat{\boldsymbol{\mathcal{G}}}_k = \frac{1}{n_k}\sum_{Y_i=k}[\![\mathbf{X}_i; \mathbf{A}_1^{\mathrm{T}}, ..., \mathbf{A}_M^{\mathrm{T}}]\!].$$

To maximize (S5.15) over a factor matrix, we need to solve the following nonconvex optimization problem,

$$\widehat{\mathbf{A}}_m = \underset{\mathbf{A}_m^{\mathrm{T}}\mathbf{A}_m=\mathbf{I}_{r_m}}{\arg\max}\ \mathrm{tr}\left(\mathbf{H}_{1m}\mathbf{A}_m\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{H}_{2m}\mathbf{A}_m^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\mathbf{A}_m\right), \qquad \text{(S5.26)}$$

where

$$\mathbf{H}_{1m} = \sum_{i=1}^{n} \boldsymbol{\mathcal{G}}_{Y_i(m)} \left(\otimes_{m' \neq m} \mathbf{A}_{m'}^{\mathrm{T}} \boldsymbol{\Sigma}_{m'}^{-1}\right) \mathbf{X}_{i(m)}^{\mathrm{T}} \boldsymbol{\Sigma}_{m}^{-1},$$
$$\mathbf{H}_{2m} = \sum_{i=1}^{n} \boldsymbol{\mathcal{G}}_{Y_i(m)} \left(\otimes_{m' \neq m} \mathbf{A}_{m'}^{\mathrm{T}} \boldsymbol{\Sigma}_{m'}^{-1} \mathbf{A}_{m'}\right) \boldsymbol{\mathcal{G}}_{Y_i(m)}^{\mathrm{T}}, \tag{S5.27}$$

which calls for delicately designed algorithms that optimize over Stiefel manifolds.

Proof of Theorem 1

For the first model which applies LDA on vectorized tensor data, the sample covariance $\widehat{\boldsymbol{\Sigma}}$ and the discriminant coefficient estimator $\widehat{\boldsymbol{\beta}}_k^{\mathrm{LDA}} = \widehat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$ are both MLEs. Naturally, we have $\sqrt{n}(\widehat{\mathbf{h}}_{\mathrm{LDA}} - \mathbf{h}^*) \to N(0, \mathbf{W}_{\boldsymbol{\beta}})$. To derive $\mathbf{W}_{\boldsymbol{\beta}}$, we first consider the asymptotic covariance matrix of vector $\widehat{\boldsymbol{\eta}}_1 = (\mathrm{vec}(\overline{\mathbf{X}}_1)^{\mathrm{T}}, ..., \mathrm{vec}(\overline{\mathbf{X}}_K)^{\mathrm{T}}, \mathrm{vech}(\widehat{\boldsymbol{\Sigma}})^{\mathrm{T}})^{\mathrm{T}}$. Let $\mathbf{J}_1^{-1} = \mathrm{avar}(\widehat{\boldsymbol{\eta}}_1)$, then

$$\mathbf{J}_1^{-1} = \mathrm{diag}\left\{\frac{1}{\pi_1}\boldsymbol{\Sigma}, ..., \frac{1}{\pi_K}\boldsymbol{\Sigma}, \left(\frac{1}{2}\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\right)^{-1}\right\},$$

where $p = \prod_{m=1}^{M} p_m$, $\mathbf{E}_p \in \mathbb{R}^{p^2 \times p(p+1)/2}$ with $\mathrm{vec}(\boldsymbol{\Sigma}) = \mathbf{E}_p \mathrm{vech}(\boldsymbol{\Sigma})$.

Next, consider the asymptotic covariance matrix of vector $\widehat{\boldsymbol{\eta}}_2 = (\mathrm{vec}(\overline{\mathbf{X}}_2 - \overline{\mathbf{X}}_1)^{\mathrm{T}}, ..., \mathrm{vec}(\overline{\mathbf{X}}_K - \overline{\mathbf{X}}_1)^{\mathrm{T}}, \mathrm{vech}(\widehat{\boldsymbol{\Sigma}})^{\mathrm{T}})^{\mathrm{T}}$. $\widehat{\boldsymbol{\eta}}_2$ is a linear transformation of $\widehat{\boldsymbol{\eta}}_1$ with $\widehat{\boldsymbol{\eta}}_2 = \mathbf{L}_1 \widehat{\boldsymbol{\eta}}_1$ where $\mathbf{L}_1 = \mathrm{diag}\{(-\mathbf{1}_{K-1}, \mathbf{I}_{K-1}) \otimes \mathbf{I}_p, \mathbf{I}_{p(p+1)/2}\}$. Let $\mathbf{J}_2^{-1} = \mathrm{avar}(\widehat{\boldsymbol{\eta}}_2)$. We have

$$\mathbf{J}_2^{-1} = \mathbf{L}_1 \mathbf{J}_1^{-1} \mathbf{L}_1^{\mathrm{T}} = \mathrm{diag}\left\{\mathbf{A} \otimes \boldsymbol{\Sigma}, \left(\frac{1}{2}\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\right)^{-1}\right\},$$

where $\mathbf{A} \in \mathbb{R}^{(K-1)\times(K-1)}$ is a symmetric matrix with diagonal elements being $(1/\pi_k + 1/\pi_1)$ and off-diagonal elements being $1/\pi_1$. On the other hand, $\boldsymbol{\eta}_2 = (\mathrm{vec}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}, ..., \mathrm{vec}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)^{\mathrm{T}}, \mathrm{vech}(\boldsymbol{\Sigma})^{\mathrm{T}})^{\mathrm{T}}$ is a function of $\mathbf{h}_{\mathrm{LDA}}$. Since

$$\mathrm{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) = \boldsymbol{\Sigma}\boldsymbol{\beta}_k = \mathrm{vec}(\mathbf{I}_p \boldsymbol{\Sigma} \boldsymbol{\beta}_k) = (\boldsymbol{\beta}_k^{\mathrm{T}} \otimes \mathbf{I}_p)\mathrm{vec}(\boldsymbol{\Sigma}) = (\boldsymbol{\beta}_k^{\mathrm{T}} \otimes \mathbf{I}_p)\mathbf{E}_p \mathrm{vech}(\boldsymbol{\Sigma}),$$

we have

$$\frac{\partial \text{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)}{\partial \boldsymbol{\beta}_k} = \boldsymbol{\Sigma}, \quad \frac{\partial \text{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)}{\partial \text{vech}(\boldsymbol{\Sigma})} = (\boldsymbol{\beta}_k^{\mathrm{T}} \otimes \mathbf{I}_p)\mathbf{E}_p,$$

and hence

$$\mathbf{G} = \frac{\partial \boldsymbol{\eta}_2}{\partial \mathbf{h}_{\text{LDA}}} = \begin{pmatrix} \mathbf{I}_{K-1} \otimes \boldsymbol{\Sigma} & (\boldsymbol{\beta}^{\mathrm{T}} \otimes \mathbf{I}_p)\mathbf{E}_p \\ \mathbf{0} & \mathbf{I}_{p(p+1)/2} \end{pmatrix},$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_K) \in \mathbb{R}^{p\times(K-1)}$. With the multivariate delta method, we have $\mathbf{J}_2^{-1} = \mathbf{G}^{\mathrm{T}}\mathbf{W}_{\boldsymbol{\beta}}\mathbf{G}$. Thus,

$$\mathbf{W}_{\boldsymbol{\beta}} = \begin{pmatrix} \mathbf{A} \otimes \boldsymbol{\Sigma}^{-1} & -(\mathbf{A}\boldsymbol{\beta}^{\mathrm{T}} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \\ -\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\beta}\mathbf{A} \otimes \boldsymbol{\Sigma}^{-1}) & \mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\beta}\mathbf{A}\boldsymbol{\beta}^{\mathrm{T}} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p + \left(\frac{1}{2}\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\right)^{-1} \end{pmatrix}. \tag{S5.28}$$

and the Fisher information matrix for $\mathbf{h}_{\text{LDA}}$ is

$$\mathbf{J}_{\mathbf{h}} = \mathbf{W}_{\boldsymbol{\beta}}^{-1} = \mathbf{G}\mathbf{J}_2\mathbf{G}^{\mathrm{T}}$$

$$= \begin{pmatrix} \mathbf{A}^{-1} \otimes \boldsymbol{\Sigma} + \frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}} \otimes \mathbf{I}_p)\mathbf{E}_p\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\beta} \otimes \mathbf{I}_p) & \frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}} \otimes \mathbf{I}_p)\mathbf{E}_p\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \\ \frac{1}{2}\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\beta} \otimes \mathbf{I}_p) & \frac{1}{2}\mathbf{E}_p^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \end{pmatrix}$$

Denote $\mathbf{H} = \partial \mathbf{h}(\boldsymbol{\psi}_1)/\partial \boldsymbol{\psi}_1$ and $\mathbf{K} = \partial \mathbf{h}(\boldsymbol{\psi}_2)/\partial \boldsymbol{\psi}_2$. Following the Proposition 4.1 in Shapiro (1986), we have

$$\sqrt{n}(\mathbf{h}(\widehat{\boldsymbol{\psi}}_1) - \mathbf{h}^*) \to N(\mathbf{0}, \mathbf{U}_{\boldsymbol{\beta}})$$

$$\sqrt{n}(\mathbf{h}(\widehat{\boldsymbol{\psi}}_2) - \mathbf{h}^*) \to N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}})$$

where

$$\mathbf{U}_{\boldsymbol{\beta}} = \mathbf{H}(\mathbf{H}^{\mathrm{T}}\mathbf{J}_h\mathbf{H})^{\dagger}\mathbf{H}^{\mathrm{T}}, \tag{S5.29}$$

$$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{K}(\mathbf{K}^{\mathrm{T}}\mathbf{J}_h\mathbf{K})^{\dagger}\mathbf{K}^{\mathrm{T}}, \tag{S5.30}$$

with

$$\mathbf{H} = \frac{\partial \mathbf{h}(\boldsymbol{\psi}_1)}{\partial \boldsymbol{\psi}} = \begin{pmatrix} \mathbf{I}_{(K-1)p} & \mathbf{0} \\ \mathbf{0} & \left(\frac{\partial \text{vech}(\boldsymbol{\Sigma})}{\partial \text{vech}(\boldsymbol{\Sigma}_1)}, ..., \frac{\partial \text{vech}(\boldsymbol{\Sigma})}{\partial \text{vech}(\boldsymbol{\Sigma}_m)} \right) \end{pmatrix}$$

and $\mathbf{K} = \partial \mathbf{h}(\boldsymbol{\psi}_2)/\partial \boldsymbol{\psi}_2$ having its elements being

$$\frac{\partial \boldsymbol{\beta}_k}{\partial \text{vec}(\boldsymbol{\Phi}_k)} = \otimes_{m=M}^{1} \mathbf{D}_m, \quad k = 2, ..., K \tag{S5.31}$$

$$\frac{\partial \boldsymbol{\beta}_k}{\partial \text{vec}(\mathbf{D}_j)} = \mathbf{T}_j \left\{ \left((\otimes_{m=M, m\neq j}^{1} \mathbf{D}_m) \boldsymbol{\Phi}_{k(j)}^{\mathrm{T}} \right) \otimes \mathbf{I}_{p_j} \right\} \tag{S5.32}$$

$$\frac{\partial \text{vech}(\boldsymbol{\Sigma})}{\partial \text{vech}(\boldsymbol{\Sigma}_m)} = \mathbf{C}_p \frac{\partial \text{vec}(\boldsymbol{\Sigma})}{\partial \text{vec}(\boldsymbol{\Sigma}_m)} \mathbf{E}_{p_m}, \quad m = 1, ..., M \tag{S5.33}$$

where $\mathbf{T}_j \in \mathbb{R}^{p\times p}$ is the permutation matrix such that $\text{vec}(\mathbf{B}) = \mathbf{T}_j \text{vec}(\mathbf{B}_{(j)})$ for tensor $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$, and $\mathbf{C}_p \in \mathbb{R}^{p(p+1)/2 \times p^2}$ is the contraction matrix such that $\text{vech}(\boldsymbol{\Sigma}) = \mathbf{C}_p \text{vec}(\boldsymbol{\Sigma})$. For (S5.33), we could obtain the explicit form derivatives of $m = 1$ and M using formulas from Fackler (2005). However, when $2 \leq m \leq (M-1)$, we only have elementwise derivatives.

For $\mathbf{U}_{\boldsymbol{\beta}}$ and $\mathbf{V}_{\boldsymbol{\beta}}$, we have

$$\mathbf{U}_{\boldsymbol{\beta}} = \mathbf{H}(\mathbf{H}^{\mathrm{T}} \mathbf{J}_h \mathbf{H})^{\dagger} \mathbf{H}^{\mathrm{T}} = \mathbf{J}_h^{-1/2} \mathbf{J}_h^{1/2} \mathbf{H} (\mathbf{H}^{\mathrm{T}} \mathbf{J}_h^{1/2} \mathbf{J}_h^{1/2} \mathbf{H})^{\dagger} \mathbf{H}^{\mathrm{T}} \mathbf{J}_h^{1/2} \mathbf{J}_h^{-1/2} = \mathbf{J}_h^{-1/2} \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}} \mathbf{J}_h^{-1/2},$$

$$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{K}(\mathbf{K}^{\mathrm{T}} \mathbf{J}_h \mathbf{K})^{\dagger} \mathbf{K}^{\mathrm{T}} = \mathbf{J}_h^{-1/2} \mathbf{J}_h^{1/2} \mathbf{K} (\mathbf{K}^{\mathrm{T}} \mathbf{J}_h^{1/2} \mathbf{J}_h^{1/2} \mathbf{K})^{\dagger} \mathbf{K}^{\mathrm{T}} \mathbf{J}_h^{1/2} \mathbf{J}_h^{-1/2} = \mathbf{J}_h^{-1/2} \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}} \mathbf{J}_h^{-1/2}.$$

Also, notice that $\partial \mathbf{h}/\partial \boldsymbol{\psi}_2 = \partial \mathbf{h}/\partial \boldsymbol{\psi}_1 \cdot \partial \boldsymbol{\psi}_1/\partial \boldsymbol{\psi}_2$, i.e., $\mathbf{K} = \mathbf{H} \cdot \partial \boldsymbol{\psi}_1/\partial \boldsymbol{\psi}_2$ according to the chain rule. Thus, we have $\text{span}(\mathbf{K}) \subseteq \text{span}(\mathbf{H})$, $\text{span}(\mathbf{J}_h^{1/2}\mathbf{K}) \subseteq \text{span}(\mathbf{J}_h^{1/2}\mathbf{H})$, and $\mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}} = \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}} \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}} = \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}} \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}}$. Therefore,

$$\begin{aligned} \mathbf{U}_{\boldsymbol{\beta}} - \mathbf{V}_{\boldsymbol{\beta}} &= \mathbf{J}_h^{-1/2} (\mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}} - \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}}) \mathbf{J}_h^{-1/2} \\ &= \mathbf{J}_h^{-1/2} (\mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}} - \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}} \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}}) \mathbf{J}_h^{-1/2} \\ &= \mathbf{J}_h^{-1/2} (\mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}} \mathbf{Q}_{\mathbf{J}_h^{1/2}\mathbf{K}}) \mathbf{J}_h^{-1/2} \geq 0 \end{aligned}$$

and

$$\begin{aligned}\mathbf{W}_{\boldsymbol{\beta}} - \mathbf{U}_{\boldsymbol{\beta}} &= \mathbf{J}_h^{-1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}})\mathbf{J}_h^{-1/2} \\ &= \mathbf{J}_h^{-1/2}\mathbf{Q}_{\mathbf{J}_h^{1/2}\mathbf{H}}\mathbf{J}_h^{-1/2} \geq 0\end{aligned}$$

where $\mathbf{Q}_{\mathbf{J}_h^{1/2}\mathbf{K}}$ and $\mathbf{Q}_{\mathbf{J}_h^{1/2}\mathbf{H}}$ are projection matrices onto the orthogonal compliment of span($\mathbf{J}_h^{1/2}\mathbf{K}$) and span($\mathbf{J}_h^{1/2}\mathbf{H}$), respectively.

S5.1 Proof of Theorem 2

To proceed with the proof of Theorem 2, we introduce some notations in advance. For a matrix $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$, let $\sigma_1(\mathbf{X}) \geq \cdots \geq \sigma_{p_1 \wedge p_2}(\mathbf{X}) \geq 0$ be its singular values in non-increasing order and $\text{SVD}_r(\mathbf{X})$ be the first r left singular vectors of $\mathbf{X}$. The max norm, spectral norm and Frobnieus norm of matrix $\mathbf{X}$ are defined to be $\|\mathbf{X}\|_{\max} = \max_{i,j} |X_{ij}|$, $\|\mathbf{X}\| = \max_{\mathbf{v} \in \mathbb{R}^{p_2}} \frac{\|\mathbf{X}\mathbf{v}\|_2}{\|\mathbf{v}\|_2} = \sigma_1(\mathbf{X})$ and $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{p_1}\sum_{j=1}^{p_2} X_{ij}^2} = \sqrt{\sum_{i=1}^{p_1 \wedge p_2} \sigma_i^2(\mathbf{X})}$, respectively. Denote the collection of $p \times r$ orthogonal matrices as $\mathbb{O}_{p \times r} = \{\mathbf{U} \in \mathbb{R}^{p \times r} : \mathbf{U}^T\mathbf{U} = \mathbf{I}_r\}$. For $\mathbf{U} \in \mathbb{O}_{p \times r}$, $\mathbf{U}_\perp \in \mathbb{O}_{p \times (p-r)}$ is its orthogonal complement such that $[\mathbf{U}, \mathbf{U}_\perp] \in \mathbb{O}_{p \times p}$. For any $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathbb{O}_{p \times r}$, we measure the distance between the two subspaces span($\mathbf{U}$) and span($\widetilde{\mathbf{U}}$) by $\sin\Theta(\mathbf{U}, \widetilde{\mathbf{U}}) = \text{diag}\,(\sin\theta_1, ..., \sin\theta_r)$ where $\theta_i = \arccos\sigma_i$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$ being singular values of $\mathbf{U}^T\widetilde{\mathbf{U}}$. To avoid potential confusion, we use $\mathcal{M}_m(\boldsymbol{\mu})$ instead of $\boldsymbol{\mu}_{(m)}$ to denote the mode-m matricization of tensor $\boldsymbol{\mu}$ in the proof. Denote $\widehat{\mathbf{A}}_m^{(0)} = \text{SVD}_{r_m}(\mathcal{M}_m(\widehat{\boldsymbol{\mu}}))$, $\eta_m = \sigma_{r_m}(\mathcal{M}_m(\widehat{\boldsymbol{\mu}}))$ as the leading r_m left singular vectors and the r_m-th singular value of $\mathcal{M}_m(\widehat{\boldsymbol{\mu}})$. Then, $\widehat{\mathbf{A}}_m^{(0)}$ is the initial factor matrix estimate. We furthre denote $\widehat{\mathbf{A}}_m$ as the mode-m factor matrix estimate after achieving convergence under Algorithm S5.

Proof of part (a)

We prove (a) in 5 steps. In step 1, we give a reformulation for the multiclass problem to facilitate the proof. In Step 2, we establish the initialization error bound for $\widehat{\mathbf{A}}_m^{(0)}$ using the perturbation theory and matrix concentration inequalities. In Step 3, we establish the error bound for the factor matrix estimate $\widehat{\mathbf{A}}_m$ obtained under Algorithm S5. In Step 4, we bound the error for the sparse estimate of $\mathbf{D}_m$ based on corollaries of results in Min and Mai (2022). In the last step, we construct and derive the error bound for $\widehat{\mathbf{B}}$ using $\widehat{\boldsymbol{\mu}}$, $\{\widehat{\mathbf{A}}_m\}_{m=1}^3$, $\{\widehat{\mathbf{D}}_m\}_{m=1}^3$, and corresponding error bounds.

Step 1 Under TLC, we could rewrite $\mathbf{X}_i$ as

$$\mathbf{X}_i|(Y=k) = \boldsymbol{\mu}_k + \mathbf{Z}_i, \quad \mathbf{Z}_i \overset{iid}{\sim} TN(\mathbf{0}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3). \tag{S5.34}$$

According to the estimation process in Section 4,

$$\widehat{\boldsymbol{\mu}}_{[:,:,:,k-1]} = \frac{1}{n_k}\sum_{Y_i=k}\mathbf{X}_i - \frac{1}{n_1}\sum_{Y_i=1}\mathbf{X}_i = \boldsymbol{\mu}_k - \boldsymbol{\mu}_1 + \mathbf{E}_{[:,:,:,k-1]}, \quad k=2,...,K, \tag{S5.35}$$

where $\mathbf{E} \in \mathbb{R}^{p_1\times p_2\times p_3\times(K-1)}$ and $\mathbf{E}_{[:,:,:,k-1]} = \frac{1}{n_k}\sum_{Y_i=k}\mathbf{Z}_i - \frac{1}{n_1}\sum_{Y_i=1}\mathbf{Z}_i$. Note that

$$\begin{aligned}\mathcal{M}_{M+1}(\mathbf{E}) &= \begin{bmatrix} \mathrm{vec}^T\left(\frac{1}{n_2}\sum_{Y_i=2}\mathbf{Z}_i - \frac{1}{n_1}\sum_{Y_i=1}\mathbf{Z}_i\right) \\ \vdots \\ \mathrm{vec}^T\left(\frac{1}{n_K}\sum_{Y_i=K}\mathbf{Z}_i - \frac{1}{n_1}\sum_{Y_i=1}\mathbf{Z}_i\right)\end{bmatrix} = \begin{bmatrix} \mathrm{vec}^T\left(\frac{1}{n_2}\sum_{Y_i=2}\mathbf{Z}_i\right) - \mathrm{vec}^T\left(\frac{1}{n_1}\sum_{Y_i=1}\mathbf{Z}_i\right) \\ \vdots \\ \mathrm{vec}^T\left(\frac{1}{n_K}\sum_{Y_i=K}\mathbf{Z}_i\right) - \mathrm{vec}^T\left(\frac{1}{n_1}\sum_{Y_i=1}\mathbf{Z}_i\right)\end{bmatrix} \\ &= \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -1 & 0 & 0 & \cdots & 1 \end{bmatrix}_{(K-1)\times K} \begin{bmatrix} \mathrm{vec}^T\left(\frac{1}{n_1}\sum_{Y_i=1}\mathbf{Z}_i\right) \\ \mathrm{vec}^T\left(\frac{1}{n_2}\sum_{Y_i=2}\mathbf{Z}_i\right) \\ \vdots \\ \mathrm{vec}^T\left(\frac{1}{n_K}\sum_{Y_i=K}\mathbf{Z}_i\right)\end{bmatrix}_{K\times p_1p_2p_3}. \end{aligned} \tag{S5.36}$$

Denote $\widetilde{\mathbf{z}}_k = \text{vec}\left(\frac{1}{n_k}\sum_{Y_i=k}\mathbf{Z}_i\right) \in \mathbb{R}^{p_1p_2p_3}$. Then, $\widetilde{\mathbf{z}}_k \overset{iid}{\sim} N(\mathbf{0}, \frac{1}{n_k}\mathbf{\Sigma})$ where $\mathbf{\Sigma} = \mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$. To simplify the presentation, we assume that $n_1 = \cdots = n_K$. Then,

$$
\begin{aligned}
&\mathrm{E}\left[\mathcal{M}_{M+1}(\mathbf{E})\mathcal{M}_{M+1}^T(\mathbf{E})\right] \\
=&\begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -1 & 0 & 0 & \cdots & 1 \end{bmatrix}
\begin{bmatrix} \mathrm{E}\widetilde{\mathbf{z}}_1^T\widetilde{\mathbf{z}}_1 & 0 & \cdots & 0 \\ 0 & \mathrm{E}\widetilde{\mathbf{z}}_2^T\widetilde{\mathbf{z}}_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mathrm{E}\widetilde{\mathbf{z}}_K^T\widetilde{\mathbf{z}}_K \end{bmatrix}
\begin{bmatrix} -1 & -1 & \cdots & -1 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \frac{2\text{tr}(\mathbf{\Sigma})}{n_k} CS(0.5).
\end{aligned} \tag{S5.37}
$$

This implies that

$$
\widehat{\boldsymbol{\mu}} \sim TN(\boldsymbol{\mu}, \frac{2}{n_k}\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3, \mathbf{\Sigma}_4),
$$

where $\boldsymbol{\mu} = [\![\mathbf{\Phi}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{I}_{K-1}]\!]$ and $\mathbf{\Sigma}_4 = CS(0.5)$ when $n_1 = \cdots = n_K$. According to Lemma 17, $\|\mathbf{\Sigma}_4\| = \sigma_1(\mathbf{\Sigma}_4) = \sqrt{K/2}$ and $\sigma_2(\mathbf{\Sigma}_4) = \cdots = \sigma_{K-1}(\mathbf{\Sigma}_4) = \sqrt{1/2}$. We could reformulate $\widehat{\boldsymbol{\mu}}$ as

$$
\widehat{\boldsymbol{\mu}} \overset{\Delta}{=} \boldsymbol{\mu} + \frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i, \tag{S5.38}
$$

where $\widetilde{\mathbf{Z}}_i \sim TN(\mathbf{0}, 2\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3, \mathbf{\Sigma}_4)$ and $\mathbf{\Sigma}_4 = CS(0.5)$.

Step 2: Error bound for $\widehat{\mathbf{A}}_m^{(0)}$. We aim to show that

$$
\begin{aligned}
&P\left(\left\|\sin\Theta(\widehat{\mathbf{A}}_m^{(0)}, \mathbf{A}_m)\right\| \le C\left(\frac{\eta_m\sqrt{\frac{p_m}{n_k}} + c}{\eta_m^2}\right)\right) \\
\ge& 1 - 2\exp\{-\widetilde{c}'p_m\} - 2\exp\{-\widetilde{c}''(n_k\sigma_{r_m}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(\prod_{l\neq m}p_l)C_{\mathbf{\Sigma}}^5)\}, \quad m = 1,2,3.
\end{aligned} \tag{S5.39}
$$

where $\widehat{\mathbf{A}}_m^{(0)}, \mathbf{A}_m \in \mathbb{O}_{p_m \times r_m}$ are the leading r_m left singular vectors of $\mathcal{M}_m(\widehat{\boldsymbol{\mu}})$ and $\mathcal{M}_m(\boldsymbol{\mu})$, respectively, and $C, c, \widetilde{c}, \widetilde{c}' > 0$ are some constants. We give the proof for (S5.39) along mode-1. The proof along other modes follows a similar fashion and hence is omitted.

According to Lemma 7, we have

$$\left\|\sin\Theta(\widehat{\mathbf{A}}_1^{(0)}, \mathbf{A}_1)\right\| \leq \frac{\sigma_{r_1}(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathbf{P}_{(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T}\right\|}{\sigma_{r_1}^2(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})) - \sigma_{r_1+1}^2(\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))} \wedge 1. \tag{S5.40}$$

Note that function $f(x,y,z) = \frac{xy}{x^2-z^2} \leq \frac{C_y}{C_x - C_z^2/C_x}$ if $x \geq C_x$, $y \leq C_y$, and $z \leq C_z$. Thus, to give an upper bound for (S5.40), our goal is to find the lower bound for $\sigma_{r_1}(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))$ and upper bounds for $\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathbf{P}_{(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T}\right\|$ and $\sigma_{r_1+1}^2(\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))$. We analyze $\sigma_{r_1}(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))$, $\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathbf{P}_{(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T}\right\|$, and $\sigma_{r_1+1}^2(\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))$ separately in Steps 1.1 - 1.3 and construct the upper bound for $\left\|\sin\Theta(\widehat{\mathbf{A}}_1^{(0)}, \mathbf{A}_1)\right\|$ in Step 1.4.

Step 2 Firstly, we aim to show that

$$\begin{aligned} &P\left(\sigma_{r_1}^2\left(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right) \geq \left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right) - \frac{2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}{n_k}\right)(1-x)\right) \\ &\geq 1 - 2\exp\left\{c_4 r_1 - c_3\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5\right)(x \wedge \frac{x^2}{8})\right\}. \end{aligned} \tag{S5.41}$$

By (S5.38), we know that

$$\mathcal{M}_1(\widehat{\boldsymbol{\mu}}) = \mathcal{M}_1(\boldsymbol{\mu}) + \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \in \mathbb{R}^{p_1 \times p_2p_3(K-1)},$$

where $\mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \overset{iid}{\sim} MN(\mathbf{0}, 2\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_4 \otimes \boldsymbol{\Sigma}_3 \otimes \boldsymbol{\Sigma}_2)$ and

$$\begin{aligned} &\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}}) \\ =&\left(\mathcal{M}_1(\boldsymbol{\mu}) + \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu}) + \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i)\right) \\ =&\mathcal{M}_1(\boldsymbol{\mu})\mathcal{M}_1^T(\boldsymbol{\mu}) + \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\mathcal{M}_1^T(\boldsymbol{\mu}) + \frac{1}{n_k}\mathcal{M}_1(\boldsymbol{\mu})\sum_{i=1}^{n_k}\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i) + \frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_j). \end{aligned}$$

It follows that

$$\begin{aligned}\mathrm{E}\left[\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\right] =&\mathcal{M}_1(\boldsymbol{\mu})\mathcal{M}_1^T(\boldsymbol{\mu}) + \frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\mathrm{E}\left[\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_j)\right]\\ =&\mathcal{M}_1(\boldsymbol{\mu})\mathcal{M}_1^T(\boldsymbol{\mu}) + \frac{1}{n_k^2}\sum_{i=1}^{n_k}\mathrm{E}\left[\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i)\right]\\ =&\mathcal{M}_1(\boldsymbol{\mu})\mathcal{M}_1^T(\boldsymbol{\mu}) + \frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\boldsymbol{\Sigma}_1, \end{aligned} \tag{S5.42}$$

$$\begin{aligned}\mathrm{E}\left[\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\right] =&\mathbf{A}_1^T\left(\mathcal{M}_1(\boldsymbol{\mu})\mathcal{M}_1^T(\boldsymbol{\mu}) + \frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\boldsymbol{\Sigma}_1\right)\mathbf{A}_1\\ =&\mathrm{diag}\left\{\sigma_1^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right),...,\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)\right\} + \frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\mathbf{A}_1^T\boldsymbol{\Sigma}_1\mathbf{A}_1. \end{aligned} \tag{S5.43}$$

Obviously, $\mathrm{E}\left[\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\right]$ is positive definite. We could find a symmetric normalization matrix $\mathbf{N}\in\mathbb{R}^{r_1\times r_1}$ such that

$$\mathbf{N} = \left(\mathrm{diag}\left\{\sigma_1^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right),...,\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)\right\} + \frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\mathbf{A}_1^T\boldsymbol{\Sigma}_1\mathbf{A}_1\right)^{-1/2}. \tag{S5.44}$$

Consequently,

$$\mathrm{E}\left[\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\mathbf{N}\right] =\mathbf{I}_{r_1},$$

and

$$\begin{aligned}\|\mathbf{N}\|^2 =& \sigma_1^2(\mathbf{N}) = \sigma_1(\mathbf{N}^2) = \sigma_{r_1}^{-1}(\mathbf{N}^{-2})\\ =&\sigma_{r_1}^{-1}\left(\mathrm{diag}\left\{\sigma_1^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right),...,\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)\right\} + \frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\mathbf{A}_1^T\boldsymbol{\Sigma}_1\mathbf{A}_1\right).\end{aligned}$$

According to Lemma 8,

$$\begin{aligned}
&\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-\frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\boldsymbol{\Sigma}_1\| \\
\leq&\sigma_{r_1}\left(\mathrm{diag}\left\{\sigma_1^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right),...,\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)\right\}+\frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\mathbf{A}_1^T\boldsymbol{\Sigma}_1\mathbf{A}_1\right) \\
\leq&\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)+\frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\boldsymbol{\Sigma}_1\|, \qquad \text{(S5.45)}
\end{aligned}$$

and hence

$$\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)+\frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\boldsymbol{\Sigma}_1\|\right)^{-1}\leq\|\mathbf{N}\|^2\leq\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-\frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\boldsymbol{\Sigma}_1\|\right)^{-1}. \qquad \text{(S5.46)}$$

It follows that

$$\begin{aligned}
\sigma_{r_1}^2\left(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right)\geq&\sigma_{r_1}^2\left(\mathbf{N}^{-1}\right)\sigma_{r_1}^2\left(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right) \qquad \text{(S5.47)}\\
=&\sigma_{r_1}\left(\mathbf{N}^{-2}\right)\sigma_{r_1}\left(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\mathbf{N}\right)\\
=&\sigma_{r_1}\left(\mathbf{N}^{-2}\right)\sigma_{r_1}\left(\mathbf{I}_{r_1}+\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\mathbf{N}-\mathbf{I}_{r_1}\right)\\
\geq&\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-\frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\boldsymbol{\Sigma}_1\|\right)\left(1-\|\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\mathbf{N}-\mathbf{I}_{r_1}\|\right), \qquad \text{(S5.48)}
\end{aligned}$$

where the inequality (S5.47) holds due to Lemma 8 and the inequality (S5.48) comes from (S5.45) and Lemma 9. For any unit vector $\mathbf{u} \in \mathbb{R}^{r_1}$, we have

$$
\begin{aligned}
&\mathbf{u}^T \left(\mathbf{N}^T \mathbf{A}_1^T \mathcal{M}_1(\widehat{\boldsymbol{\mu}}) \mathcal{M}_1^T(\widehat{\boldsymbol{\mu}}) \mathbf{A}_1 \mathbf{N} - \mathbf{I}_{r_1} \right) \mathbf{u} \\
=&\mathbf{u}^T \mathbf{N}^T \mathbf{A}_1^T \mathcal{M}_1(\widehat{\boldsymbol{\mu}}) \mathcal{M}_1^T(\widehat{\boldsymbol{\mu}}) \mathbf{A}_1 \mathbf{N} \mathbf{u} - \mathbf{u}^T \mathrm{E} \left[\mathbf{N}^T \mathbf{A}_1^T \mathcal{M}_1(\widehat{\boldsymbol{\mu}}) \mathcal{M}_1^T(\widehat{\boldsymbol{\mu}}) \mathbf{A}_1 \mathbf{N} \right] \mathbf{u} \\
=&\mathbf{u}^T \mathbf{N}^T \mathbf{A}_1^T \mathcal{M}_1(\boldsymbol{\mu}) \mathcal{M}_1^T(\boldsymbol{\mu}) \mathbf{A}_1 \mathbf{N} \mathbf{u} - \mathbf{u}^T \mathrm{E} \left[\mathbf{N}^T \mathbf{A}_1^T \mathcal{M}_1(\boldsymbol{\mu}) \mathcal{M}_1^T(\boldsymbol{\mu}) \mathbf{A}_1 \mathbf{N} \right] \mathbf{u} \\
&+ 2\mathbf{u}^T \mathbf{N}^T \mathbf{A}_1^T \mathcal{M}_1(\boldsymbol{\mu}) \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i) \right) \mathbf{A}_1 \mathbf{N} \mathbf{u} - 2\mathrm{E} \left[\mathbf{u}^T \mathbf{N}^T \mathbf{A}_1^T \mathcal{M}_1(\boldsymbol{\mu}) \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i) \right) \mathbf{A}_1 \mathbf{N} \mathbf{u} \right] \\
&+ \mathbf{u}^T \mathbf{N}^T \mathbf{A}_1^T \left(\frac{1}{n_k^2} \sum_{i,j=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \mathcal{M}_1^T(\widetilde{\mathbf{Z}}_j) \right) \mathbf{A}_1 \mathbf{N} \mathbf{u} - \mathrm{E} \left[\mathbf{u}^T \mathbf{N}^T \mathbf{A}_1^T \left(\frac{1}{n_k^2} \sum_{i,j=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \mathcal{M}_1^T(\widetilde{\mathbf{Z}}_j) \right) \mathbf{A}_1 \mathbf{N} \mathbf{u} \right] \\
=&\underbrace{2\,\mathbf{u}^T \mathbf{N}^T \mathbf{A}_1^T \mathcal{M}_1(\boldsymbol{\mu}) \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i) \right) \mathbf{A}_1 \mathbf{N} \mathbf{u}}_{I_2} \\
&\underbrace{+ \mathbf{u}^T \mathbf{N}^T \mathbf{A}_1^T \left(\frac{1}{n_k^2} \sum_{i,j=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \mathcal{M}_1^T(\widetilde{\mathbf{Z}}_j) \right) \mathbf{A}_1 \mathbf{N} \mathbf{u} - \mathrm{E} \left[\mathbf{u}^T \mathbf{N}^T \mathbf{A}_1^T \left(\frac{1}{n_k^2} \sum_{i,j=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \mathcal{M}_1^T(\widetilde{\mathbf{Z}}_j) \right) \mathbf{A}_1 \mathbf{N} \mathbf{u} \right]}_{I_1}.
\end{aligned}
\tag{S5.49}
$$

We analyze I_1 and I_2 separately.

Step 2.1 - Analysis on I_1 Set $\mathbf{F} = \mathbf{I}_{p_2 p_3 (K-1)} \otimes \left((\mathbf{A}_1 \mathbf{N} \mathbf{u})(\mathbf{A}_1 \mathbf{N} \mathbf{u})^T \right)$. Note that

$$
\begin{aligned}
&(\mathbf{A}_1 \mathbf{N} \mathbf{u})^T \left(\frac{1}{n_k^2} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \mathcal{M}_1^T(\widetilde{\mathbf{Z}}_j) \right) (\mathbf{A}_1 \mathbf{N} \mathbf{u}) \\
=&\left\langle \frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i), (\mathbf{A}_1 \mathbf{N} \mathbf{u})(\mathbf{A}_1 \mathbf{N} \mathbf{u})^T \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \right) \right\rangle \\
=&\mathrm{vec}^T \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \right) \mathrm{vec} \left((\mathbf{A}_1 \mathbf{N} \mathbf{u})(\mathbf{A}_1 \mathbf{N} \mathbf{u})^T \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \right) \right) \\
=&\mathrm{vec}^T \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \right) \left(\mathbf{I}_{p_2 p_3 (K-1)} \otimes (\mathbf{A}_1 \mathbf{N} \mathbf{u})(\mathbf{A}_1 \mathbf{N} \mathbf{u})^T \right) \mathrm{vec} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \right) \\
=&\mathrm{vec}^T \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \right) \mathbf{F} \mathrm{vec} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1(\widetilde{\mathbf{Z}}_i) \right).
\end{aligned}
$$

Thus, I_1 could be rewritten as

$$I_1 = \text{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\mathbf{F}\text{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right) - \text{E}\left[\text{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\mathbf{F}\text{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\right].$$

Note that $\mathbf{F} \in \mathbb{R}^{p_1p_2p_3(K-1)\times p_1p_2p_3(K-1)}$, and

$$\|\mathbf{F}\| = \|(\mathbf{A}_1\mathbf{N}\mathbf{u})(\mathbf{A}_1\mathbf{N}\mathbf{u})^T\| = \|\mathbf{N}\mathbf{u}\|_2^2 \leq \|\mathbf{N}\|^2\|\mathbf{u}\|_2^2 \leq \left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right) - \frac{2(K-1)\text{tr}(\boldsymbol{\Sigma}_2)\text{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\boldsymbol{\Sigma}_1\|\right)^{-1} \tag{S5.50}$$

$$\begin{aligned}\|\mathbf{F}\|_F^2 &= p_2p_3(K-1)\|(\mathbf{A}_1\mathbf{N}\mathbf{u})(\mathbf{A}_1\mathbf{N}\mathbf{u})^T\|_F^2 = p_2p_3(K-1)\|\mathbf{N}\mathbf{u}\|_2^4 \leq p_2p_3(K-1)\|\mathbf{N}\|^4\|\mathbf{u}\|_2^4 \\ &\leq p_2p_3(K-1)\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right) - \frac{2(K-1)\text{tr}(\boldsymbol{\Sigma}_2)\text{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\boldsymbol{\Sigma}_1\|\right)^{-2},\end{aligned} \tag{S5.51}$$

where the first equality holds due to Lemma 5 and the last inequality comes from (S5.46) for both (S5.50) and (S5.51). Let $\check{\mathbf{Z}}_i \overset{iid}{\sim} MN\left(\mathbf{0}, \mathbf{I}_{p_1}, \mathbf{I}_{p_2p_3(K-1)}\right)$. Then, $\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}\text{vec}(\check{\mathbf{Z}}_i)\right\|_{\psi_2} \leq \frac{c_0}{\sqrt{n_k}}$,

$$\mathcal{M}_1(\widetilde{\mathbf{Z}}_i) = \sqrt{2}\boldsymbol{\Sigma}_1^{1/2}\check{\mathbf{Z}}_i\left(\boldsymbol{\Sigma}_4\otimes\boldsymbol{\Sigma}_3\otimes\boldsymbol{\Sigma}_2\right)^{1/2}, \quad \text{vec}\left(\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right) = \sqrt{2}\left[\left(\boldsymbol{\Sigma}_4\otimes\boldsymbol{\Sigma}_3\otimes\boldsymbol{\Sigma}_2\right)^{1/2}\otimes\boldsymbol{\Sigma}_1^{1/2}\right]\text{vec}\left(\check{\mathbf{Z}}_i\right).$$

Set $\widetilde{\boldsymbol{\Sigma}}^{1/2} = \left[\left(\boldsymbol{\Sigma}_4\otimes\boldsymbol{\Sigma}_3\otimes\boldsymbol{\Sigma}_2\right)^{1/2}\otimes\boldsymbol{\Sigma}_1^{1/2}\right]$. Then,

$$\|\widetilde{\boldsymbol{\Sigma}}^{1/2}\mathbf{F}\widetilde{\boldsymbol{\Sigma}}^{1/2}\| \leq \|\mathbf{F}\|\cdot\|\widetilde{\boldsymbol{\Sigma}}\| = \|\mathbf{F}\|\cdot\|\boldsymbol{\Sigma}_1\|\cdot\|\boldsymbol{\Sigma}_2\|\cdot\|\boldsymbol{\Sigma}_3\|\cdot\|\boldsymbol{\Sigma}_4\| \leq C_{\boldsymbol{\Sigma}}^3\sqrt{\frac{K}{2}}\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right) - \frac{2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}{n_k}\right)^{-1} \tag{S5.52}$$

$$\begin{aligned}\|\boldsymbol{\Sigma}^{1/2}\mathbf{F}\boldsymbol{\Sigma}^{1/2}\|_F^2 &\leq \left(\|\widetilde{\boldsymbol{\Sigma}}^{1/2}\|\cdot\|\mathbf{F}\|_F\cdot\|\widetilde{\boldsymbol{\Sigma}}^{1/2}\|\right)^2 = \|\mathbf{F}\|_F^2\cdot\|\boldsymbol{\Sigma}\|^2 = \|\mathbf{F}\|_F^2\cdot\|\boldsymbol{\Sigma}_1\|^2\cdot\|\boldsymbol{\Sigma}_2\|^2\cdot\|\boldsymbol{\Sigma}_3\|^2\cdot\|\boldsymbol{\Sigma}_4\|^2 \\ &\leq \frac{K(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^6}{2}\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right) - \frac{2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}{n_k}\right)^{-2},\end{aligned} \tag{S5.53}$$

where the first inequality holds due to Lemma 4 for both (S5.52) and (S5.53). By the Hanson-Wright inequality in Lemma 11,

$$
\begin{aligned}
&P\left(|I_1|>\frac{x}{2}\right)\\
=&P\left(\left|\frac{1}{n_k^2}\mathrm{vec}^T\left(\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\mathbf{F}\mathrm{vec}\left(\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)-\mathrm{E}\left[\frac{1}{n_k^2}\mathrm{vec}^T\left(\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\mathbf{F}\mathrm{vec}\left(\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\right]\right|>\frac{x}{2}\right)\\
=&P\left(\left|\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{F}\widetilde{\mathbf{\Sigma}}^{1/2}\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)-\mathrm{E}\left[\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i^T\right)\widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{F}\widetilde{\mathbf{\Sigma}}^{1/2}\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\right]\right|>\frac{x}{2}\right)\\
\leq&2\exp\left\{-c_1\min\left(\frac{x^2\left(n_k\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)^2}{2K(K-1)p_2p_3c_0^4C_{\mathbf{\Sigma}}^6},\frac{x\left(n_k\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)}{\sqrt{2K}c_0^2C_{\mathbf{\Sigma}}^3}\right)\right\}.
\end{aligned}
\tag{S5.54}
$$

Step 2.1 - Analysis on I_2 Note that

$$
\begin{aligned}
I_2=&\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i)\right)\mathbf{A}_1\mathbf{N}\mathbf{u}=\mathrm{tr}\left\{\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i)\right)\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\right\}\\
=&\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i)\right)\mathrm{vec}\left(\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\right),
\end{aligned}
$$

and

$$
\begin{aligned}
\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}=&\mathbf{W}_1\mathrm{diag}\left\{\sigma_1(\mathcal{M}_1(\boldsymbol{\mu})),...,\sigma_{r_1}(\mathcal{M}_1(\boldsymbol{\mu}))\right\}\cdot\\
&\left[\mathrm{diag}\left\{\sigma_1^2(\mathcal{M}_1(\boldsymbol{\mu})),...,\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))\right\}+\frac{2(K-1)\mathrm{tr}(\mathbf{\Sigma}_2)\mathrm{tr}(\mathbf{\Sigma}_3)}{n_k}\mathbf{A}_1^T\mathbf{\Sigma}_1\mathbf{A}_1\right]^{-1/2}\mathbf{u},
\end{aligned}
$$

which implies that $\left\|\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right\|_2^2\leq 1$. Thus,

$$
\begin{aligned}
&\left\|\mathrm{vec}\left(\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\right)\right\|_2^2=\left\|\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\right\|_F^2=\|\mathbf{N}\mathbf{u}\|_2^2\left\|\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right\|_2^2\\
\leq&\|\mathbf{N}\|^2\leq\left(\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-\frac{2(K-1)\mathrm{tr}(\mathbf{\Sigma}_2)\mathrm{tr}(\mathbf{\Sigma}_3)}{n_k}\|\mathbf{\Sigma}_1\|\right)^{-1},
\end{aligned}
$$

and hence

$$\begin{aligned}
&\left\|\widetilde{\mathbf{\Sigma}}^{1/2}\mathrm{vec}\left(\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\right)\right\|_2^2 \leq \left\|\widetilde{\mathbf{\Sigma}}\right\| \cdot \left\|\mathrm{vec}\left(\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\right)\right\|_2^2 \\
=&\|\mathbf{\Sigma}_1\| \cdot \|\mathbf{\Sigma}_2\| \cdot \|\mathbf{\Sigma}_3\| \cdot \|\mathbf{\Sigma}_4\| \cdot \left\|\mathrm{vec}\left(\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\right)\right\|_2^2 \\
\leq& C_{\mathbf{\Sigma}}^3\sqrt{\frac{K}{2}}\left(\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - \frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}\right)^{-1}.
\end{aligned}$$

By Lemma 12, we have

$$\begin{aligned}
P\left(|2I_2| > \frac{x}{2}\right) =& P\left(\left|\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i)\right)\mathbf{A}_1\mathbf{N}\mathbf{u}\right| > \frac{x}{4}\right) \\
=& P\left(\left|\frac{1}{n_k}\sum_{i=1}^{n_k}\mathrm{vec}^T\left(\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_i)\right)\mathrm{vec}\left(\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\right)\right| > \frac{x}{4}\right) \\
=& P\left(\left|\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\widetilde{\mathbf{\Sigma}}^{1/2}\mathrm{vec}\left(\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\right)\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right)^T\right)\right| > \frac{x}{4}\right) \\
\leq& 2\exp\left\{-\frac{c_2x^2\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)}{8\sqrt{2K}c_0^2C_{\mathbf{\Sigma}}^3}\right\}. \qquad \text{(S5.55)}
\end{aligned}$$

Combining (S5.49), (S5.54) and (S5.55) gives

$$P\left(\left|\mathbf{u}^T\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\mathbf{N}\mathbf{u} - \mathbf{u}^T\mathbf{I}_{r_1}\mathbf{u}\right| > x\right) \leq 2\exp\left\{-c_3\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)(x \wedge \frac{x^2}{8})\right\}$$

By the ϵ-net argument in Lemma 10, we have

$$P\left(\left\|\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\mathbf{N} - \mathbf{I}_{r_1}\right\| > x\right) \leq 2\exp\left\{c_4r_1 - c_3\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)(x \wedge \frac{x^2}{8})\right\}, \qquad \text{(S5.56)}$$

which implies that

$$\begin{aligned}
&P\left(\sigma_{r_1}^2\left(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right) \geq \left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right) - \frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}\right)(1-x)\right) \\
\geq& 1 - 2\exp\left\{c_4r_1 - c_3\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)(x \wedge \frac{x^2}{8})\right\}.
\end{aligned}$$

Step 2.2 Secondly, we aim to show

$$P\left(\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathbf{P}_{(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T}\right\| \le x + \frac{2\sqrt{2}(K-1)C_{\boldsymbol{\Sigma}}^5}{n_k\sqrt{\sigma_{r_1}^2(\mathcal{M}(\boldsymbol{\mu})) - \frac{2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}{n_k}}}\right)$$
$$\ge 1 - 2\exp\left\{c_4p_1 - c_3'n_kx^2\right\} - 2\exp\left\{-c_5\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5\right)\right\}. \tag{S5.57}$$

Note that

$$\begin{aligned}\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathbf{P}_{(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T}\right\| &= \left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathbf{P}_{(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T}\right\| \\ &= \left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T\left[\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\mathbf{N}\right]^{-1}(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))\right\| \\ &\le \left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T\right\|\sigma_{r_1}^{-1}(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})),\end{aligned} \tag{S5.58}$$

where the inequality in (S5.58) holds due to Lemmas 4 and 1. We analyze $\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T\right\|$ and $\sigma_{r_1}^{-1}(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))$ separately.

Step 2.2 - Analysis on $\sigma_{r_1}^{-1}(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))$ In Step 1.1 we know that

$$\sigma_{r_1}^2\left(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right) = \sigma_{r_1}\left(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\mathbf{N}\right) \ge 1 - \left\|\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_1\mathbf{N} - \mathbf{I}_{r_1}\right\|,$$

and by (S5.56), we have

$$P\left(\sigma_{r_1}^2\left(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right) \ge 1 - x\right) \ge 1 - 2\exp\left\{c_4r_1 - c_3\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5\right)(x \wedge \frac{x^2}{8})\right\}.$$

Set $x = 1/2$. Since $\eta^2 \geq C_{\text{gap}} p^{5/2}/n_k$ and $r_m \leq C_0 p^{1/2}$ for large constants $C_{\text{gap}}, C_0 > 0$,

$$\begin{aligned} P\left(\sigma_{r_1}^{-1}\left(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right) \leq \sqrt{2}\right) &= P\left(\sigma_{r_1}^2\left(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right) \geq \frac{1}{2}\right) \\ &\geq 1 - 2\exp\left\{c_4 r_1 - c_3'\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5\right)\right\} \\ &\geq 1 - 2\exp\left\{-c_5\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu})) - 2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5\right)\right\}. \end{aligned} \tag{S5.59}$$

Step 2.2 - Analysis on $\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T\|$ Note that

$$\begin{aligned} &\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T \\ =&\mathbf{A}_{1\perp}^T\mathcal{M}_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N} + \mathbf{A}_{1\perp}^T\mathcal{M}_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\mathcal{M}_1^T\left(\frac{1}{n_k}\sum_{j=1}^{n_k}\widetilde{\mathbf{Z}}_j\right)\mathbf{A}_1\mathbf{N} \\ &- \mathrm{E}\left[\mathbf{A}_{1\perp}^T\mathcal{M}_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\mathcal{M}_1^T\left(\frac{1}{n_k}\sum_{j=1}^{n_k}\widetilde{\mathbf{Z}}_j\right)\mathbf{A}_1\mathbf{N}\right] + \frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\mathbf{A}_{1\perp}^T\boldsymbol{\Sigma}_1\mathbf{A}_1\mathbf{N}. \end{aligned}$$

For any unit vector $\mathbf{u} \in \mathbb{R}^{r_1}$ and $\mathbf{v} \in \mathbb{R}^{p_1 - r_1}$,

$$\begin{aligned} &\mathbf{v}^T\mathbf{A}_{1\perp}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u} = \mathrm{tr}\left\{\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}^T\right\} \\ =&\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\mathrm{vec}\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}^T\right) = \mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\widetilde{\boldsymbol{\Sigma}}^{1/2}\mathrm{vec}\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}^T\right), \end{aligned}$$

and

$$\begin{aligned} \|\widetilde{\boldsymbol{\Sigma}}^{1/2}\mathrm{vec}\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}^T\right)\|_2^2 \leq&\|\widetilde{\boldsymbol{\Sigma}}\| \cdot \|\mathrm{vec}\left(\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}^T\right)\|_2^2 \\ \leq&\|\boldsymbol{\Sigma}_1\| \cdot \|\boldsymbol{\Sigma}_2\| \cdot \|\boldsymbol{\Sigma}_3\| \cdot \|\boldsymbol{\Sigma}_4\| \cdot \|\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\|_2^2 \cdot \|\mathbf{A}_{1\perp}\mathbf{v}\|_2^2 \leq C_{\boldsymbol{\Sigma}}^3\sqrt{\frac{K}{2}}. \end{aligned}$$

By Lemma 12, it follows that

$$P\left(\left|\mathbf{v}^T\mathbf{A}_{1\perp}^T\left(\frac{1}{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\mathbf{u}\right| \geq \frac{x}{2\sqrt{2}}\right) \leq 2\exp\left\{-\frac{c_3 n_k x^2}{4\sqrt{2K}c_0^2C_{\boldsymbol{\Sigma}}^3}\right\}. \tag{S5.60}$$

By Lemma 10, we have

$$P\left(\left\|\mathbf{A}_{1\perp}^T\left(\frac{1}{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\mathcal{M}_1^T(\boldsymbol{\mu})\mathbf{A}_1\mathbf{N}\right\|<\frac{x}{2\sqrt{2}}\right)\geq 1-2\exp\left\{c_4p_1-\frac{c_3n_kx^2}{4\sqrt{2K}c_0^2C_{\boldsymbol{\Sigma}}^3}\right\}. \tag{S5.61}$$

Moreover,

$$\begin{aligned}
&\mathbf{v}^T\mathbf{A}_{1\perp}^T\mathcal{M}_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\mathcal{M}_1^T\left(\frac{1}{n_k}\sum_{j=1}^{n_k}\widetilde{\mathbf{Z}}_j\right)\mathbf{A}_1\mathbf{N}\mathbf{u}=\mathrm{tr}\left\{\mathcal{M}_1^T\left(\frac{1}{n_k}\sum_{j=1}^{n_k}\widetilde{\mathbf{Z}}_j\right)\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}^T\mathcal{M}_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\right\}\\
=&\mathrm{vec}^T\left(\mathcal{M}_1\left(\frac{1}{n_k}\sum_{j=1}^{n_k}\widetilde{\mathbf{Z}}_j\right)\right)\mathrm{vec}\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}\mathcal{M}_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\right)\\
=&\mathrm{vec}^T\left(\mathcal{M}_1\left(\frac{1}{n_k}\sum_{j=1}^{n_k}\widetilde{\mathbf{Z}}_j\right)\right)\left(\mathbf{I}_{p_2p_3(K-1)}\otimes\left(\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}^T\right)\right)\mathrm{vec}\left(\mathcal{M}_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\right)\\
=&\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{j=1}^{n_k}\check{\mathbf{Z}}_j^T\right)\widetilde{\boldsymbol{\Sigma}}^{1/2}\widetilde{\mathbf{F}}\boldsymbol{\Sigma}^{1/2}\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right),
\end{aligned}$$

where $\widetilde{\mathbf{F}}=\mathbf{I}_{p_2p_3(K-1)}\otimes(\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}^T)$ and

$$\|\widetilde{\mathbf{F}}\|=\|\mathbf{A}_1\mathbf{N}\mathbf{u}\mathbf{v}^T\mathbf{A}_{1\perp}^T\|=\|\mathbf{N}\mathbf{u}\mathbf{v}^T\|\leq\|\mathbf{N}\|\cdot\|\mathbf{u}\mathbf{v}^T\|\leq\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-\frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\boldsymbol{\Sigma}_1\|\right)^{-1/2} \tag{S5.62}$$

$$\begin{aligned}
\|\widetilde{\mathbf{F}}\|_F^2=&p_2p_3(K-1)\|(\mathbf{A}_{1\perp}\mathbf{v})(\mathbf{A}_1\mathbf{N}\mathbf{u})^T\|_F^2=p_2p_3(K-1)\|\mathbf{A}_{1\perp}\mathbf{v}\|_2^2\cdot\|\mathbf{A}_1\mathbf{N}\mathbf{u}\|_2^2\leq p_2p_3(K-1)\|\mathbf{N}\|^2\\
\leq&p_2p_3(K-1)\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-\frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\boldsymbol{\Sigma}_1\|\right)^{-1},
\end{aligned} \tag{S5.63}$$

$$\|\widetilde{\boldsymbol{\Sigma}}^{1/2}\widetilde{\mathbf{F}}\widetilde{\boldsymbol{\Sigma}}^{1/2}\|\leq\|\widetilde{\mathbf{F}}\|\cdot\|\widetilde{\boldsymbol{\Sigma}}\|=\|\widetilde{\mathbf{F}}\|\cdot\|\boldsymbol{\Sigma}_1\|\cdot\|\boldsymbol{\Sigma}_2\|\cdot\|\boldsymbol{\Sigma}_3\|\cdot\|\boldsymbol{\Sigma}_4\|\leq C_{\boldsymbol{\Sigma}}^3\sqrt{\frac{K}{2}}\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-\frac{2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}{n_k}\right)^{-1/} \tag{S5.64}$$

$$\begin{aligned}
\|\widetilde{\boldsymbol{\Sigma}}^{1/2}\widetilde{\mathbf{F}}\widetilde{\boldsymbol{\Sigma}}^{1/2}\|_F^2\leq&\left(\|\widetilde{\boldsymbol{\Sigma}}^{1/2}\|\cdot\|\widetilde{\mathbf{F}}\|_F\cdot\|\widetilde{\boldsymbol{\Sigma}}^{1/2}\|\right)^2=\|\widetilde{\mathbf{F}}\|_F^2\cdot\|\widetilde{\boldsymbol{\Sigma}}\|^2=\|\widetilde{\mathbf{F}}\|_F^2\cdot\|\boldsymbol{\Sigma}_1\|^2\cdot\|\boldsymbol{\Sigma}_2\|^2\cdot\|\boldsymbol{\Sigma}_3\|^2\cdot\|\boldsymbol{\Sigma}_4\|^2\\
\leq&p_2p_3(K-1)C_{\boldsymbol{\Sigma}}^6\left(\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-\frac{2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}{n_k}\right)^{-1}.
\end{aligned} \tag{S5.65}$$

By Lemma 11, we have

$$
\begin{aligned}
&P\left(\left|\mathbf{v}^T\mathbf{A}_{1\perp}^T\left[\mathcal{M}_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\mathcal{M}_1^T\left(\frac{1}{n_k}\sum_{j=1}^{n_k}\widetilde{\mathbf{Z}}_j\right)-\frac{2(K-1)\mathrm{tr}(\mathbf{\Sigma}_2)\mathrm{tr}(\mathbf{\Sigma}_3)}{n_k}\mathbf{\Sigma}_1\right]\mathbf{A}_1\mathbf{N}\mathbf{u}\right|\geq\frac{x}{2\sqrt{2}}\right)\\
=&P\left(\left|\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i^T\right)\mathbf{\Sigma}^{1/2}\widetilde{\mathbf{F}}\mathbf{\Sigma}^{1/2}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)-\mathrm{E}\left[\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i^T\right)\mathbf{\Sigma}^{1/2}\widetilde{\mathbf{F}}\mathbf{\Sigma}^{1/2}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\right]\right|\geq\frac{x}{2\sqrt{2}}\right)\\
\leq&2\exp\left\{-c_1\min\left(\frac{x^2n_k\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)}{8(K-1)p_2p_3c_0^4C_{\mathbf{\Sigma}}^6},\frac{x\sqrt{n_k}}{2\sqrt{K}c_0^2C_{\mathbf{\Sigma}}^3}\sqrt{n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}\right)\right.
\end{aligned}
$$

Again, by the ϵ-net argument in Lemma 10, we have

$$
\begin{aligned}
&P\left(\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T-\mathrm{E}\left[\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})(\mathbf{N}^T\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T\right]\right\|\geq\frac{x}{2\sqrt{2}}\right)\\
\leq&2\exp\left\{c_4p_1-\frac{c_1'x^2n_k\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)}{(K-1)p_2p_3}\right\}.
\end{aligned}
\tag{S5.66}
$$

Besides,

$$
\left\|\mathbf{A}_{1\perp}^T\mathbf{\Sigma}_1\mathbf{A}_1\mathbf{N}\right\|\leq\|\mathbf{\Sigma}_1\|\cdot\|\mathbf{N}\|\leq C_{\mathbf{\Sigma}}\left(\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-\frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}\right)^{-1/2}.
\tag{S5.67}
$$

Combining (S5.58), (S5.59), (S5.66), and (S5.67), we have

$$
\begin{aligned}
&P\left(\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathbf{P}_{(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T}\right\|\leq x+\frac{2\sqrt{2}(K-1)C_{\mathbf{\Sigma}}^5}{n_k\sqrt{\sigma_{r_1}^2(\mathcal{M}(\boldsymbol{\mu}))-\frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}}}\right)\\
\geq&1-2\exp\left\{c_4p_1-c_3'n_kx^2\right\}-2\exp\left\{-c_5\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)\right\}.
\end{aligned}
$$

Step 2.3 Thirdly, we aim to prove that

$$
\begin{aligned}
&P\left(\sigma_{r_1+1}^2(\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))\leq\frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}(1+x)\right)\\
\geq&1-2\exp\left\{c_4(p_1-r_1)-2c_1p_2p_3\left(\frac{\sqrt{2}C_{\mathbf{\Sigma}}^2x(K-1)}{c_0^2\sqrt{K}}\wedge\left(\frac{2C_{\mathbf{\Sigma}}^2x(K-1)}{c_0^2\sqrt{K}}\right)^2\right)\right\}
\end{aligned}
\tag{S5.68}
$$

Note that

$$
\sigma_{r_1+1}\left(\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right)=\min_{\mathrm{rank}(\mathbf{G})=r_1}\|\mathcal{M}_1(\widehat{\boldsymbol{\mu}})-\mathbf{G}\|\leq\|\mathcal{M}_1(\widehat{\boldsymbol{\mu}})-\mathbf{P}_{\mathbf{A}_1}\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\|\leq\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right\|,
$$

and

$$\begin{aligned}\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right\|^2 &= \left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp}\right\| \\ &\leq \frac{2(K-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)\mathrm{tr}(\boldsymbol{\Sigma}_3)}{n_k}\|\mathbf{A}_{1\perp}^T\boldsymbol{\Sigma}_1\mathbf{A}_{1\perp}\| + \left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp} - \mathrm{E}\left[\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp}\right.\right. \\ &\leq \frac{2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}{n_k} + \left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp} - \mathrm{E}\left[\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp}\right]\right\|\end{aligned}$$

Similar to the proof of (S5.41), for any unit vector $\mathbf{v} \in \mathbb{R}^{p_1-r_1}$,

$$\begin{aligned}&\mathbf{v}^T\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp}\mathbf{v} - \mathbf{v}^T\mathrm{E}\left[\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp}\right]\mathbf{v} \\ =&\mathbf{v}^T\mathbf{A}_{1\perp}^T\left(\frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_j)\right)\mathbf{A}_{1\perp}\mathbf{v} - \mathbf{v}^T\mathbf{A}_{1\perp}^T\mathrm{E}\left[\frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\mathcal{M}_1^T(\widetilde{\mathbf{Z}}_j)\right]\mathbf{A}_{1\perp}\mathbf{v} \\ =&\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\check{\mathbf{F}}\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right) - \mathrm{E}\left[\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\check{\mathbf{F}}\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{M}_1(\widetilde{\mathbf{Z}}_i)\right)\right] \\ =&\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\widetilde{\boldsymbol{\Sigma}}^{1/2}\check{\mathbf{F}}\widetilde{\boldsymbol{\Sigma}}^{1/2}\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right) - \mathrm{E}\left[\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\widetilde{\boldsymbol{\Sigma}}^{1/2}\check{\mathbf{F}}\widetilde{\boldsymbol{\Sigma}}^{1/2}\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\right]\end{aligned}$$

where $\check{\mathbf{F}} = \mathbf{I}_{p_2p_3(K-1)} \otimes \left((\mathbf{A}_{1\perp}\mathbf{v})(\mathbf{A}_{1\perp}\mathbf{v})^T\right)$ with

$$\|\check{\mathbf{F}}\| \leq 1, \quad \|\widetilde{\boldsymbol{\Sigma}}^{1/2}\check{\mathbf{F}}\widetilde{\boldsymbol{\Sigma}}^{1/2}\| \leq \|\widetilde{\boldsymbol{\Sigma}}\| \cdot \|\check{\mathbf{F}}\| = \|\boldsymbol{\Sigma}_1\| \cdot \|\boldsymbol{\Sigma}_2\| \cdot \|\boldsymbol{\Sigma}_3\| \cdot \|\boldsymbol{\Sigma}_4\| \cdot \|\check{\mathbf{F}}\| \leq C_{\boldsymbol{\Sigma}}^3\sqrt{\frac{K}{2}}, \quad \|\check{\mathbf{F}}\|_F^2 \leq p_2p_3(K-1),$$

$$\|\widetilde{\boldsymbol{\Sigma}}^{1/2}\check{\mathbf{F}}\widetilde{\boldsymbol{\Sigma}}^{1/2}\|_F^2 \leq \|\widetilde{\boldsymbol{\Sigma}}^{1/2}\|^2 \cdot \|\check{\mathbf{F}}\|_F^2 \cdot \|\widetilde{\boldsymbol{\Sigma}}^{1/2}\|^2 = \|\boldsymbol{\Sigma}_1\|^2 \cdot \|\boldsymbol{\Sigma}_2\|^2 \cdot \|\boldsymbol{\Sigma}_3\|^2 \cdot \|\boldsymbol{\Sigma}_4\|^2 \cdot \|\check{\mathbf{F}}\|_F^2 \leq \frac{p_2p_3K(K-1)C_{\boldsymbol{\Sigma}}^6}{2}.$$

Again, by Lemma 11,

$$\begin{aligned}&P\left(\left|\mathbf{v}^T\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp}\mathbf{v} - \mathbf{v}^T\mathrm{E}\left[\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp}\right]\mathbf{v}\right| > x\right) \\ =&P\left(\left|\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\widetilde{\boldsymbol{\Sigma}}^{1/2}\check{\mathbf{F}}\widetilde{\boldsymbol{\Sigma}}^{1/2}\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right) - \mathrm{E}\left[\mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i^T\right)\widetilde{\boldsymbol{\Sigma}}^{1/2}\check{\mathbf{F}}\widetilde{\boldsymbol{\Sigma}}^{1/2}\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\check{\mathbf{Z}}_i\right)\right]\right| > x\right. \\ \leq&2\exp\left\{-c_1\min\left(\frac{2x^2n_k^2}{p_2p_3K(K-1)c_0^4C_{\boldsymbol{\Sigma}}^6}, \frac{\sqrt{2}xn_k}{\sqrt{K}c_0^2C_{\boldsymbol{\Sigma}}^3}\right)\right\}.\end{aligned}$$

Then, by Lemma 10, we have

$$P\left(\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp}-\mathrm{E}\left[\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathcal{M}_1^T(\widehat{\boldsymbol{\mu}})\mathbf{A}_{1\perp}\right]\right\|>\frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}x\right)$$
$$\leq 2\exp\left\{c_4(p_1-r_1)-2c_1''p_2p_3\left(\frac{\sqrt{2}C_{\mathbf{\Sigma}}^2x}{c_0^2}\wedge\left(\frac{2C_{\mathbf{\Sigma}}^2x}{c_0^2}\right)^2\right)\right\}. \tag{S5.69}$$

It follows that

$$P\left(\sigma_{r_1+1}^2(\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))\leq\frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}(1+x)\right)$$
$$\geq 1-2\exp\left\{c_4(p_1-r_1)-2c_1''p_2p_3\left(\frac{\sqrt{2}C_{\mathbf{\Sigma}}^2x}{c_0^2}\wedge\left(\frac{2C_{\mathbf{\Sigma}}^2x}{c_0^2}\right)^2\right)\right\}.$$

Step 2.4 For (S5.41), setting $x=\frac{\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))}{3\left(\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-\frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}\right)}$ gives

$$P\left(\sigma_{r_1}^2\left(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right)\geq\frac{2}{3}\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-\frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}\right)$$
$$\geq 1-2\exp\left\{c_4r_1-\frac{c_3n_k^2\sigma_{r_1}^4\left(\mathcal{M}_1(\boldsymbol{\mu})\right)}{72\left[n_k\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right]}\right\}. \tag{S5.70}$$

For (S5.68), set $x=\frac{n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))}{6(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}-2$. We have

$$P\left(\sigma_{r_1+1}^2\left(\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\right)\leq\frac{\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))}{3}-\frac{2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}{n_k}\right)$$
$$\geq 1-2\exp\left\{c_4(p_1-r_1)-c_1''\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-12(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)\right\}. \tag{S5.71}$$

When $\eta^2\geq C_{\mathrm{gap}}(K-1)p^{5/2}/n_k$ with C_{gap} being large enough, for some small $\widetilde{c}>0$,

$$\frac{c_3n_k^2\sigma_{r_1}^4\left(\mathcal{M}_1(\boldsymbol{\mu})\right)}{72\left[n_k\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right]}-c_4r_1\geq\frac{\widetilde{c}n_k^2\sigma_{r_1}^4\left(\mathcal{M}_1(\boldsymbol{\mu})\right)}{n_k\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}, \tag{S5.72}$$

$$c_1''\left(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-12(K-1)p_2p_3C_{\mathbf{\Sigma}}^5\right)-c_4(p_1-r_1)\geq\frac{\widetilde{c}n_k^2\sigma_{r_1}^4\left(\mathcal{M}_1(\boldsymbol{\mu})\right)}{n_k\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-2(K-1)p_2p_3C_{\mathbf{\Sigma}}^5}, \tag{S5.73}$$

which implies that

$$P\left(\left\|\sin\Theta(\widehat{\mathbf{A}}_1^{(0)},\mathbf{A}_1)\right\|\le\frac{\sqrt{6\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-\frac{18(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}{n_k}}\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathbf{P}_{(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T}\right\|}{\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)}\wedge 1\right)$$

$$\ge 1-2\exp\left\{-\frac{\widetilde{c}n_k^2\sigma_{r_1}^4\left(\mathcal{M}_1(\boldsymbol{\mu})\right)}{n_k\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)-2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}\right\}. \tag{S5.74}$$

Further set $x=\sqrt{p_1/n_k}$ for (S5.57). Then,

$$P\left(\left\|\mathbf{A}_{1\perp}^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}})\mathbf{P}_{(\mathbf{A}_1^T\mathcal{M}_1(\widehat{\boldsymbol{\mu}}))^T}\right\|\le\sqrt{\frac{p_1}{n_k}}+\frac{2\sqrt{2}(K-1)C_{\boldsymbol{\Sigma}}^5}{n_k\sqrt{\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-\frac{2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5}{n_k}}}\right)$$

$$\ge 1-2\exp\left\{-(c_3'-c_4)p_1\right\}-2\exp\left\{-c_5(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-2(K-1)p_2p_3C_{\boldsymbol{\Sigma}}^5)\right\}. \tag{S5.75}$$

It follows that

$$P\left(\left\|\sin\Theta(\widehat{\mathbf{A}}_1^{(0)},\mathbf{A}_1)\right\|\le C\left(\frac{\sigma_{r_1}\left(\mathcal{M}_1(\boldsymbol{\mu})\right)\sqrt{\frac{p_1}{n_k}}+c}{\sigma_{r_1}^2\left(\mathcal{M}_1(\boldsymbol{\mu})\right)}\right)\right)$$

$$\ge 1-2\exp\{-\widetilde{c}'p_1\}-2\exp\{-\widetilde{c}''(n_k\sigma_{r_1}^2(\mathcal{M}_1(\boldsymbol{\mu}))-2p_2p_3C_{\boldsymbol{\Sigma}}^5)\}.$$

where $c>0$ is is a constant depends on $C_{\boldsymbol{\Sigma}}$, C_{gap}, and n_k. Similar inequalities hold for $\left\|\sin\Theta(\widehat{\mathbf{A}}_2^{(0)},\mathbf{A}_2)\right\|$ and $\left\|\sin\Theta(\widehat{\mathbf{A}}_3^{(0)},\mathbf{A}_3)\right\|$ as well. Hence, we have

$$P\left(\left\|\sin\Theta(\widehat{\mathbf{A}}_m^{(0)},\mathbf{A}_m)\right\|\le C\left(\frac{\eta_m\sqrt{\frac{p_m}{n_k}}+c}{\eta_m^2}\right)\right)$$

$$\ge 1-2\exp\{-\widetilde{c}'p_m\}-2\exp\{-\widetilde{c}''(n_k\sigma_{r_m}^2(\mathcal{M}_1(\boldsymbol{\mu}))-2(\prod_{l\neq m}p_l)C_{\boldsymbol{\Sigma}}^5)\},\quad m=1,2,3.$$

Step 3: Error bound for $\widehat{\mathbf{A}}_m$. Our goal is to show that

$$\left\|\sin\Theta(\widehat{\mathbf{A}}_m,\mathbf{A}_m)\right\|\le\frac{C}{\eta}\sqrt{\frac{p\log p}{n_k}},\quad m=1,2,3, \tag{S5.76}$$

with probability at least $1 - O(p^{-1})$ as long as $t_{\max} > 1$. Denote

$$e_t = \max_{m=1,2,3} e_{t,m}, \quad e_{t,m} = \left\| \sin \Theta(\widehat{\mathbf{A}}_m^{(t)}, \mathbf{A}_m) \right\|, \quad t = 0, 1, 2, ...$$

We aim to show that

$$e_{t+1} \leq \frac{2C_1}{\eta} \sqrt{\frac{p \log p}{n_k}} + \frac{6C}{\eta} \sqrt{\frac{p^2 r}{n_k}} e_t \leq \frac{\sqrt{2}}{2} \tag{S5.77}$$

for $t = 0, 1, 2, ...$ We conduct the proof using induction. First, we prove that (S5.77) holds when $t = 0$. Next, we prove that (S5.77) holds for $t + 1$ if it holds for t. Then, we use (S5.77) to derive (S5.76).

Step 3.1 - Induction: $t = 0$ Based on (S5.39), there exist $C_{\mathbf{\Sigma}} > 0$, n_k, and a large $C_{\text{gap}} > 0$ such that $e_0 \leq \sqrt{2}/2$ with probability at least $1 - 2\exp\{-\widetilde{c}p\} - 2\exp\{-\widetilde{c}' n_k \eta^2\}$. By Lemma 19, we have

$$\begin{aligned}
\max_{\mathbf{V}_2 \in \mathbb{R}^{p_2 \times r_2}, \mathbf{V}_3 \in \mathbb{R}^{p_3 \times r_3}} \frac{\left\| \mathcal{M}_1 \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \widetilde{\mathbf{Z}}_i \right) (\mathbf{I}_{K-1} \otimes \mathbf{V}_3 \otimes \mathbf{V}_2) \right\|}{\|\mathbf{V}_2\| \cdot \|\mathbf{V}_3\|} &\leq C \sqrt{\frac{p^2 r}{n_k}}, \\
\max_{\mathbf{V}_1 \in \mathbb{R}^{p_1 \times r_1}, \mathbf{V}_3 \in \mathbb{R}^{p_3 \times r_3}} \frac{\left\| \mathcal{M}_2 \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \widetilde{\mathbf{Z}}_i \right) (\mathbf{I}_{K-1} \otimes \mathbf{V}_3 \otimes \mathbf{V}_1) \right\|}{\|\mathbf{V}_1\| \cdot \|\mathbf{V}_3\|} &\leq C \sqrt{\frac{p^2 r}{n_k}}, \\
\max_{\mathbf{V}_1 \in \mathbb{R}^{p_1 \times r_1}, \mathbf{V}_2 \in \mathbb{R}^{p_2 \times r_2}} \frac{\left\| \mathcal{M}_3 \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \widetilde{\mathbf{Z}}_i \right) (\mathbf{I}_{K-1} \otimes \mathbf{V}_2 \otimes \mathbf{V}_1) \right\|}{\|\mathbf{V}_1\| \cdot \|\mathbf{V}_2\|} &\leq C \sqrt{\frac{p^2 r}{n_k}},
\end{aligned} \tag{S5.78}$$

with probability at least $1 - C\exp\{-C'pr\}$. By Lemma 16,

$$\begin{aligned}
&\left\| \frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_1 \left(\widetilde{\mathbf{Z}}_i \times_2 \mathbf{A}_2^T \times_3 \mathbf{A}_3^T \right) \right\| \leq C_1 \sqrt{\frac{p \log p}{n_k}}, \quad \left\| \frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_2 \left(\widetilde{\mathbf{Z}}_i \times_1 \mathbf{A}_1^T \times_3 \mathbf{A}_3^T \right) \right\| \leq C_1 \sqrt{\frac{p \log p}{n_k}}, \\
&\left\| \frac{1}{n_k} \sum_{i=1}^{n_k} \mathcal{M}_3 \left(\widetilde{\mathbf{Z}}_i \times_1 \mathbf{A}_1^T \times_2 \mathbf{A}_2^T \right) \right\| \leq C_1 \sqrt{\frac{p \log p}{n_k}},
\end{aligned} \tag{S5.79}$$

with probability at least $1 - O(p^{-1})$.

Under Algorithm S5, the estimate of $\mathbf{A}_1$ in the $(t+1)$-th iteration is

$$\begin{aligned}\widehat{\mathbf{A}}_1^{(t+1)} &= \mathrm{SVD}_{r_1}\left(\mathcal{M}_1(\widehat{\boldsymbol{\mu}} \times_2 \widehat{\mathbf{A}}_2^{(t)T} \times_3 \widehat{\mathbf{A}}_3^{(3)T} \times_4 \mathbf{I}_{K-1})\right) \\ &= \mathrm{SVD}_{r_1}\left(\mathcal{M}_1\left(\boldsymbol{\mu} + \frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\left(\mathbf{I}_{K-1} \otimes \widehat{\mathbf{A}}_3^{(t)} \otimes \widehat{\mathbf{A}}_2^{(t)}\right)\right).\end{aligned}$$

By Wedin's $\sin\Theta$ theorem in Lemma 30, we have

$$e_{t+1,1} \leq \frac{2\left\|\mathcal{M}_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\left(\mathbf{I}_{K-1} \otimes \widehat{\mathbf{A}}_3^{(t)} \otimes \widehat{\mathbf{A}}_2^{(t)}\right)\right\|}{\sigma_{r_1}\left(M_1(\boldsymbol{\mu})\left(\mathbf{I}_{K-1} \otimes \widehat{\mathbf{A}}_3^{(t)} \otimes \widehat{\mathbf{A}}_2^{(t)}\right)\right)}. \tag{S5.80}$$

Moreover,

$$\sigma_{r_1}\left(M_1(\boldsymbol{\mu})\left(\mathbf{I}_{K-1} \otimes \widehat{\mathbf{A}}_3^{(t)} \otimes \widehat{\mathbf{A}}_2^{(t)}\right)\right) = \sigma_{r_1}\left(M_1(\boldsymbol{\mu}) \cdot \mathbf{P}_{\mathbf{I}_{K-1}\otimes\mathbf{A}_3\otimes\mathbf{A}_2} \cdot \left(\mathbf{I}_{K-1} \otimes \widehat{\mathbf{A}}_3^{(t)} \otimes \widehat{\mathbf{A}}_2^{(t)}\right)\right) \tag{S5.81}$$

$$\begin{aligned}=&\sigma_{r_1}\left(M_1(\boldsymbol{\mu})\left(\mathbf{I}_{K-1} \otimes \mathbf{A}_3 \otimes \mathbf{A}_2\right)\left(\mathbf{I}_{K-1} \otimes \mathbf{A}_3 \otimes \mathbf{A}_2\right)^T\left(\mathbf{I}_{K-1} \otimes \widehat{\mathbf{A}}_3^{(t)} \otimes \widehat{\mathbf{A}}_2^{(t)}\right)\right) \\ =&\sigma_{r_1}\left(M_1(\boldsymbol{\mu})\left(\mathbf{I}_{K-1} \otimes \mathbf{A}_3 \otimes \mathbf{A}_2\right)\left(\mathbf{I}_{K-1} \otimes \mathbf{A}_3^T\widehat{\mathbf{A}}_3^{(t)} \otimes \mathbf{A}_2^T\widehat{\mathbf{A}}_2^{(t)}\right)\right) \\ \geq&\sigma_{r_1}\left(M_1(\boldsymbol{\mu})\right)\sigma_{\min}\left(\mathbf{A}_3\right)\sigma_{\min}\left(\mathbf{A}_2\right)\sigma_{\min}\left(\mathbf{A}_3^T\widehat{\mathbf{A}}_3^{(t)}\right)\sigma_{\min}\left(\mathbf{A}_2^T\widehat{\mathbf{A}}_2^{(t)}\right) \qquad \text{(S5.82)} \\ \geq&\sigma_{r_1}\left(M_1(\boldsymbol{\mu})\right)(1-e_t^2), \qquad \text{(S5.83)}\end{aligned}$$

where the equation in (S5.81) holds because $\mathbf{I}_{K-1}\otimes\mathbf{A}_3\otimes\mathbf{A}_2$ spans the right singular space of $M_1(\boldsymbol{\mu})$, the inequality in (S5.82) holds due to lemma 8, and the inequality in (S5.83)

holds due to lemma 6. Meanwhile,

$$
\begin{aligned}
&\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\left(\mathbf{I}_{K-1}\otimes\mathbf{A}_3^{(t)}\otimes\mathbf{A}_2^{(t)}\right)\right\| \\
=&\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\left(\mathbf{P}_{\mathbf{I}_{K-1}\otimes\mathbf{A}_3\otimes\mathbf{A}_2}+\mathbf{P}_{\mathbf{I}_{K-1}\otimes\mathbf{A}_{3\perp}\otimes\mathbf{A}_2}+\mathbf{P}_{\mathbf{I}_{K-1}\otimes\mathbf{I}_{p_3}\otimes\mathbf{A}_2}\right)(\mathbf{I}_{K-1}\otimes\mathbf{A}_3\otimes\mathbf{A}_2)\right\| \\
\le&\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\mathbf{P}_{\mathbf{I}_{K-1}\otimes\mathbf{A}_3\otimes\mathbf{A}_2}\left(\mathbf{I}_{K-1}\otimes\mathbf{A}_3^{(t)}\otimes\mathbf{A}_2^{(t)}\right)\right\| \\
&+\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\mathbf{P}_{\mathbf{I}_{K-1}\otimes\mathbf{A}_{3\perp}\otimes\mathbf{A}_2}\left(\mathbf{I}_{K-1}\otimes\mathbf{A}_3^{(t)}\otimes\mathbf{A}_2^{(t)}\right)\right\| \\
&+\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\mathbf{P}_{\mathbf{I}_{K-1}\otimes\mathbf{A}_3\otimes\mathbf{A}_{2\perp}}\left(\mathbf{I}_{K-1}\otimes\mathbf{A}_3^{(t)}\otimes\mathbf{A}_2^{(t)}\right)\right\| \\
&+\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\mathbf{P}_{\mathbf{I}_{K-1}\otimes\mathbf{A}_{3\perp}\otimes\mathbf{A}_{2\perp}}\left(\mathbf{I}_{K-1}\otimes\mathbf{A}_3^{(t)}\otimes\mathbf{A}_2^{(t)}\right)\right\| \\
=&\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)(\mathbf{I}_{K-1}\otimes\mathbf{A}_3\otimes\mathbf{A}_2)(\mathbf{I}_{K-1}\otimes\mathbf{A}_3\otimes\mathbf{A}_2)^T\left(\mathbf{I}_{K-1}\otimes\mathbf{A}_3^{(t)}\otimes\mathbf{A}_2^{(t)}\right)\right\| \\
&+\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\left(\mathbf{I}_{K-1}\otimes(\mathbf{P}_{\mathbf{A}_{3\perp}}\mathbf{A}_3^{(t)})\otimes(\mathbf{P}_{\mathbf{A}_2}\mathbf{A}_2^{(t)})\right)\right\| \\
&+\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\left(\mathbf{I}_{K-1}\otimes(\mathbf{P}_{\mathbf{A}_3}\mathbf{A}_3^{(t)})\otimes(\mathbf{P}_{\mathbf{A}_{2\perp}}\mathbf{A}_2^{(t)})\right)\right\| \\
&+\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\left(\mathbf{I}_{K-1}\otimes(\mathbf{P}_{\mathbf{A}_{3\perp}}\mathbf{A}_3^{(t)})\otimes(\mathbf{P}_{\mathbf{A}_{2\perp}}\mathbf{A}_2^{(t)})\right)\right\| \\
\le&\left\| M_1\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)(\mathbf{I}_{K-1}\otimes\mathbf{A}_3\otimes\mathbf{A}_2)\right\|+C\sqrt{\frac{p^2r}{n_k}}\left\|\mathbf{P}_{\mathbf{A}_{3\perp}}\mathbf{A}_3^{(t)}\right\|\cdot\left\|\mathbf{P}_{\mathbf{A}_2}\mathbf{A}_2^{(t)}\right\| \qquad \text{(S5.84)} \\
&+C\sqrt{\frac{p^2r}{n_k}}\left\|\mathbf{P}_{\mathbf{A}_3}\mathbf{A}_3^{(t)}\right\|\cdot\left\|\mathbf{P}_{\mathbf{A}_{2\perp}}\mathbf{A}_2^{(t)}\right\|+C\sqrt{\frac{p^2r}{n_k}}\left\|\mathbf{P}_{\mathbf{A}_{3\perp}}\mathbf{A}_3^{(t)}\right\|\cdot\left\|\mathbf{P}_{\mathbf{A}_{2\perp}}\mathbf{A}_2^{(t)}\right\| \\
\le&C_1\sqrt{\frac{p\log p}{n_k}}+2C\sqrt{\frac{p^2r}{n_k}}e_t+C\sqrt{\frac{p^2r}{n_k}}e_t^2\le C_1\sqrt{\frac{p\log p}{n_k}}+3C\sqrt{\frac{p^2r}{n_k}}e_t, \qquad \text{(S5.85)}
\end{aligned}
$$

with probability at least $1-O(p^{-1})$ where the inequality in (S5.84) holds due to (S5.78) and the first inequality in (S5.85) holds due to (S5.79) and Lemma 6. Thus,

$$
e_{t+1,1}\le\frac{C_1\sqrt{\frac{p\log p}{n_k}}+3C\sqrt{\frac{p^2r}{n_k}}e_t}{\sigma_{r_1}(\mathcal{M}_1(\boldsymbol{\mu}))(1-e_t^2)}. \qquad \text{(S5.86)}
$$

When $t = 0$, we have

$$e_{1,1} \leq \frac{2C_1}{\eta}\sqrt{\frac{p\log p}{n_k}} + \frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}}e_0.$$

Similarly, we can obtain

$$e_{1,2} \leq \frac{2C_1}{\eta}\sqrt{\frac{p\log p}{n_k}} + \frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}}e_0, \quad \text{and} \quad e_{1,3} \leq \frac{2C_1}{\eta}\sqrt{\frac{p\log p}{n_k}} + \frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}}e_0.$$

Since $\eta^2 > C_{\text{gap}} p^{5/2}/n_k$ and $r_m \leq C_0 p^{1/2}$, there exists a large constant $C_{\text{gap}} > 0$ such that

$$e_1 \leq \frac{2C_1}{\eta}\sqrt{\frac{p\log p}{n_k}} + \frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}}e_0 \leq \frac{\sqrt{2}}{2} \quad \text{and} \quad \frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}} \leq \frac{1}{2}.$$

Step 3.2 - Induction $t > 0$ Assume that (S5.77) holds for t. By (S5.86),

$$e_{t+1} \leq \frac{C_1\sqrt{\frac{p\log p}{n_k}} + 3C\sqrt{\frac{p^2 r}{n_k}}e_t}{\sigma_{r_1}(\mathcal{M}_1(\boldsymbol{\mu}))(1-e_t^2)} \leq \frac{2C_1}{\eta}\sqrt{\frac{p\log p}{n_k}} + \frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}}e_t \leq \frac{\sqrt{2}}{2},$$

where the last inequality holds because $e_t \leq \sqrt{2}/2$. Consequently, (S5.77) holds for $t = 0, 1, 2, ...$ Combining (S5.77) and the condition $\frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}} \leq \frac{1}{2}$ yields

$$\begin{aligned}
e_{t+1} &\leq \frac{4C_1}{\eta}\sqrt{\frac{p\log p}{n_k}}\left(1 - \frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}}\right) + \frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}}e_t \\
\Rightarrow e_{t+1} - \frac{4C_1}{\eta}\sqrt{\frac{p\log p}{n_k}} &\leq \frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}}\left(e_t - \frac{4C_1}{\eta}\sqrt{\frac{p\log p}{n_k}}\right) \\
\Rightarrow e_{t_{\max}} - \frac{4C_1}{\eta}\sqrt{\frac{p\log p}{n_k}} &\leq \left(\frac{6C}{\eta}\sqrt{\frac{p^2 r}{n_k}}\right)^{t_{\max}}\left(e_0 - \frac{4C_1}{\eta}\sqrt{\frac{p\log p}{n_k}}\right) \\
\Rightarrow e_{t_{\max}} &\leq \frac{4C_1}{\eta}\sqrt{\frac{p\log p}{n_k}} + \frac{e_0}{2^{t_{\max}}} \leq \frac{C}{\eta}\sqrt{\frac{p\log p}{n_k}}
\end{aligned}$$

as long as $t_{\max} > 1$. Therefore, we have

$$\left\|\sin\Theta(\widehat{\mathbf{A}}_m, \mathbf{A}_m)\right\| \leq e_{\max} \leq \frac{C}{\eta}\sqrt{\frac{p\log p}{n_k}}, \quad m = 1, 2, 3,$$

with probability at least $1 - O(p^{-1})$.

Step 4: Error bound for $\widehat{\mathbf{D}}_m$. Specifically, we aim to show that

$$\|\widehat{\mathbf{D}}_m - \mathbf{D}_m\|_F \leq \frac{C}{\eta}\sqrt{\frac{srp\log p}{n}}, \quad m = 1, 2, 3 \tag{S5.87}$$

with probability at least $1 - O(p^{-1})$. Note that Tucker decomposition is unidentifiable to orthogonal transformations of factor matrices. We consider the factor matrices, $\mathbf{A}_m$'s, that have the the minimum Frobenius norm error with the obtained estimates. Then,

$$\|\widehat{\mathbf{a}}_l^{(m)} - \mathbf{a}_l^{(m)}\|_\infty \leq \left\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\right\|_{\max} \leq \left\|\widehat{\mathbf{A}}_m - \mathbf{A}_m\right\| \leq \sqrt{2}\left\|\sin\Theta\left(\widehat{\mathbf{A}}_m, \mathbf{A}_m\right)\right\| \leq \frac{C}{\eta}\sqrt{\frac{p\log p}{n}} \tag{S5.88}$$

with probability at least $1 - O(p^{-1})$ where $\widehat{\mathbf{a}}_l^{(m)}, \mathbf{a}_l^{(m)} \in \mathbb{R}^{p_m}$ are the l-th column of $\widehat{\mathbf{A}}_m$ and $\mathbf{A}_m$, respectively. Meanwhile, note that

$$\begin{aligned}
\widehat{\mathbf{D}}_m &= \underset{\mathbf{D}\in\mathbb{R}^{p_m\times r_m}}{\arg\min}\left\{\mathrm{tr}\left(\frac{1}{2}\mathbf{D}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_m\mathbf{D} - \widehat{\mathbf{A}}_m^{\mathrm{T}}\mathbf{D}\right) + \lambda\sum_{l=1}^{p_m}\|\mathbf{D}_{l\cdot}\|_2\right\} \\
&= \underset{\substack{\mathbf{D}=[\mathbf{d}_1,\ldots,\mathbf{d}_{r_m}]\\ \mathbf{d}_1,\ldots,\mathbf{d}_{r_m}\in\mathbb{R}^{p_m}}}{\arg\min}\left\{\sum_{l=1}^{r}\left(\frac{1}{2}\mathbf{d}_l^T\widehat{\mathbf{\Sigma}}_m\mathbf{d}_l - \widehat{\mathbf{a}}_l^{(m)T}\mathbf{d}_l\right) + \lambda\sum_{l=1}^{p_m}\|\mathbf{D}_{l\cdot}\|_2\right\},
\end{aligned}$$

which suggests that we are essentially solving (S7.127) along each mode. Hence, we could apply Lemma 21 to obtain the upper bound for $\|\widehat{\mathbf{D}}_m - \mathbf{D}_m\|_F$. Specifically, (S5.88) suggests that condition (i) holds with probability at least $1 - O(p^{-1})$. By Lemma 20condition (ii) holds with probability at least $1 - O(p^{-1})$. Moreover, condition (iii) holds due to Lemma 22. Since $\lambda \asymp \sqrt{\frac{rp\log p}{n}}$, by Lemma 23, we have

$$\|\widehat{\mathbf{D}}_m - \mathbf{D}_m\|_F \leq \frac{C}{\eta}\sqrt{\frac{srp\log p}{n}}, \quad m = 1, 2, 3$$

with probability at least $1 - O(p^{-1})$.

Step 5: Error bound for $\|\widehat{\mathbf{B}}-\mathbf{B}\|_F$. Since $\widehat{\mathbf{B}}$ is constructed by

$$\widehat{\mathbf{B}}=\left[\!\left[[\![\widehat{\boldsymbol{\mu}};\widehat{\mathbf{A}}_1^T,\widehat{\mathbf{A}}_2^T,\widehat{\mathbf{A}}_3^T]\!];\widehat{\mathbf{D}}_1,\widehat{\mathbf{D}}_2,\widehat{\mathbf{D}}_3\right]\!\right]=\widehat{\boldsymbol{\mu}}\times_1\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T\times_2\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T,$$

we have

$$\begin{aligned}
&\widehat{\mathbf{B}}-\mathbf{B}\\
=&\widehat{\boldsymbol{\mu}}\times_1\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T\times_2\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T-\boldsymbol{\mu}\times_1\mathbf{D}_1\mathbf{A}_1^T\times_2\mathbf{D}_2\mathbf{A}_2^T\times_3\mathbf{D}_3\mathbf{A}_3^T\\
=&\boldsymbol{\mu}\times_1\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T\times_2\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T-\boldsymbol{\mu}\times_1\mathbf{D}_1\mathbf{A}_1^T\times_2\mathbf{D}_2\mathbf{A}_2^T\times_3\mathbf{D}_3\mathbf{A}_3^T\\
&+\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\times_1\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T\times_2\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\\
=&\boldsymbol{\mu}\times_1(\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T-\mathbf{D}_1\mathbf{A}_1^T)\times_2\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T+\boldsymbol{\mu}\times_1\mathbf{D}_1\mathbf{A}_1^T\times_2(\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T-\mathbf{D}_2\mathbf{A}_2^T)\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\\
&+\boldsymbol{\mu}\times_1\mathbf{D}_1\mathbf{A}_1^T\times_2\mathbf{D}_2\mathbf{A}_2^T\times_3(\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T-\mathbf{D}_3\mathbf{A}_3^T)+\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\times_1\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T\times_2\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T.
\end{aligned}$$

Then,

$$\|\widehat{\mathbf{B}}-\mathbf{B}\|_F\leq I_1+I_2+I_3+I_4,$$

where

$$\begin{aligned}
I_1&=\|\boldsymbol{\mu}\times_1(\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T-\mathbf{D}_1\mathbf{A}_1^T)\times_2\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\|_F,\\
I_2&=\|\boldsymbol{\mu}\times_1\boldsymbol{\Sigma}_1^{-1}\mathbf{A}_1\mathbf{A}_1^T\times_2(\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T-\mathbf{D}_2\mathbf{A}_2^T)\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\|_F,\\
I_3&=\|\boldsymbol{\mu}\times_1\boldsymbol{\Sigma}_1^{-1}\mathbf{A}_1\mathbf{A}_1^T\times_2\boldsymbol{\Sigma}_2^{-1}\mathbf{A}_2\mathbf{A}_2^T\times_3(\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T-\mathbf{D}_3\mathbf{A}_3^T)\|_F,\\
I_4&=\left\|\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\times_1\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T\times_2\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\times_3\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\right\|_F.
\end{aligned}$$

Note that

$$\begin{aligned}\|\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T-\mathbf{D}_m\mathbf{A}_m^T\|_F=&\|(\widehat{\mathbf{D}}_m-\mathbf{D}_m)\widehat{\mathbf{A}}_m^T+\mathbf{D}_m(\widehat{\mathbf{A}}_m-\mathbf{A}_m)^T\|_F\\ \le&\|\widehat{\mathbf{A}}_m(\widehat{\mathbf{D}}_m-\mathbf{D}_m)^T\|_F+\|\mathbf{D}_m(\widehat{\mathbf{A}}_m-\mathbf{A}_m)^T\|_F\\ \le&\|\widehat{\mathbf{A}}_m\|\cdot\|(\widehat{\mathbf{D}}_m-\mathbf{D}_m)^T\|_F+\|\mathbf{\Sigma}_m^{-1}\|\cdot\|\widehat{\mathbf{A}}_m\|\cdot\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_F\\ \le&\|\widehat{\mathbf{D}}_m-\mathbf{D}_m\|_F+C_{\mathbf{\Sigma}}\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\|_F\\ \le&\frac{C_1}{\eta}\left(\sqrt{\frac{srp\log p}{n}}+\sqrt{\frac{rp\log p}{n}}\right),\end{aligned}\tag{S5.89}$$

$$\begin{aligned}\|\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T\|_F=\|(\widehat{\mathbf{D}}_m-\mathbf{D}_m)\widehat{\mathbf{A}}_m^T+\mathbf{D}_m\widehat{\mathbf{A}}_m^T\|_F&\le\|\widehat{\mathbf{A}}_m(\widehat{\mathbf{D}}_m-\mathbf{D}_m)^T\|_F+\|\mathbf{D}_m\widehat{\mathbf{A}}_m^T\|_F\\ &\le\|\widehat{\mathbf{A}}_m\|\cdot\|\widehat{\mathbf{D}}_m-\mathbf{D}_m\|_F+\|\widehat{\mathbf{A}}_m\|\cdot\|\mathbf{\Sigma}_m^{-1}\|\cdot\|\mathbf{A}_m\|_F\\ &\le\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\mathbf{\Sigma}}\sqrt{r},\end{aligned}\tag{S5.90}$$

and

$$\|\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T\|\le\|\widehat{\mathbf{D}}_m\|\le\|\widehat{\mathbf{D}}_m-\mathbf{D}_m\|+\|\mathbf{D}_m\|\le\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\mathbf{\Sigma}}.\tag{S5.91}$$

It follows that

$$\begin{aligned}I_1&=\left\|(\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T-\mathbf{D}_1\mathbf{A}_1^T)\mathcal{M}_1(\boldsymbol{\mu})(\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\otimes\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T)^T\right\|_F\\ &\le\left\|\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T-\mathbf{D}_1\mathbf{A}_1^T\right\|_F\cdot\|\boldsymbol{\mu}\|_F\cdot\left\|\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\right\|\cdot\left\|\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\right\|\\ &\le\frac{C_3}{\eta}\left(\sqrt{\frac{srp\log p}{n}}+\sqrt{\frac{rp\log p}{n}}\right)\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\mathbf{\Sigma}}\right)^2,\end{aligned}$$

$$\begin{aligned}I_2&=\left\|(\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T-\mathbf{D}_2\mathbf{A}_2^T)\mathcal{M}_2(\boldsymbol{\mu})(\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\otimes\mathbf{\Sigma}_1^{-1}\mathbf{A}_1\mathbf{A}_1^T)^T\right\|_F\\ &\le\left\|\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T-\mathbf{D}_2\mathbf{A}_2^T\right\|_F\cdot\|\boldsymbol{\mu}\|_F\cdot\left\|\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\right\|\cdot\left\|\mathbf{\Sigma}_1^{-1}\mathbf{A}_1\mathbf{A}_1^T\right\|\\ &\le\frac{C_4}{\eta}\left(\sqrt{\frac{srp\log p}{n}}+\sqrt{\frac{rp\log p}{n}}\right)\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\mathbf{\Sigma}}\right),\end{aligned}$$

$$
\begin{aligned}
I_3 &= \left\|(\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T - \mathbf{D}_3\mathbf{A}_3^T)\mathcal{M}_3(\boldsymbol{\mu})(\boldsymbol{\Sigma}_2^{-1}\mathbf{A}_2\mathbf{A}_2^T \otimes \boldsymbol{\Sigma}_1^{-1}\mathbf{A}_1\mathbf{A}_1^T)^T\right\|_F \\
&\leq \left\|\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T - \mathbf{D}_3\mathbf{A}_3^T\right\|_F \cdot \|\boldsymbol{\mu}\|_F \cdot \|\boldsymbol{\Sigma}_2^{-1}\mathbf{A}_2\mathbf{A}_2^T\| \cdot \|\boldsymbol{\Sigma}_1^{-1}\mathbf{A}_1\mathbf{A}_1^T\| \\
&\leq \frac{C_5}{\eta}\left(\sqrt{\frac{srp\log p}{n}} + \sqrt{\frac{rp\log p}{n}}\right),
\end{aligned}
$$

with probability at least $1 - O(p^{-1})$, and

$$
\begin{aligned}
I_4^2 =& \left\|\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right) \times_1 \widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T \times_2 \widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T \times_3 \widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\right\|_F^2 \\
=& \mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\left(\mathbf{I}_{K-1} \otimes (\otimes_{m=3}^1 \widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)^T\left(\mathbf{I}_{K-1} \otimes (\otimes_{m=3}^1 \widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right) \\
=& \mathrm{vec}^T\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right)\left(\mathbf{I}_{K-1} \otimes (\otimes_{m=3}^1 \widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)\mathrm{vec}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{\mathbf{Z}}_i\right) \\
=& 2\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i^T\right)\boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_{K-1} \otimes (\otimes_{m=3}^1 \widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)\boldsymbol{\Sigma}^{1/2}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i\right),
\end{aligned}
$$

where $\underline{\mathbf{Z}}_i \overset{iid}{\sim} N(\mathbf{0}, \mathbf{I}_{p_1p_2p_3(K-1)})$, $\boldsymbol{\Sigma} = \otimes_{m=4}^1\boldsymbol{\Sigma}_m$, and

$$
\begin{aligned}
&\left\|\boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_{K-1} \otimes (\otimes_{m=3}^1 \widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)\boldsymbol{\Sigma}^{1/2}\right\| \\
\leq& \|\boldsymbol{\Sigma}\| \cdot \left\|\widehat{\mathbf{A}}_1\widehat{\mathbf{D}}_1^T\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T\right\| \cdot \left\|\widehat{\mathbf{A}}_2\widehat{\mathbf{D}}_2^T\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\right\| \cdot \left\|\widehat{\mathbf{A}}_3\widehat{\mathbf{D}}_3^T\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\right\| \\
\leq& \|\boldsymbol{\Sigma}\| \cdot \left\|\widehat{\mathbf{D}}_1\right\|^2 \cdot \left\|\widehat{\mathbf{D}}_2\right\|^2 \cdot \left\|\widehat{\mathbf{D}}_3\right\|^2 \\
\leq& C_{\boldsymbol{\Sigma}}^3\sqrt{\frac{K}{2}}\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}} + C_{\boldsymbol{\Sigma}}\right)^6,
\end{aligned}
$$

$$
\begin{aligned}
&\left\|\boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_{K-1} \otimes (\otimes_{m=3}^1 \widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)\boldsymbol{\Sigma}^{1/2}\right\|_F^2 \\
\leq& (K-1)\|\boldsymbol{\Sigma}\|^2 \cdot \left\|\widehat{\mathbf{A}}_1\widehat{\mathbf{D}}_1^T\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T\right\|_F^2 \cdot \left\|\widehat{\mathbf{A}}_2\widehat{\mathbf{D}}_2^T\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\right\|_F^2 \cdot \left\|\widehat{\mathbf{A}}_3\widehat{\mathbf{D}}_3^T\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\right\|_F^2 \\
\leq& (K-1)\|\boldsymbol{\Sigma}\|^2 \cdot \left\|\widehat{\mathbf{D}}_1\widehat{\mathbf{A}}_1^T\right\|_F^4 \cdot \left\|\widehat{\mathbf{D}}_2\widehat{\mathbf{A}}_2^T\right\|_F^4 \cdot \left\|\widehat{\mathbf{D}}_3\widehat{\mathbf{A}}_3^T\right\|_F^4 \\
\leq& \frac{K(K-1)C_{\boldsymbol{\Sigma}}^6}{2}\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}} + C_{\boldsymbol{\Sigma}}\sqrt{r}\right)^{12},
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i^T\right)\boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_{K-1}\otimes(\otimes_{m=3}^{1}\widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)\boldsymbol{\Sigma}^{1/2}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i\right)\right]\\
=&\frac{1}{n_k}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_{K-1}\otimes(\otimes_{m=3}^{1}\widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)\boldsymbol{\Sigma}^{1/2}\right\}\\
=&\frac{1}{n_k}\mathrm{tr}\left\{\left(\mathbf{I}_{K-1}\otimes(\otimes_{m=3}^{1}\widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)\left(\otimes_{m=4}^{1}\boldsymbol{\Sigma}_m\right)\right\}\\
=&\frac{1}{n_k}\mathrm{tr}\left\{\boldsymbol{\Sigma}_4\otimes\left(\otimes_{m=3}^{1}\widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T\boldsymbol{\Sigma}_m\right)\right\}=\frac{1}{n_k}\mathrm{tr}(\boldsymbol{\Sigma}_4)\prod_{m=1}^{3}\mathrm{tr}\left(\widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T\boldsymbol{\Sigma}_m\right)\\
=&\frac{K-1}{n_k}\prod_{m=1}^{3}\mathrm{tr}\left(\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T\boldsymbol{\Sigma}_m\widehat{\mathbf{A}}_m\right)\le\frac{K-1}{n_k}\prod_{m=1}^{3}r_m\left\|\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T\boldsymbol{\Sigma}_m\widehat{\mathbf{A}}_m\right\|^2\\
\le&\frac{r_1r_2r_3(K-1)C_{\boldsymbol{\Sigma}}^6}{n_k}\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\boldsymbol{\Sigma}}\right)^6.
\end{aligned}
$$

Denote $\underline{\mathbf{F}}=\boldsymbol{\Sigma}^{1/2}\left(\mathbf{I}_{K-1}\otimes(\otimes_{m=3}^{1}\widehat{\mathbf{A}}_m\widehat{\mathbf{D}}_m^T\widehat{\mathbf{D}}_m\widehat{\mathbf{A}}_m^T)\right)\boldsymbol{\Sigma}^{1/2}$. By Lemma 11,

$$
\begin{aligned}
&P\left(\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i^T\right)\underline{\mathbf{F}}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i\right)>\frac{r_1r_2r_3(K-1)C_{\boldsymbol{\Sigma}}^6}{n_k}\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\boldsymbol{\Sigma}}\right)^6+x\right)\\
\le&P\left(\left\|\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i^T\right)\underline{\mathbf{F}}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i\right)\right\|_F^2-\mathrm{E}\left\|\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i^T\right)\underline{\mathbf{F}}\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\underline{\mathbf{Z}}_i\right)\right\|_F^2>x\right)\\
\le&\exp\left\{-c\min\left(\frac{2n_k^2x^2}{K(K-1)C_{\boldsymbol{\Sigma}}^6c_0^4\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\boldsymbol{\Sigma}}\sqrt{r}\right)^{12}},\frac{\sqrt{2}n_kx}{C_{\boldsymbol{\Sigma}}^3\sqrt{K}c_0^2\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\boldsymbol{\Sigma}}\right)^6}\right)\right\}.
\end{aligned}
$$

Set $x=\frac{2srp\log p}{n\eta^2}$. Then,

$$
\begin{aligned}
&P\left(I_4\le\sqrt{\frac{r_1r_2r_3(K-1)C_{\boldsymbol{\Sigma}}^6}{2n_k}\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\boldsymbol{\Sigma}}\right)^6+\frac{srp\log p}{n\eta^2}}\right)\\
>&1-\exp\left\{-c\min\left(\frac{2\left(\frac{2srp\log p}{\eta^2}\right)^2}{K(K-1)C_{\boldsymbol{\Sigma}}^6c_0^4\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\boldsymbol{\Sigma}}\sqrt{r}\right)^{12}},\frac{\sqrt{2}\frac{2srp\log p}{n\eta^2}}{C_{\boldsymbol{\Sigma}}^3\sqrt{K}c_0^2\left(\frac{C_2}{\eta}\sqrt{\frac{srp\log p}{n}}+C_{\boldsymbol{\Sigma}}\right)^6}\right)\right\}.
\end{aligned}
$$

Consequently,

$$\left\|\widehat{\mathbf{B}}-\mathbf{B}\right\|_F \leq I_1+I_2+I_3+I_4 \leq C\sqrt{\frac{srp\log p}{n\eta^2}}$$

with probability at least $1-O\left(p^{-1}\right)$.

Proof of part (b)

Under conditions (C1) to (C5), we have (S5.88) and hence

$$|(\widehat{A}_{mj}-\widehat{A}_{1j})-(A_{mj}-A_{1j})| \leq |\widehat{A}_{mj}-A_{mj}|+|\widehat{A}_{1j}-A_{1j}| \leq 2\left\|\widehat{\mathbf{A}}_m-\mathbf{A}_m\right\|_{\max} \leq C\sqrt{\frac{p\log p}{n\eta^2}}$$

with probability at least $1-O(p^{-1})$. Furthermore, according to Lemma A.2 in Min and Mai (2022),

$$|\widehat{\sigma}_{m,ij}-\sigma_{m,ij}| \leq \left\|\widehat{\mathbf{\Sigma}}_m-\mathbf{\Sigma}_m\right\|_{\max} \lesssim \sqrt{\frac{\log p}{np^2}}$$

with probability at least $1-O(p^{-1})$. Let $\epsilon=\sqrt{\frac{p\log p}{n\eta^2}}$, then by Lemma 29, we have $\widehat{\mathcal{S}}_m=\mathcal{S}_m$ with probability at least $1-O(p^{-1})$ if there exist constants ψ_1,ψ_2 such that $\psi_1\sqrt{\frac{p\log p}{n\eta^2}}<\lambda<\min\{\frac{D_{m,\min}}{8\phi},\psi_2(1-\kappa_m)\}$.

Proof of Theorem 3

Denote $\widetilde{s}=\prod_{m=1}^{M}s_m$. By Lemmas 26 and 27, we know that

$$\|\mathrm{vec}(\widehat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)\|_{2,\widetilde{s}} \leq C\sqrt{\frac{\widetilde{s}\log p}{n}}. \tag{S5.92}$$

The misclassification risk $R_{\boldsymbol{\theta}}(\widehat{C})$ equals to

$$R_{\boldsymbol{\theta}}(\widehat{C})=\pi_1 P_{\boldsymbol{\theta}}(\widehat{C}(\mathbf{X})=2|\mathrm{label}(\mathbf{X})=1)+\pi_2 P_{\boldsymbol{\theta}}(\widehat{C}(\mathbf{X})=1|\mathrm{label}(\mathbf{X})=2). \tag{S5.93}$$

Denote $\Delta = \sqrt{\text{vec}^T(\mathbf{B}_2)\mathbf{\Sigma}\text{vec}(\mathbf{B}_2)}$ and $\widehat{\Delta} = \sqrt{\text{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{\Sigma}\text{vec}(\widehat{\mathbf{B}}_2)}$. Then, $\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\mathbf{X} - \frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}\right) + \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1} \sim N\left(\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\boldsymbol{\mu}_k - \frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}\right) + \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}, \widehat{\Delta}^2\right)$, and

$$
\begin{aligned}
&P\left(\widehat{C}(\mathbf{X}) = 2|\text{label}(\mathbf{X}) = 1\right)\\
=&P\left(\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1} + \text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\mathbf{X} - \frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}\right) > 0\middle|\text{vec}(\mathbf{X}) \sim N(\text{vec}(\boldsymbol{\mu}_1), \mathbf{\Sigma})\right)\\
=&1 - \Phi\left(-\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\boldsymbol{\mu}_1 - \frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}\right) + \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right) = \Phi\left(\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\boldsymbol{\mu}_1 - \frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}\right) + \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right)\\
=&\Phi\left(-\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2} - \boldsymbol{\mu}_1\right) - \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right),
\end{aligned}
$$

$$
\begin{aligned}
&P\left(\widehat{C}(\mathbf{X}) = 1|\text{label}(\mathbf{X}) = 2\right)\\
=&P\left(\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1} + \text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\mathbf{X} - \frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}\right) \le 0\middle|\text{vec}(\mathbf{X}) \sim N(\text{vec}(\boldsymbol{\mu}_2), \mathbf{\Sigma})\right)\\
=&\Phi\left(\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\boldsymbol{\mu}_2 - \frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}\right) + \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right) = 1 - \Phi\left(-\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2} - \boldsymbol{\mu}_2\right) - \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right)\\
=&\bar{\Phi}\left(-\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2} - \boldsymbol{\mu}_2\right) - \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right),
\end{aligned}
$$

where $\bar{\Phi}(x) = 1 - \Phi(x)$ and $\Phi(x)$ is the cumulative density function for $N(0,1)$. It follows that

$$
R_{\boldsymbol{\theta}}(\widehat{C}) = \pi_1\Phi\left(-\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2} - \boldsymbol{\mu}_1\right) - \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right) + \pi_2\bar{\Phi}\left(-\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2} - \boldsymbol{\mu}_2\right) - \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right).
$$

Replacing the parameter estimates with corresponding true values yields the misclassification risk of the oracle Bayes rule,

$$
\begin{aligned}
R_{opt}(\boldsymbol{\theta}) &= \pi_1\Phi\left(-\frac{\text{vec}^T(\mathbf{B}_2)\text{vec}\left(\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2} - \boldsymbol{\mu}_1\right) - \log\frac{\pi_2}{\pi_1}}{\Delta}\right) + \pi_2\bar{\Phi}\left(-\frac{\text{vec}^T(\mathbf{B}_2)\text{vec}\left(\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2} - \boldsymbol{\mu}_2\right) - \log\frac{\pi_2}{\pi_1}}{\Delta}\right)\\
&= \pi_1\Phi\left(\frac{-\frac{\Delta^2}{2} + \log\frac{\pi_2}{\pi_1}}{\Delta}\right) + \pi_2\bar{\Phi}\left(\frac{\frac{\Delta^2}{2} + \log\frac{\pi_2}{\pi_1}}{\Delta}\right).
\end{aligned}
$$

To bound $R_{\boldsymbol{\theta}}(\widehat{C}) - R_{opt}(\boldsymbol{\theta})$, we introduce an intermediate quantity,

$$R^* = \pi_1 \Phi\left(\frac{-\frac{1}{2}\mathrm{vec}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2) + \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right) + \pi_2 \bar{\Phi}\left(\frac{\frac{1}{2}\mathrm{vec}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2) + \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right), \tag{S5.94}$$

and complete the proof in two steps which bound $R_{\boldsymbol{\theta}}(\widehat{C}) - R^*$ and $R^* - R_{opt}(\boldsymbol{\theta})$ respectively.

Denote $\delta_n = \|\widehat{\mathbf{B}}_2 - \mathbf{B}_2\|_F \vee \|\mathrm{vec}(\widehat{\boldsymbol{\mu}}_1) - \mathrm{vec}(\boldsymbol{\mu}_1)\|_{2,\widetilde{s}} \vee \|\mathrm{vec}(\widehat{\boldsymbol{\mu}}_2) - \mathrm{vec}(\boldsymbol{\mu}_2)\|_{2,\widetilde{s}}$.

Step 1 We aim to show that

$$R^* - R_{opt}(\boldsymbol{\theta}) \lesssim \Delta\delta_n^2 \exp\left\{-\frac{\Delta^2}{8}\right\}. \tag{S5.95}$$

Applying Taylor's expansion to the two terms of R^* at points $\frac{-\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}$ and $\frac{\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}$ respectively yields

$$\begin{aligned}
&\Phi\left(\frac{-\frac{1}{2}\mathrm{vec}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2) + \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right) \\
=&\Phi\left(\frac{-\frac{\Delta^2}{2} + \log\frac{\pi_2}{\pi_1}}{\Delta}\right) + \Phi'\left(\frac{-\frac{\Delta^2}{2} + \log\frac{\pi_2}{\pi_1}}{\Delta}\right)\left[\frac{\Delta}{2} - \frac{\mathrm{vec}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}} + \log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta}\right)\right] \\
&+ \frac{1}{2}\Phi''(t_{1,n})\left[\frac{\Delta}{2} - \frac{\mathrm{vec}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}} + \log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta}\right)\right]^2,
\end{aligned} \tag{S5.96}$$

and

$$\begin{aligned}
&\bar{\Phi}\left(\frac{\frac{1}{2}\mathrm{vec}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2) + \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right) \\
=&\bar{\Phi}\left(\frac{\frac{\Delta^2}{2} + \log\frac{\pi_2}{\pi_1}}{\Delta}\right) + \bar{\Phi}'\left(\frac{\frac{\Delta^2}{2} + \log\frac{\pi_2}{\pi_1}}{\Delta}\right)\left[-\frac{\Delta}{2} + \frac{\mathrm{vec}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}} + \log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta}\right)\right] \\
&+ \frac{1}{2}\bar{\Phi}''(t_{2,n})\left[-\frac{\Delta}{2} + \frac{\mathrm{vec}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}} + \log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta}\right)\right]^2,
\end{aligned} \tag{S5.97}$$

where $t_{1,n}$ is some constant between $\frac{-\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}$ and $\frac{-\frac{1}{2}\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)+\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}$ and $t_{2,n}$ is some constant between $\frac{\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}$ and $\frac{\frac{1}{2}\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)+\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}$. Since

$$\frac{-\frac{1}{2}\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)+\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}} \to \frac{-\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta} \quad \text{as} \quad \delta_n \to 0,$$

we have

$$|\Phi''(t_{1,n})| \asymp \left|\Phi''\left(\frac{-\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}\right)\right| \asymp \left|\left(\frac{\frac{\Delta^2}{2}-\log\frac{\pi_2}{\pi_1}}{\Delta}\right)\exp\left\{-\frac{1}{2}\left(\frac{-\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}\right)^2\right\}\right| \asymp \left|\frac{\Delta}{2}\exp\left\{-\frac{\Delta^2}{8}\right\}\right|. \tag{S5.98}$$

Let $\boldsymbol{\gamma}=\boldsymbol{\Sigma}^{1/2}\mathrm{vec}(\mathbf{B}_2)$ and $\widehat{\boldsymbol{\gamma}}=\boldsymbol{\Sigma}^{1/2}\mathrm{vec}(\widehat{\mathbf{B}}_2)$. Then,

$$\begin{aligned}\left|\Delta-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{\widehat{\Delta}}\right| &= \left|\|\boldsymbol{\gamma}\|_2-\frac{\boldsymbol{\gamma}^T\widehat{\boldsymbol{\gamma}}}{\|\widehat{\boldsymbol{\gamma}}\|_2}\right| = \left|\frac{\|\boldsymbol{\gamma}\|_2\cdot\|\widehat{\boldsymbol{\gamma}}\|_2-\boldsymbol{\gamma}^T\widehat{\boldsymbol{\gamma}}}{\|\widehat{\boldsymbol{\gamma}}\|_2}\right| \\ &\lesssim \frac{1}{\Delta}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_2^2 \lesssim \frac{1}{\Delta}\|\widehat{\mathbf{B}}_2-\mathbf{B}_2\|_F^2 \lesssim \frac{1}{\Delta}\delta_n^2, \end{aligned} \tag{S5.99}$$

where the first inequality in (S5.99) holds due to Lemma 28. Besides,

$$\left|\widehat{\Delta}-\Delta\right| \le \sqrt{\mathrm{vec}^T(\widehat{\mathbf{B}}_2-\mathbf{B}_2)\boldsymbol{\Sigma}\mathrm{vec}(\widehat{\mathbf{B}}_2-\mathbf{B}_2)} \lesssim \left\|\widehat{\mathbf{B}}_2-\mathbf{B}_2\right\|_F \lesssim \delta_n. \tag{S5.100}$$

Combining (S5.99) and (S5.100), when $\delta_n=o(1)$,

$$\begin{aligned}&\frac{\Delta}{2}-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}+\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right) \\ \le&\left|\frac{\Delta}{2}-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}\right|+\left|\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right| \lesssim \frac{1}{\Delta}\delta_n^2+\delta_n \lesssim \delta_n. \end{aligned} \tag{S5.101}$$

It follows that

$$\frac{1}{2}\Phi''(t_{1,n})\left[\frac{\Delta}{2}-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}+\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right]^2 \lesssim \Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}.$$

In a similar fashion,

$$\frac{1}{2}\bar{\Phi}''(t_{2,n})\left[-\frac{\Delta}{2}+\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}+\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right]^2\lesssim\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}.$$

Thus,

$$\begin{aligned}
&R^*-R_{opt}(\boldsymbol{\theta})\\
=&\pi_1\Phi'\left(\frac{-\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}\right)\left[\frac{\Delta}{2}-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}+\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right]\\
&+\pi_2\Phi'\left(\frac{\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}\right)\left[\frac{\Delta}{2}-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}-\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right]+O\left(\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{R^*-R_{opt}(\boldsymbol{\theta})}{\sqrt{\pi_1\pi_2}}\\
=&\exp\left\{-\frac{1}{2}\left(\frac{-\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}\right)^2-\frac{1}{2}\log\frac{\pi_2}{\pi_1}\right\}\left[\frac{\Delta}{2}-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}+\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right]\\
&+\exp\left\{-\frac{1}{2}\left(\frac{\frac{\Delta^2}{2}+\log\frac{\pi_2}{\pi_1}}{\Delta}\right)^2+\frac{1}{2}\log\frac{\pi_2}{\pi_1}\right\}\left[\frac{\Delta}{2}-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}-\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right]\\
&+O\left(\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}\right)\\
=&\exp\left\{-\frac{\Delta^2}{8}+\frac{1}{2\Delta^2}\left(\log\frac{\pi_2}{\pi_1}\right)^2\right\}\left[\frac{\Delta}{2}-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}\right]+O\left(\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}\right)\\
\lesssim&\exp\left\{-\frac{\Delta^2}{8}\right\}\left|\frac{\Delta}{2}-\frac{\mathrm{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\mathrm{vec}(\widehat{\mathbf{B}}_2)}{2\widehat{\Delta}}\right|+O\left(\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}\right)\\
\lesssim&\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}.
\end{aligned}$$

Hence (S5.95) is proved.

Step 2 We aim to show that

$$R_{\boldsymbol{\theta}}(\widehat{C})-R^*\lesssim\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}. \tag{S5.102}$$

Similar to step 1, applying Taylor's expansion to $R_{\boldsymbol{\theta}}(\widehat{C})$ at $\frac{-\frac{1}{2}\text{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\text{vec}(\widehat{\mathbf{B}}_2)+\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}$ and $\frac{\frac{1}{2}\text{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\text{vec}(\widehat{\mathbf{B}}_2)+\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}$ respectively yields

$$
\begin{aligned}
&R_{\boldsymbol{\theta}}(\widehat{C})-R^* \\
=&\pi_1\Phi'\left(\frac{-\frac{1}{2}\text{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\text{vec}(\widehat{\mathbf{B}}_2)+\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right)\frac{-\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)+\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}-\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}} \\
&-\pi_2\Phi'\left(\frac{\frac{1}{2}\text{vec}^T(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\text{vec}(\widehat{\mathbf{B}}_2)+\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right)\frac{-\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)+\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}-\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}} \\
&+O\left(\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}\right). \qquad \text{(S5.103)}
\end{aligned}
$$

As $n\to\infty$, we have $\left|\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}-\log\frac{\pi_2}{\pi_1}\right|\to 0$ and

$$
\begin{aligned}
&\left|\frac{-\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)+\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}-\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right|\lesssim\left|\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\frac{\widehat{\boldsymbol{\mu}}_1+\widehat{\boldsymbol{\mu}}_2}{2}-\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)}{\Delta}\right| \\
\lesssim&\frac{1}{\Delta}\left[\left\|\widehat{\mathbf{B}}_2-\mathbf{B}_2\right\|_F\cdot\left\|\text{vec}\left(\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1+\widehat{\boldsymbol{\mu}}_2-\boldsymbol{\mu}_2\right)\right\|_{2,\widetilde{s}}+\left|\text{vec}^T(\widehat{\mathbf{B}}_2)\text{vec}\left(\widehat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1+\widehat{\boldsymbol{\mu}}_2-\boldsymbol{\mu}_2\right)\right|\right] \\
\lesssim&\delta_n.
\end{aligned}
$$

Then, denote $\mathbf{v}=\text{vec}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)$. We have

$$
\begin{aligned}
&\frac{R_{\boldsymbol{\theta}}(\widehat{C})-R^*}{\sqrt{\pi_1\pi_2}} \\
\lesssim&\delta_n\left|\exp\left\{-\frac{\left(\frac{1}{2}\text{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}-\log\frac{\pi_2}{\pi_1}\right)^2}{2\widehat{\Delta}^2}-\frac{1}{2}\log\frac{\pi_2}{\pi_1}\right\}-\exp\left\{-\frac{\left(\frac{1}{2}\text{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}+\log\frac{\pi_2}{\pi_1}\right)^2}{2\widehat{\Delta}^2}+\frac{1}{2}\log\frac{\pi_2}{\pi_1}\right\}\right| \\
&+O\left(\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}\right) \\
=&\delta_n\exp\left\{-\frac{\left(\frac{1}{2}\text{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}\right)^2+\left(\log\frac{\pi_2}{\pi_1}\right)^2}{2\widehat{\Delta}^2}\right\}\cdot \\
&\left|\exp\left\{\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}\log\frac{\pi_2}{\pi_1}}{2\widehat{\Delta}^2}-\frac{1}{2}\log\frac{\pi_2}{\pi_1}\right\}-\exp\left\{-\frac{\text{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}\log\frac{\pi_2}{\pi_1}}{2\widehat{\Delta}^2}+\frac{1}{2}\log\frac{\pi_2}{\pi_1}\right\}\right|+O\left(\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}\right).
\end{aligned}
$$

Note that $\frac{\left(\frac{1}{2}\mathrm{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}\right)^2+\left(\log\frac{\pi_2}{\pi_1}\right)^2}{2\widehat{\Delta}^2} \to \frac{\Delta^2}{8}+\frac{\left(\log\frac{\pi_2}{\pi_1}\right)^2}{2\Delta^2}$ as $\delta_n \to 0$ and $\exp\{-\frac{\Delta^2}{8}-\frac{\left(\log\frac{\pi_2}{\pi_1}\right)^2}{2\Delta^2}\} \lesssim \exp\{-\frac{\Delta^2}{8}\}$. Besides,

$$\frac{\mathrm{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}\log\frac{\pi_2}{\pi_1}}{2\widehat{\Delta}^2}-\frac{1}{2}\log\frac{\pi_2}{\pi_1}=\frac{1}{2}\log\frac{\pi_2}{\pi_1}\left(\frac{\mathrm{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}}{\widehat{\Delta}^2}-1\right)\lesssim\frac{1}{\Delta}\left|\frac{\mathrm{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}}{\widehat{\Delta}}-\Delta\right|\lesssim\frac{\delta_n^2}{\Delta^2}.$$

Let $x=\frac{\mathrm{vec}^T(\widehat{\mathbf{B}}_2)\mathbf{v}\log\frac{\pi_2}{\pi_1}}{2\widehat{\Delta}^2}-\frac{1}{2}\log\frac{\pi_2}{\pi_1}$. Then, $|\exp\{x\}-\exp\{-x\}|\lesssim x\lesssim\frac{\delta_n^2}{\Delta^2}$. Consequently,

$$\frac{R_{\boldsymbol{\theta}}(\widehat{C})-R^*}{\sqrt{\pi_1\pi_2}}\lesssim\delta_n\exp\left\{-\frac{\Delta^2}{8}\right\}\frac{\delta_n^2}{\Delta^2}+O\left(\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}\right)\lesssim\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}.$$

Next, combining (S5.95) and (S5.102) gives

$$R_{\boldsymbol{\theta}}(\widehat{C})-R_{opt}(\boldsymbol{\theta})=R_{\boldsymbol{\theta}}(\widehat{C})-R^*+R^*-R_{opt}(\boldsymbol{\theta})\lesssim\Delta\delta_n^2\exp\left\{-\frac{\Delta^2}{8}\right\}.$$

In Theorem 2, we have proved that $\|\widehat{\mathbf{B}}_2-\mathbf{B}_2\|_F\lesssim\sqrt{\frac{srp\log p}{n\eta^2}}$ with probability at least $1-O(p^{-1})$. According to the definition of δ_n, we have

$$R_{\boldsymbol{\theta}}(\widehat{C})-R_{opt}(\boldsymbol{\theta})\lesssim\frac{(s^3\vee srp)\log p}{n\eta^2} \tag{S5.104}$$

with probability at least $1-O(p^{-1})$.

S6 Review of Tucker Decomposition

In this section, we present the review of Tucker decomposition and provide two well established algorithms to implement the decomposition. Tucker decomposition is a widely used decomposition method for tensors. It was first introduced in Tucker (1963) and then more formally discussed as a three-mode factor analysis method in Tucker (1966). Although Tucker decomposition was originally proposed for three-way tensors, it generalizes well to M-way tensors, $M\geq 3$. For an M-way tensor $\mathbf{A}\in\mathbb{R}^{p_1\times\cdots\times p_M}$, its Tucker decomposition

could be written as

$$\mathbf{A} = \mathbf{C} \times_1 \mathbf{U}_1 \times_2 \cdots \times_M \mathbf{U}_M = [\![\mathbf{C}; \mathbf{U}_1, ..., \mathbf{U}_M]\!] \tag{S6.105}$$

where $\mathbf{C} \in \mathbb{R}^{q_1 \times \cdots \times q_M}$ is called the core tensor and $\mathbf{U}_m \in \mathbb{R}^{p_m \times q_m}, m = 1, ..., M$ are factor matrices.

Note that Tucker decomposition (S6.105) is not identifiable. Given nonsingular matrices $\mathbf{O}_m \in \mathbb{R}^{q_m \times q_m}$, it is easy to find

$$\mathbf{C} \times_1 \mathbf{U}_1 \cdots \times_M \mathbf{U}_M = (\mathbf{C} \times_1 \mathbf{O}_1 \cdots \times_M \mathbf{O}_M) \times_1 \mathbf{U}_1\mathbf{O}_1^{-1} \cdots \times_M \mathbf{U}_M\mathbf{O}_M^{-1}, \tag{S6.106}$$

which implies that the core tensor and factor matrices are not unique. To solve the identifiability issue, we require the factor matrices to be orthogonal unless otherwise specified. It is easy to find that if $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ admits a Tucker decomposition as in (S6.105), its number of free parameters would be $\sum_{m=1}^M p_m q_m + \prod_{m=1}^M q_m - \sum_{m=1}^M q_m^2$, in which $-\sum_{m=1}^M q_m^2$ adjusts for the nonsingular indeterminacy.

If $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ could be decomposed as in (S6.105), then the following result holds as well,

$$\mathbf{A}_{(m)} = \mathbf{U}_m\mathbf{C}_{(m)}(\mathbf{U}_M \otimes \cdots \otimes \mathbf{U}_{m+1} \otimes \mathbf{U}_{m-1} \otimes \cdots \otimes \mathbf{U}_1)^{\mathrm{T}}, \quad m = 1, ..., M, \tag{S6.107}$$

where $\mathbf{A}_{(m)}$ and $\mathbf{C}_{(m)}$ represent the matricization of tensor $\mathbf{A}$ and $\mathbf{C}$ along mode m, and $\otimes$ denotes the Kronecker product. $\mathbf{A}$ is said to have Tucker rank $\boldsymbol{r} = (r_1, ..., r_M)$ if $r_m = \mathrm{rank}(\mathbf{A}_{(m)}), m = 1, ..., M$. When $r_m < p_m, m = 1, ..., M$, we say that $\mathbf{A}$ has a Tucker low-rank structure. Considering (S6.107) and the identifiability constraint, $\mathbf{A}$ naturally admits a Tucker decomposition with $\mathbf{C} \in \mathbb{R}^{r_1 \times \cdots \times r_M}$ and $\mathbf{U}_m \in \mathbb{O}^{p_m \times r_m}, m = 1, ..., M$ if it is of Tucker rank $\boldsymbol{r} = (r_1, ..., r_M)$.

Given a tensor $\mathbf{A}$ and its rank $\boldsymbol{r}$, the Tucker decomposition of $\mathbf{A}$ can be obtained by solving the problem as follows,

$$\min_{\substack{\mathbf{C}\in\mathbb{R}^{r_1\times\cdots\times r_M}\\ \mathbf{U}_m\in\mathbb{O}^{p_m\times r_m},m=1,\ldots,M}} \|\mathbf{A}-\mathbf{C}\times_1\mathbf{U}_1\cdots\times_M\mathbf{U}_M\|_{\mathrm{F}}^2, \tag{S6.108}$$

which is equivalent to

$$\max_{\mathbf{U}_m\in\mathbb{O}^{p_m\times r_m},m=1,\ldots,M} \|\mathbf{A}\times_1\mathbf{U}_1^{\mathrm{T}}\cdots\times_M\mathbf{U}_M^{\mathrm{T}}\|_{\mathrm{F}}^2. \tag{S6.109}$$

Algorithms to solve this optimization problem have been well studied. The most popular two methods are Higher-order SVD (HOSVD) (Tucker 1966; De Lathauwer et al. 2000a) and Higher-order Orthogonal Iteration (HOOI) (De Lathauwer et al., 2000b).

HOSVD adopts the first r_m left singular vectors of $\mathbf{A}_{(m)}$ as its factor matrix along mode-m, and calculates the core tensor after obtaining all factor matrices. The estimation procedure of HOSVD is outlined in Algorithm S4. Based on HOSVD, HOOI further improves the estimation accuracy by using an iterative alternative least squares (ALS) method to further refine the obtained factor matrices and the core tensor. Steps in HOOI are summerized in Algorithm S5.

Algorithm S4 Higher-order Singular Value Decomposition (HOSVD)

1. Input $\mathbf{A}$, rank $r=(r_1,\ldots,r_M)$.
2. For $m=1,\ldots,M$, do:
   $\widehat{\mathbf{U}}_m = r_m$ leading left singular vectors of $\mathbf{A}_{(m)}$.
3. Calculate $\widehat{\mathbf{C}}=\mathbf{A}\times_1\widehat{\mathbf{U}}_1^{\mathrm{T}}\cdots\times_M\widehat{\mathbf{U}}_M$.
4. Output $\widehat{\mathbf{C}},\widehat{\mathbf{U}}_1,\ldots,\widehat{\mathbf{U}}_M$.

S7 Technical Lemmas

Lemma 1. *For any full rank matrix $\mathbf{X}\in\mathbb{R}^{m\times n}$, $m>n$, if its SVD is $\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^T$ where $\mathbf{U}\in\mathbb{O}_{m\times n}$, and $\mathbf{V}\in\mathbb{O}_n$, $\mathbf{D}\in\mathbb{R}^{n\times n}$ is the diagonal matrix containing the singular values*

Algorithm S5 Higher-order Orthogonal Iteration (HOOI)

1. Input $\mathbf{A}$, rank $r = (r_1, ..., r_M)$.
2. Initialize $\widehat{\mathbf{U}}_m^{(0)}, m = 1, ..., M$, with HOSVD.
3. While the stopping criterion is not satisfied, do:
   For $m = 1, ..., M$, do:
   $\widetilde{\mathbf{A}} = \mathbf{A} \times_1 \widehat{\mathbf{U}}_1^{\mathrm{T}} \cdots \times_{m-1} \widehat{\mathbf{U}}_{m-1}^{\mathrm{T}} \times_{m+1} \widehat{\mathbf{U}}_{m+1}^{\mathrm{T}} \cdots \times_M \widehat{\mathbf{U}}_M^{\mathrm{T}}$;
   $\widehat{\mathbf{U}}_m = r_m$ leading left singular vectors of $\widetilde{\mathbf{A}}_{(m)}$.
4. Calculate $\widehat{\mathbf{C}} = \mathbf{A} \times_1 \widehat{\mathbf{U}}_1^{\mathrm{T}} \cdots \times_M \widehat{\mathbf{U}}_M^{\mathrm{T}}$.
5. Output $\widehat{\mathbf{C}}, \widehat{\mathbf{U}}_1, ..., \widehat{\mathbf{U}}_M$.

of $\mathbf{X}$, $\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\| = \sigma_{\min}^{-1}(\mathbf{X})$.

Proof. $\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\| = \|(\mathbf{V}\mathbf{D}^2\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\| = \|\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\| = \sigma_{\min}^{-1}(\mathbf{X})$. □

Lemma 2 (Lemma 2. Pan et al. 2019b)**.** *If* $\mathbf{W} \sim TN(\mathbf{0}, \mathbf{\Omega}_1, ..., \mathbf{\Omega}_M)$, *then* $\mathrm{E}[\mathcal{M}_m(\mathbf{W})\mathcal{M}_m^T(\mathbf{W})] = \mathbf{\Omega}_m \cdot \prod_{l \neq m} tr(\mathbf{\Omega}_l)$, $m = 1, ..., M$.

Lemma 3. *For* $\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}} \in \mathbb{R}^{p \times p}$ *and* $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\boldsymbol{\beta}\|_\infty \leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}\|\boldsymbol{\beta}\|_2 \leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}\|\boldsymbol{\beta}\|_1.$$

Lemma 4. *For* $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$, $\mathbf{Y} \in \mathbb{R}^{p_2 \times p_3}$, $r = rank(\mathbf{X})$, *then*

$$\begin{aligned} &\|\mathbf{X}\| \leq \|\mathbf{X}\|_F \leq \sqrt{r}\|\mathbf{X}\|, \quad \|\mathbf{X}\|_{\max} \leq \|\mathbf{X}\| \leq \sqrt{p_1 p_2}\|\mathbf{X}\|_{\max}, \\ &\|\mathbf{X}\mathbf{Y}\|_F \leq \|\mathbf{X}\|_F\|\mathbf{Y}\|_F, \quad \sigma_{\min}(\mathbf{X})\|\mathbf{Y}\|_F \leq \|\mathbf{X}\mathbf{Y}\|_F \leq \|\mathbf{X}\|\|\mathbf{Y}\|_F \\ &\sigma_{\min}(\mathbf{X})\|\mathbf{Y}\| \leq \|\mathbf{X}\mathbf{Y}\| \leq \|\mathbf{X}\|\|\mathbf{Y}\|. \end{aligned} \tag{S7.110}$$

Lemma 5 (Lemma 4. Zhang and Xia 2018)**.** *Suppose we have* $\mathbf{U}_1 \in \mathbb{R}^{p_1 \times r_1}$ *and* $\mathbf{U}_2 \in \mathbb{R}^{p_2 \times r_2}$, *then*

$$\|\mathbf{U}_1 \otimes \mathbf{U}_2\| = \|\mathbf{U}_1\| \cdot \|\mathbf{U}_2\|, \quad \|\mathbf{U}_1 \otimes \mathbf{U}_2\|_F = \|\mathbf{U}_1\|_F \cdot \|\mathbf{U}_2\|_F, \quad \sigma_{\min}(\mathbf{U}_1 \otimes \mathbf{U}_2) = \sigma_{\min}(\mathbf{U}_1)\sigma_{\min}(\mathbf{U}_2) \tag{S7.111}$$

Lemma 6 (Lemma 2.5. & 2.6. Chen et al. 2021)**.** *For any* $\mathbf{U}, \widetilde{\mathbf{U}} \in \mathbb{O}_{p\times r}$*, one has*

$$\|\mathbf{U}\mathbf{U}^T - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T\| = \|\sin\Theta(\mathbf{U}, \widetilde{\mathbf{U}})\| = \|\mathbf{U}_\perp^T\widetilde{\mathbf{U}}\| = \|\mathbf{U}^T\widetilde{\mathbf{U}}_\perp\| = \sqrt{1-\sigma_{\min}^2(\widetilde{\mathbf{U}}^T\mathbf{U})}$$

$$\frac{1}{\sqrt{2}}\|\mathbf{U}\mathbf{U}^T - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T\|_F = \|\sin\Theta(\mathbf{U}, \widetilde{\mathbf{U}})\|_F = \|\mathbf{U}_\perp^T\widetilde{\mathbf{U}}\|_F = \|\mathbf{U}^T\widetilde{\mathbf{U}}_\perp\|_F = \sqrt{r-\|\widetilde{\mathbf{U}}^T\mathbf{U}\|_F^2}$$

$$\|\mathbf{U}\mathbf{U}^T - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T\| \le \min_{\mathbf{R}\in\mathbb{O}_{r\times r}}\|\mathbf{U}\mathbf{R} - \widetilde{\mathbf{U}}\| \le \sqrt{2}\|\mathbf{U}\mathbf{U}^T - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T\|$$

$$\frac{1}{\sqrt{2}}\|\mathbf{U}\mathbf{U}^T - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T\|_F \le \min_{\mathbf{R}\in\mathbb{O}_{r\times r}}\|\mathbf{U}\mathbf{R} - \widetilde{\mathbf{U}}\|_F \le \|\mathbf{U}\mathbf{U}^T - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T\|_F$$

Lemma 7 (Proposition 1. Cai and Zhang 2018)**.** *Suppose that* $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ *is a rank-*r *matrix with its full SVD decomposition being* $\mathbf{X} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^T$*, where* $\widetilde{\mathbf{U}} = [\mathbf{U}\ \mathbf{U}_\perp] \in \mathbb{O}_{p_1}$*,* $\mathbf{U} \in \mathbb{O}_{p_1\times r}, \mathbf{U}_\perp \in \mathbb{O}_{p_1\times(p_1-r)}$ *are the leading* r *and the last* $p_1 - r$ *left singular vectors,* $\widetilde{\mathbf{V}} = [\mathbf{V}\ \mathbf{V}_\perp] \in \mathbb{O}_{p_2}$*,* $\mathbf{V} \in \mathbb{O}_{p_2\times r}, \mathbf{V}_\perp \in \mathbb{O}_{p_2\times(p_2-r)}$ *are the leading* r *and the last* $p_2 - r$ *right singular vectors,* $\widetilde{\mathbf{D}} = \begin{bmatrix} \mathbf{D} & \mathbf{0}_{r\times(p_2-r)} \\ \mathbf{0}_{(p_1-r)\times r} & \mathbf{0}_{(p_1-r)\times(p_2-r)} \end{bmatrix} \in \mathbb{R}^{p_1\times p_2}$*, and* $\mathbf{D} = diag\{\sigma_1(\mathbf{X}), ..., \sigma_r(\mathbf{X})\}$*. Let* $\widetilde{\mathbf{W}} = [\mathbf{W}\ \mathbf{W}_\perp] \in \mathbb{O}_{p_2}$ *be any orthogonal matrix with* $\mathbf{W} \in \mathbb{O}_{p_1\times r}$ *and* $\mathbf{W}_\perp \in \mathbb{O}_{p_1\times(p_1-r)}$*. Given that* $\sigma_r(\mathbf{W}^T\mathbf{X}) > \sigma_{r+1}(\mathbf{X})$*, we have*

$$\|\sin\Theta(\mathbf{U}, \mathbf{W})\| \le \frac{\sigma_r(\mathbf{W}^T\mathbf{X})\left\|\mathbf{W}_\perp^T\mathbf{X}\mathbf{P}_{(\mathbf{W}^T\mathbf{X})^T}\right\|}{\sigma_r^2(\mathbf{W}^T\mathbf{X}) - \sigma_{r+1}^2(\mathbf{X})} \wedge 1. \tag{S7.112}$$

Lemma 8. *Let* $\mathbf{X} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{Y} \in \mathbb{R}^{n\times k}$*. Then*

$$\sigma_{\min}(\mathbf{X}\mathbf{Y}) \ge \sigma_{\min}(\mathbf{X})\sigma_{\min}(\mathbf{Y}), \tag{S7.113}$$

where $\sigma_{\min}(\mathbf{X})$ *is the minimum nonzero singular value of* $\mathbf{X}$*.*

Proof. By definition,

$$\sigma_{\min}(\mathbf{X}\mathbf{Y}) = \min_{\substack{\mathbf{u}\in\mathbb{R}^n\\ \mathbf{u}\neq\mathbf{0}}}\frac{\|\mathbf{X}\mathbf{Y}\mathbf{u}\|_2}{\|\mathbf{u}\|_2} = \min_{\substack{\mathbf{u}\in\mathbb{R}^n\\ \mathbf{u}\neq\mathbf{0}}}\frac{\|\mathbf{X}\mathbf{Y}\mathbf{u}\|_2}{\|\mathbf{Y}\mathbf{u}\|_2}\frac{\|\mathbf{Y}\mathbf{u}\|_2}{\|\mathbf{u}\|_2} \ge \min_{\substack{\mathbf{v}\in\mathbb{R}^n\\ \mathbf{v}\neq\mathbf{0}}}\frac{\|\mathbf{X}\mathbf{v}\|_2}{\|\mathbf{v}\|_2}\min_{\substack{\mathbf{u}\in\mathbb{R}^n\\ \mathbf{u}\neq\mathbf{0}}}\frac{\|\mathbf{Y}\mathbf{u}\|_2}{\|\mathbf{u}\|_2} = \sigma_{\min}(\mathbf{X})\sigma_{\min}(\mathbf{Y}).$$

□

Lemma 9 (Weyl's inequality for singular values, Lemma 2.3. Chen et al. 2021)**.** *For* $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^{p_1 \times p_2}$ *and every* $1 \leq i \leq \min\{p_1, p_2\}$*, the* i*-th largest singular values of* $\mathbf{X}$ *and* $\mathbf{X} + \mathbf{Z}$ *obey*

$$|\sigma_i(\mathbf{X} + \mathbf{Z}) - \sigma_i(\mathbf{X})| \leq \|\mathbf{Z}\|. \tag{S7.114}$$

Lemma 10 (Lemma 5. Cai and Zhang 2018)**.** *For any* $p \geq 1$*, denote* $\mathbb{B}^p = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_2 \leq 1\}$ *as the* p*-dimensional unit ball. Suppose that* $\mathbf{A} \in \mathbb{R}^{p_1 \times p_2}$ *is a random matrix. Then, we have for any* $t > 0$*,*

$$P\left(\|\mathbf{A}\| \geq 3t\right) \leq 7^{p_1+p_2} \cdot \max_{\substack{\mathbf{u}\in\mathbb{B}^{p_1} \\ \mathbf{v}\in\mathbb{B}^{p_2}}} P\left(|\mathbf{u}^T\mathbf{A}\mathbf{v}| \geq t\right). \tag{S7.115}$$

Lemma 11 (Theorem 1.1. Rudelson and Vershynin 2013, Hanson-Wright inequality)**.** *Let* $\mathbf{X} = (X_1, ..., X_p) \in \imath$ *be a random vector with independent components* X_i *which satisfy* $\mathrm{E}X_i = 0$ *and* $\|X_i\|_{\psi_2} \leq K$*. Let* $\mathbf{A}$ *be an* $p \times p$ *matrix. Then, for every* $t \geq 0$*,*

$$P\left\{\left|\mathbf{X}^T\mathbf{A}\mathbf{X} - \mathrm{E}\left[\mathbf{X}^T\mathbf{A}\mathbf{X}\right]\right| > x\right\} \leq 2\exp\left\{-c\min\left(\frac{x^2}{K^4\|\mathbf{A}\|_{HS}^2}, \frac{x}{K^2\|\mathbf{A}\|}\right)\right\}. \tag{S7.116}$$

Lemma 12 (Theorem 2.6.3. Vershynin 2018)**.** *Let* $X_1, ..., X_n$ *be independent, mean zero, sub-Gaussian random variables, and* $\mathbf{a} = (a_1, ..., a_n) \in \mathbb{R}^n$*. Then, for every* $t \geq 0$*, we have*

$$P\left\{\left|\sum_{i=1}^n a_i X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{K^2\|\mathbf{a}\|_2^2}\right) \tag{S7.117}$$

where $K = \max_i \|X_i\|_{\psi_2}$*.*

Lemma 13 (Corollary 5.35. Vershynin 2010)**.** *Let* $\mathbf{X}$ *be a* $p_1 \times p_2$*,* $p_1 \geq p_2$*, random matrix whose entries* X_{ij} *are independent standard normal random variables. Then, for any* $t \geq 0$*, we have*

$$\sqrt{p_1} - \sqrt{p_2} - t \leq \sigma_{p_2}(\mathbf{X}) \leq \|\mathbf{X}\| \leq \sqrt{p_1} + \sqrt{p_2} + t \tag{S7.118}$$

with probability at least $1 - 2\exp(-t^2/2)$.

Lemma 14 (Corollary of Lemma 13)**.** *Let* $\mathbf{A} \in \mathbb{R}^{p_1 \times p_2}$ *be a matrix with* $A_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $1 \le i \le p_1, 1 \le j \le p_2$. *Then for any* $t \ge 0$, *one has*

$$P\left(\|\mathbf{A}\| \ge \sigma(\sqrt{p_1} + \sqrt{p_2} + t)\right) \le \exp\left(-\frac{\sigma^2 t^2}{2}\right). \tag{S7.119}$$

Proof. By definition, $\|\mathbf{A}\| = \max_{\substack{\mathbf{u}\in\mathcal{S}^{p_2-1} \\ \mathbf{v}\in\mathcal{S}^{p_1-1}}} \langle \mathbf{A}\mathbf{u}, \mathbf{v}\rangle$, which is also the supremum of the Gaussian process $\mathbf{X}_{\mathbf{u},\mathbf{v}} = \langle \mathbf{A}\mathbf{u}, \mathbf{v}\rangle = \mathbf{v}^T\mathbf{A}\mathbf{u}$ indexed by $(\mathbf{u}, \mathbf{v}) \in \mathcal{S}^{p_2-1} \times \mathcal{S}^{p_1-1}$. Consider another process $\mathbf{Y}_{\mathbf{u},\mathbf{v}} = \langle \mathbf{g}, \mathbf{u}\rangle + \langle \mathbf{h}, \mathbf{v}\rangle$ where $\mathbf{g} \in \mathbb{R}^{p_2}$ and $\mathbf{h} \in \mathbb{R}^{p_1}$ are independent Gaussian random vectors with each element following $N(0, \sigma^2)$ identically and independently. Then, for any $(\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}') \in \mathcal{S}^{p_2-1} \times \mathcal{S}^{p_1-1}$,

$$\begin{aligned}
\mathrm{E}\left|\mathbf{X}_{\mathbf{u},\mathbf{v}} - \mathbf{X}_{\mathbf{u}',\mathbf{v}'}\right|^2 &= \mathrm{E}\left|\sum_{i=1}^{p_1}\sum_{j=1}^{p_2} A_{ij}u_jv_i - \sum_{i=1}^{p_1}\sum_{j=1}^{p_2} A_{ij}u'_jv'_i\right|^2 = \mathrm{E}\left|\sum_{i=1}^{p_1}\sum_{j=1}^{p_2} A_{ij}(u_jv_i - u'_jv'_i)\right|^2 \\
&= \mathrm{E}\sum_{i=1}^{p_1}\sum_{j=1}^{p_2} A_{ij}^2(u_jv_i - u'_jv'_i)^2 = \sigma^2\left(\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}(u_jv_i - u'_jv'_i)^2\right) \\
&\le \sigma^2\|\mathbf{u} - \mathbf{u}'\|_2^2 + \sigma^2\|\mathbf{v} - \mathbf{v}'\|_2^2 = \mathrm{E}\left|\mathbf{Y}_{\mathbf{u},\mathbf{v}} - \mathbf{Y}_{\mathbf{u}',\mathbf{v}'}\right|^2
\end{aligned}$$

Applying Lemma 5.33 in Vershynin (2010) results in

$$\mathrm{E}\|\mathbf{A}\| = \mathrm{E}\left[\max_{(\mathbf{u},\mathbf{v})} \mathbf{X}_{\mathbf{u},\mathbf{v}}\right] \le \mathrm{E}\left[\max_{(\mathbf{u},\mathbf{v})} \mathbf{Y}_{\mathbf{u},\mathbf{v}}\right] = \mathrm{E}\|\mathbf{g}\|_2 + \mathrm{E}\|\mathbf{h}\|_2 \le \sigma(\sqrt{p_1} + \sqrt{p_2}). \tag{S7.120}$$

By Proposition 5.34 in Vershynin (2010), we have

$$P\left(\|\mathbf{A}\| - \mathrm{E}\|\mathbf{A}\| \ge \sigma t\right) \le \exp(-\sigma^2 t^2/2). \tag{S7.121}$$

Combining (S7.120) and (S7.121), one has

$$P\left(\|\mathbf{A}\| \ge \sigma(\sqrt{p_1} + \sqrt{p_2} + t)\right) \le \exp(-\sigma^2 t^2/2).$$

□

Lemma 15 (Proposition 2, Koltchinskii et al. 2011)**.** *Let* $\mathbf{Z}_1, ..., \mathbf{Z}_n$ *be i.i.d. random matrices with dimensions* $p_1 \times p_2$ *that satisfy* $\mathrm{E}(\mathbf{Z}) = \mathbf{0}$*. Define*

$$\sigma_Z = \max\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\mathbf{Z}_i \mathbf{Z}_i^T) \right\|^{1/2}, \left\| \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\mathbf{Z}_i^T \mathbf{Z}_i) \right\|^{1/2} \right\},$$

and the ψ_α *norms of* $\|\mathbf{Z}\|$ *as*

$$U_Z^{(\alpha)} = \inf\left\{ u > 0 : \mathrm{E}\exp\left(\frac{\|\mathbf{Z}\|^\alpha}{u^\alpha} \right) \le 2 \right\}, \quad \alpha \ge 1.$$

Suppose that $U_Z^{(\alpha)} < \infty$ *for some* $\alpha \ge 1$*. Then there exists a constant* $C > 0$ *such that, for all* $t > 0$*, with probability at least* $1 - e^{-t}$*,*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i \right\| \le C \max\left\{ \sigma_Z \sqrt{\frac{\log(p_1+p_2)+t}{n}}, U_Z^{(\alpha)} \left(\log \frac{U_Z^{(\alpha)}}{\sigma_Z} \right)^{1/\alpha} \frac{\log(p_1+p_2)+t}{n} \right\}. \tag{S7.122}$$

Lemma 16 (Bernstein Inequality)**.** *For tensors* $\mathbf{Z}_i \in \mathbb{R}^{p_1 \times p_2 \times p_3 \times (K-1)}$ *which follows* $TN(\mathbf{0}, 2\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3, \mathbf{\Sigma}_4)$ *where* $\mathbf{\Sigma}_m \succ 0$*,* $C_{\mathbf{\Sigma}}^{-1} \le \sigma_{\min}(\mathbf{\Sigma}_m) \le \sigma_{\max}(\mathbf{\Sigma}_m) \le C_{\mathbf{\Sigma}}$*,* $m = 1, 2, 3$*, and* $\mathbf{\Sigma}_4 = CS(0.5)$*, we have the following tail bound*

$$P\left(\left\| \frac{1}{n}\sum_{i=1}^{n} \mathcal{M}_1(\mathbf{Z}_i) \right\| \ge C(\sqrt{p_1} + \sqrt{p_2 p_3 (K-1)}) \left(\sqrt{\frac{\log(p_1 + p_2 p_3 (K-1)) + t}{n}} + \frac{\log(p_1 + p_2 p_3 (K-1)) + t}{n} \right) \right)$$
$$\le \exp(-t). \tag{S7.123}$$

Proof. Let $\mathbf{E}_i \overset{iid}{\sim} TN(\mathbf{0}, \mathbf{I}_{p_1}, \mathbf{I}_{p_2}, \mathbf{I}_{p_3}, \mathbf{I}_{K-1})$. Then,

$$\begin{aligned} \|\mathcal{M}_1(\mathbf{Z}_i)\| &= \left\| \mathcal{M}_1(\mathbf{E}_i) \left(\sqrt{2}\mathbf{\Sigma}_4^{1/2} \otimes \mathbf{\Sigma}_3^{1/2} \otimes \mathbf{\Sigma}_2^{1/2} \otimes \mathbf{\Sigma}_1^{1/2} \right) \right\| \\ &\le \sqrt{2}\, \|\mathcal{M}_1(\mathbf{E}_i)\|\, \|\mathbf{\Sigma}_4\|^{1/2} \|\mathbf{\Sigma}_3\|^{1/2} \|\mathbf{\Sigma}_2\|^{1/2} \|\mathbf{\Sigma}_1\|^{1/2} \\ &\le C_{\mathbf{\Sigma}}^{3/2} \sqrt{K}\, \|\mathcal{M}_1(\mathbf{E}_i)\| \end{aligned}$$

By lemma 13, we know that

$$P(\|\mathcal{M}_1(\mathbf{E}_i)\| \geq (\sqrt{p_1} + \sqrt{p_2 p_3 (K-1)} + t)) \leq \exp(-t^2/2).$$

Let $u = C_0(\sqrt{p_1} + \sqrt{p_2 p_3 (K-1)})$ where $C_0 > 0$ is a large uniform constant. Then, for any $x \geq 1$,

$$\begin{aligned}
P\left(\frac{\|\mathcal{M}_1(\mathbf{Z}_i)\|}{u} \geq x\right) &\leq P\left(\frac{C_{\mathbf{\Sigma}}^{3/2}\sqrt{K}\|\mathcal{M}_1(\mathbf{E}_i)\|}{u} \geq x\right) \\
&= P\left(\|\mathcal{M}_1(\mathbf{E}_i)\| \geq \frac{C_0}{C_{\mathbf{\Sigma}}^{3/2}\sqrt{K}}(\sqrt{p_1} + \sqrt{p_2 p_3 (K-1)})x\right) \\
&\leq P\left(\|\mathcal{M}_1(\mathbf{E}_i)\| \geq \sqrt{p_1} + \sqrt{p_2 p_3 (K-1)} + \widetilde{C}_0 x\right) \\
&\leq \exp\left(-\frac{\widetilde{C}_0^2 x^2}{2}\right).
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathrm{E}\exp\left(\frac{\|\mathcal{M}_1(\mathbf{Z}_i)\|}{u}\right) &= \int_0^\infty \exp(x) P\left(\frac{\|\mathcal{M}_1(\mathbf{Z}_i)\|}{u} \geq x\right) dx \\
&\leq \int_0^1 \exp(x) dx + \int_1^\infty \exp(x) \cdot \exp(-\frac{C_0^2 x^2}{2}) dx \\
&= \exp(1) - 1 + \int_1^\infty \exp\left(x - \frac{C_0^2 x^2}{2}\right) dx \\
&\leq \exp(1) - 1 + \int_1^\infty \exp\left((1 - \frac{C_0^2}{2})x\right) dx \\
&= \exp(1) - 1 + \frac{2}{2 - C_0^2} \exp(\frac{2 - C_0^2}{2}) \leq 2,
\end{aligned}$$

which suggests that $U_Z^{(1)} = \|\|\mathcal{M}_1(\mathbf{Z}_i)\|\|_{\psi_1} = \inf\{u > 0 : \mathrm{E}\exp(\|\mathcal{M}_1(\mathbf{Z}_i)\|/u) \leq 2\} \leq C_0(\sqrt{p_1} + \sqrt{p_2 p_3 (K-1)})$ for some uniform constant $C_0 > 0$. Then, by Lemma 15,

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{M}_1(\mathbf{Z}_i)\right\| \leq C \max\{&\sigma_Z \sqrt{\frac{\log(p_1 + p_2 p_3 (K-1)) + t}{n}}, \\
&C_0(\sqrt{p_1} + \sqrt{p_2 p_3 (K-1)}) \log\left(\frac{C_0(\sqrt{p_1} + \sqrt{p_2 p_3 (K-1)})}{\sigma_Z}\right) \frac{\log(p_1 + p_2 p_3 (K-1)) + t}{n}\}
\end{aligned}$$

with probability at least $1-\exp(-t)$ where

$$\begin{aligned}\sigma_Z &= \max\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\mathcal{M}_1(\mathbf{Z}_i)\mathcal{M}_1^T(\mathbf{Z}_i)\right]\right\|^{1/2}, \left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\mathcal{M}_1^T(\mathbf{Z}_i)\mathcal{M}_1(\mathbf{Z}_i)]\right\|^{1/2}\right\} \\ &= \max\left\{\|2\mathrm{tr}(\mathbf{\Sigma}_4\otimes\mathbf{\Sigma}_3\otimes\mathbf{\Sigma}_2)\mathbf{\Sigma}_1\|^{1/2}, \|2\mathrm{tr}(\mathbf{\Sigma}_1)\mathbf{\Sigma}_4\otimes\mathbf{\Sigma}_3\otimes\mathbf{\Sigma}_2\|^{1/2}\right\} \\ &\leq \max\left\{\sqrt{p_2p_3(K-1)K}C_{\mathbf{\Sigma}}^{5/2}, \sqrt{2p_1}\left(\frac{K}{2}\right)^{1/4}C_{\mathbf{\Sigma}}^2\right\}.\end{aligned}$$

Thus,

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_1(\mathbf{Z}_i)\right\| \geq C(\sqrt{p_1}+\sqrt{p_2p_3(K-1)})\left(\sqrt{\frac{\log(p_1+p_2p_3(K-1))+t}{n}}+\frac{\log(p_1+p_2p_3(K-1))+t}{n}\right)\right)$$
$$\leq \exp(-t)$$

□

Lemma 17. *For $\mathbf{\Sigma} = CS(\rho)\in\mathbb{R}^{p\times p}$ where $\Sigma_{ii}=1$ and $\Sigma_{ij}=\rho, i\neq j, i,j\in\{1,...,p\}$, the largest eigenvalue is $\lambda_1(\mathbf{\Sigma})=1+(p-1)\rho$ and remaining $(p-1)$ eigenvalues are $\lambda_2(\mathbf{\Sigma})=\cdots=\lambda_p(\mathbf{\Sigma})=1-\rho$.*

Lemma 18 (Lemma 7. Zhang and Xia 2018)**.** *Let $\boldsymbol{\mathcal{X}}_{p_1,p_2}=\{\mathbf{X}\in\mathbb{R}^{p_1\times p_2}:\|\mathbf{X}\|\leq 1\}$ be the unit ball around the center in spectral norm. Then there exists an ϵ-net $\overline{\boldsymbol{\mathcal{X}}}_{p_1,p_2}$ in spectral norm with cardinality at most $(\frac{2+\epsilon}{\epsilon})^{p_1p_2}$ for $\boldsymbol{\mathcal{X}}_{p_1,p_2}$. To be specific, there exists $\mathbf{X}^{(1)},...,\mathbf{X}^{(N)}$ with $N\leq(\frac{2+\epsilon}{\epsilon})^{p_1p_2}$ such that for all $\mathbf{X}\in\boldsymbol{\mathcal{X}}_{p_1,p_2}$, there exists $i\in\{1,...,N\}$ satisfying $\|\mathbf{X}^{(i)}-\mathbf{X}\|\leq\epsilon$.*

Lemma 19. *For tensors $\mathbf{Z}_i\in\mathbb{R}^{p_1\times p_2\times p_3\times(K-1)}$ which follows $TN(\mathbf{0},2\mathbf{\Sigma}_1,\mathbf{\Sigma}_2,\mathbf{\Sigma}_3,\mathbf{\Sigma}_4)$ where $K\geq 2$ is an integer, $\mathbf{\Sigma}_m\succ 0$, $C_{\mathbf{\Sigma}}^{-1}\leq\sigma_{\min}(\mathbf{\Sigma}_m)\leq\sigma_{\max}(\mathbf{\Sigma}_m)\leq C_{\mathbf{\Sigma}}$, $m=1,2,3$ and*

$\boldsymbol{\Sigma}_4 = CS(0.5)$, *we have*

$$\max_{\substack{\mathbf{V}_2\in\mathbb{R}^{p_2\times r_2},\mathbf{V}_3\in\mathbb{R}^{p_3\times r_3}\\ \|\mathbf{V}_2\|\le 1,\|\mathbf{V}_3\|\le 1}} \left\|\mathcal{M}_1\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\times_2\mathbf{V}_2^T\times_3\mathbf{V}_3^T\right)\right\| \le C'\left(\sqrt{p_1}+\sqrt{r_2r_3(K-1)}\right)\sqrt{\frac{\log(p_1+r_2r_3(K-1))+x}{n}} \tag{S7.124}$$

with probability at least $1-C''\exp(-x)$.

Proof. By Lemma 18, for $m=1,2,3$, there exist ϵ-nets: $\mathbf{V}_m^{(1)}$, ..., $\mathbf{V}_m^{(N_m)}$ for $\{\mathbf{V}_m\in\mathbb{R}^{p_m\times r_m}:\|\mathbf{V}_m\|\le 1\}$, $N_m\le(\frac{2+\epsilon}{\epsilon})^{p_mr_m}$ such that for any $\mathbf{V}\in\mathbb{R}^{p_m\times r_m}$ with $\|\mathbf{V}\|\le 1$, there exists a $\mathbf{V}_m^{(j)}$ which satisfies $\|\mathbf{V}_m^{(j)}-\mathbf{V}\|\le\epsilon$. For any fixed $\mathbf{V}_2^{(s)}\in\mathbb{R}^{p_2\times r_2}$ and $\mathbf{V}_3^{(q)}\in\mathbb{R}^{p_3\times r_3}$ in such ϵ-nets, $\mathcal{M}_1\left(\mathbf{Z}_i\times_2(\mathbf{V}_2^{(s)})^T\times_3(\mathbf{V}_3^{(q)})^T\right)\sim MN\left(\mathbf{0},2\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_4\otimes\mathbf{V}_3^{(q)T}\boldsymbol{\Sigma}_3\mathbf{V}_3^{(q)}\otimes\mathbf{V}_2^{(s)T}\boldsymbol{\Sigma}_2\mathbf{V}_2^{(s)}\right)$. Let $\widetilde{\mathbf{Z}}_i\overset{iid}{\sim}MN\left(\mathbf{I}_{p_1},\mathbf{I}_{p_2p_3(K-1)}\right)$. It follows that

$$\mathcal{M}_1\left(\mathbf{Z}_i\times_2(\mathbf{V}_2^{(s)})^T\times_3(\mathbf{V}_3^{(q)})^T\right)=\sqrt{2}\boldsymbol{\Sigma}_1^{1/2}\widetilde{\mathbf{Z}}_i\left(\boldsymbol{\Sigma}_4\otimes\mathbf{V}_3^{(q)T}\boldsymbol{\Sigma}_3\mathbf{V}_3^{(q)}\otimes\mathbf{V}_2^{(s)T}\boldsymbol{\Sigma}_2\mathbf{V}_2^{(s)}\right)^{1/2}.$$

Then,

$$\begin{aligned}\mathbf{E}_1^{(sq)}&\triangleq\mathcal{M}_1\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\times_2(\mathbf{V}_2^{(s)})^T\times_3(\mathbf{V}_3^{(q)})^T\right)\\&=\sqrt{2}\boldsymbol{\Sigma}_1^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i\right)\left(\boldsymbol{\Sigma}_4\otimes\mathbf{V}_3^{(q)T}\boldsymbol{\Sigma}_3\mathbf{V}_3^{(q)}\otimes\mathbf{V}_2^{(s)T}\boldsymbol{\Sigma}_2\mathbf{V}_2^{(s)}\right)^{1/2}\in\mathbb{R}^{p_1\times r_2r_3(K-1)},\end{aligned}$$

with

$$\begin{aligned}\left\|\mathbf{E}_1^{(sq)}\right\|&\le\sqrt{2}\left\|\boldsymbol{\Sigma}_1\right\|^{1/2}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i\right\|\left\|\boldsymbol{\Sigma}_4\right\|^{1/2}\left\|\boldsymbol{\Sigma}_3\right\|^{1/2}\left\|\mathbf{V}_3^{(q)}\right\|\left\|\boldsymbol{\Sigma}_2\right\|^{1/2}\left\|\mathbf{V}_2^{(s)}\right\|\\&\le C_{\boldsymbol{\Sigma}}^{3/2}(2K)^{1/4}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i\right\|.\end{aligned}$$

By Lemma 16, we know that

$$P\left(\left\|\mathbf{E}_1^{(sq)}\right\|>C\left(\sqrt{p_1}+\sqrt{r_2r_3(K-1)}\right)\sqrt{\frac{\log(p_1+r_2r_3(K-1))+x}{n}}\right)<\exp(-x).$$

By Lemma 18, we have

$$P\left(\max_{s,q}\left\|\mathbf{E}_1^{(sq)}\right\| \le C\left(\sqrt{p_1}+\sqrt{r_2r_3(K-1)}\right)\sqrt{\frac{\log(p_1+r_2r_3(K-1))+x}{n}}\right) \ge 1-2\left(\frac{2+\epsilon}{\epsilon}\right)^{p_2r_2+p_3r_3}\exp\left(-x\right) \tag{S7.125}$$

for all $x>0$. Let

$$(\mathbf{V}_2^*,\mathbf{V}_3^*) \triangleq \underset{\substack{\mathbf{V}_2\in\mathbb{R}^{p_2\times r_2},\mathbf{V}_3\in\mathbb{R}^{p_3\times r_3}\\ \|\mathbf{V}_2\|\le 1,\|\mathbf{V}_3\|\le 1}}{\arg\max}\left\|\mathcal{M}_1\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\times_2\mathbf{V}_2^T\times_3\mathbf{V}_3^T\right)\right\|,$$
$$M \triangleq \max_{\substack{\mathbf{V}_2\in\mathbb{R}^{p_2\times r_2},\mathbf{V}_3\in\mathbb{R}^{p_3\times r_3}\\ \|\mathbf{V}_2\|\le 1,\|\mathbf{V}_3\|\le 1}}\left\|\mathcal{M}_1\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\times_2\mathbf{V}_2^T\times_3\mathbf{V}_3^T\right)\right\|.$$

By the definition of ϵ-net, we can find $1\le s\le N_2$, $1\le q\le N_3$ such that $\|\mathbf{V}_2^{(s)}-\mathbf{V}_2^*\|\le\epsilon$ and $\|\mathbf{V}_3^{(q)}-\mathbf{V}_3^*\|\le\epsilon$. Then,

$$\begin{aligned}
M &= \left\|\mathcal{M}_1\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\times_2(\mathbf{V}_2^*)^T\times_3(\mathbf{V}_3^*)^T\right)\right\| \\
&\le \left\|\mathcal{M}_1\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\times_2\left(\mathbf{V}_2^{(s)}\right)^T\times_3\left(\mathbf{V}_3^{(q)}\right)^T\right)\right\| + \left\|\mathcal{M}_1\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\times_2\left(\mathbf{V}_2^*-\mathbf{V}_2^{(s)}\right)^T\times_3\left(\mathbf{V}_3^{(q)}\right)^T\right)\right\| \\
&+ \left\|\mathcal{M}_1\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\times_2(\mathbf{V}_2^*)^T\times_3\left(\mathbf{V}_3^*-\mathbf{V}_3^{(q)}\right)^T\right)\right\| \\
&\le C\left(\sqrt{p_1}+\sqrt{r_2r_3(K-1)}\right)\sqrt{\frac{\log(p_1+r_2r_3(K-1))+x}{n}}+2\epsilon M
\end{aligned}$$

with probability at least $1-2\left(\frac{2+\epsilon}{\epsilon}\right)^{p_2r_2+p_3r_3}\exp\left(-x\right)$, i.e.,

$$P\left(M\le\frac{C\left(\sqrt{p_1}+\sqrt{r_2r_3(K-1)}\right)\sqrt{\frac{\log(p_1+r_2r_3(K-1))+x}{n}}}{1-2\epsilon}\right)\ge 1-2\left(\frac{2+\epsilon}{\epsilon}\right)^{p_2r_2+p_3r_3}\exp\left(-x\right).$$

Let $\epsilon=1/3$ and $x=C_0(p_2r_2+p_3r_3)t$. We have

$$\begin{aligned}
&\max_{\substack{\mathbf{V}_2\in\mathbb{R}^{p_2\times r_2},\mathbf{V}_3\in\mathbb{R}^{p_3\times r_3}\\ \|\mathbf{V}_2\|\le 1,\|\mathbf{V}_3\|\le 1}}\left\|\mathcal{M}_1\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\times_2\mathbf{V}_2^T\times_3\mathbf{V}_3^T\right)\right\| \\
&\le C'\left(\sqrt{p_1}+\sqrt{r_2r_3(K-1)}\right)\sqrt{\frac{\log(p_1+r_2r_3(K-1))+C_0(p_2r_2+p_3r_3)t}{n}}
\end{aligned}$$

with probability at least $1-\exp\{-C_0'(p_2r_2+p_3r_3)t\}$. $\square$

Lemma 20 (Lemma A.3. Min and Mai 2022)**.** *Under the TLC model* (3.1)*&*(3.2)*, the method of moment estimator for* $\mathbf{\Sigma}_m$ *proposed in Section 4.2 satisfies*

$$\left\|(\widehat{\mathbf{\Sigma}}_m-\mathbf{\Sigma}_m)\mathbf{d}_l\right\|_\infty \lesssim \|\mathbf{d}_l\|_2\sqrt{\frac{\log p_m}{n}},\quad l=1,...,r_m,\quad m=1,...,M. \tag{S7.126}$$

Lemma 21 (Corollary of Lemma A.4 in Min and Mai 2022)**.** *Consider* $\widehat{\mathbf{D}}=\left[\widehat{\mathbf{d}}_1,...,\widehat{\mathbf{d}}_r\right]$ *which satisfies*

$$\begin{aligned}\widehat{\mathbf{D}} &= \underset{\mathbf{D}\in\mathbb{R}^{p\times r}}{\arg\min}\left\{tr\left(\frac{1}{2}\mathbf{D}^T\widehat{\mathbf{\Sigma}}\mathbf{D}-\widehat{\mathbf{A}}^{\mathrm{T}}\mathbf{D}\right)+\lambda\sum_{l=1}^{p}\|\mathbf{D}_{l\cdot}\|_2\right\}\\ &= \underset{\substack{\mathbf{D}=[\mathbf{d}_1,...,\mathbf{d}_r]\\ \mathbf{d}_1,...,\mathbf{d}_r\in\mathbb{R}^p}}{\arg\min}\left\{\sum_{l=1}^{r}\left(\frac{1}{2}\mathbf{d}_l^T\widehat{\mathbf{\Sigma}}\mathbf{d}_l-\widehat{\mathbf{a}}_l^T\mathbf{d}_l\right)+\lambda\sum_{l=1}^{p}\|\mathbf{D}_{l\cdot}\|_2\right\},\end{aligned} \tag{S7.127}$$

where $\widehat{\mathbf{a}}_l,\widehat{\mathbf{d}}_l\in\mathbb{R}^p$*,* $\widehat{\mathbf{A}}=[\widehat{\mathbf{a}}_1,...,\widehat{\mathbf{a}}_r],\mathbf{D}=[\mathbf{d}_1,...,\mathbf{d}_r]\in\mathbb{R}^{p\times r}$*,* $\mathbf{D}_{j\cdot}\in\mathbb{R}^r$ *is the* j*th row of* $\mathbf{D}$*, and* $\|\mathbf{D}_{j\cdot}\|_2=\sqrt{\sum_{l=1}^r d_{jl}^2}$ *is the* l_2 *norm of* $\mathbf{D}_{j\cdot}$*. With properly chosen* $\lambda\asymp\frac{1}{\eta}\sqrt{\frac{r p\log p}{n}}$*, if*

(i) $\|\widehat{\mathbf{a}}_l-\mathbf{a}_l\|_\infty\le\frac{C_1}{\eta}\sqrt{\frac{p\log p}{n}}$*;*

(ii) $\|(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{d}_l\|_\infty\le C_2\sqrt{\frac{\log p}{n}}$*,*

where C_1,C_2 *are some positive constants, we have*

$$\sum_{j\in S^c}\|\Upsilon_{j\cdot}\|_2\le 2\sum_{j\in S}\|\Upsilon_{j\cdot}\|_2 \tag{S7.128}$$

where $\Upsilon=\widehat{\mathbf{D}}-\mathbf{D}$ *and* $S=\{j:d_{jl}\neq 0\ \text{for some } l\}$*.*

Proof. When conditions (i) and (ii) both hold, we have

$$\|\widehat{\mathbf{\Sigma}}\mathbf{d}_l-\widehat{\mathbf{a}}_l\|_\infty=\|\widehat{\mathbf{\Sigma}}\mathbf{d}_l-\mathbf{\Sigma}\mathbf{d}_l+\mathbf{\Sigma}\mathbf{d}_l-\widehat{\mathbf{a}}_l\|_\infty\le\|\widehat{\mathbf{\Sigma}}\mathbf{d}_l-\mathbf{\Sigma}\mathbf{d}_l\|_\infty+\|\mathbf{a}_l-\widehat{\mathbf{a}}_l\|_\infty\le\frac{C_1'}{\eta}\sqrt{\frac{p\log p}{n}}.$$

Since $\widehat{\mathbf{D}}$ minimizes the objective function in (S7.127), we have

$$\begin{aligned}
\lambda\sum_{j=1}^{p}\left(\|\widehat{\mathbf{D}}_{j\cdot}\|_2-\|\mathbf{D}_{j\cdot}\|_2\right) &\le \sum_{l=1}^{r}\left[-\frac{1}{2}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)^T\widehat{\boldsymbol{\Sigma}}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)-\langle\widehat{\boldsymbol{\Sigma}}\mathbf{d}_l-\widehat{\mathbf{a}}_l,\widehat{\mathbf{d}}_l-\mathbf{d}_l\rangle\right]\\
&\le \frac{1}{2}\sum_{l=1}^{r}|\langle\widehat{\boldsymbol{\Sigma}}\mathbf{d}_l-\widehat{\mathbf{a}}_l,\widehat{\mathbf{d}}_l-\mathbf{d}_l\rangle|\\
&\le \frac{1}{2}\sum_{l=1}^{r}\|\widehat{\boldsymbol{\Sigma}}\mathbf{d}_l-\widehat{\mathbf{a}}_l\|_\infty\|\widehat{\mathbf{d}}_l-\mathbf{d}_l\|_1\\
&\le \frac{C_1'}{2\eta}\sqrt{\frac{p\log p}{n}}\sum_{l=1}^{r}\|\widehat{\mathbf{d}}_l-\mathbf{d}_l\|_1=\frac{C_1'}{2\eta}\sqrt{\frac{p\log p}{n}}\sum_{j=1}^{p}\|\widehat{\mathbf{D}}_{j\cdot}-\mathbf{D}_{j\cdot}\|_1\\
&\le \frac{C_1'}{2\eta}\sqrt{\frac{rp\log p}{n}}\sum_{j=1}^{p}\|\widehat{\mathbf{D}}_{j\cdot}-\mathbf{D}_{j\cdot}\|_2\\
&= \frac{C_1'}{2\eta}\sqrt{\frac{rp\log p}{n}}\sum_{j=1}^{p}\|\boldsymbol{\Upsilon}_{j\cdot}\|_2. \qquad \text{(S7.129)}
\end{aligned}$$

Besides,

$$\begin{aligned}
\lambda\sum_{j=1}^{p}\left(\|\widehat{\mathbf{D}}_{j\cdot}\|_2-\|\mathbf{D}_{j\cdot}\|_2\right) &= \lambda\sum_{j\in S}\left(\|\widehat{\mathbf{d}}_{j\cdot}\|_2-\|\mathbf{d}_{j\cdot}\|_2\right)+\lambda\sum_{j\in S^c}\|\widehat{\mathbf{d}}_{j\cdot}\|_2\\
&\ge \lambda\sum_{j\in S^c}\|\widehat{\mathbf{d}}_{j\cdot}\|_2-\lambda\sum_{j\in S}\left|\|\widehat{\mathbf{d}}_{j\cdot}\|_2-\|\mathbf{d}_{j\cdot}\|_2\right|\\
&\ge \lambda\sum_{j\in S^c}\|\boldsymbol{\Upsilon}_{j\cdot}\|_2-\lambda\sum_{j\in S}\|\boldsymbol{\Upsilon}_{j\cdot}\|_2. \qquad \text{(S7.130)}
\end{aligned}$$

Combining (S7.129) and (S7.130) and setting $\lambda\ge\frac{3C_1'}{2\eta}\sqrt{\frac{rp\log p}{n}}$, we have

$$\sum_{j\in S^c}\|\boldsymbol{\Upsilon}_{j\cdot}\|_2-\sum_{j\in S}\|\boldsymbol{\Upsilon}_{j\cdot}\|_2\le\sum_{j=1}^{p}\left(\|\widehat{\mathbf{D}}_{j\cdot}\|_2-\|\mathbf{D}_{j\cdot}\|_2\right)\le\frac{1}{3}\left(\sum_{j\in S^c}\|\boldsymbol{\Upsilon}_{j\cdot}\|_2+\lambda\sum_{j\in S}\|\boldsymbol{\Upsilon}_{j\cdot}\|_2\right),$$

which implies that

$$\sum_{j\in S^c}\|\boldsymbol{\Upsilon}_{j\cdot}\|_2\le 2\sum_{j\in S}\|\boldsymbol{\Upsilon}_{j\cdot}\|_2.$$

□

Lemma 22 (Lemma B.3 & A.6 Min and Mai 2022)**.** *Assume that* $C_{\boldsymbol{\Sigma}}^{-2}\le\lambda_{\min}(\boldsymbol{\Sigma}_m)\le$

$\lambda_{\max}(\mathbf{\Sigma}_m) \leq C_{\mathbf{\Sigma}}^2$ *for some constant* $C_{\mathbf{\Sigma}} > 0$. *Let* $p_{-m} = \prod_{l \neq m} p_l$. *If* $s_m \log p_m/(n_k p_{-m}) < C$ *for some constant* $C > 0$, *the MOM estimators* $\widehat{\mathbf{\Sigma}}_m$, $m = 1, ..., M$, *satisfy*

$$C_{\mathbf{\Sigma}}^{-2} - C'\epsilon_s \leq \phi_{\min}^{\widehat{\mathbf{\Sigma}}_m}(s) \leq \phi_{\max}^{\widehat{\mathbf{\Sigma}}_m}(s) \leq C_{\mathbf{\Sigma}}^2 + C'\epsilon_s, \tag{S7.131}$$

with probability at least $1 - C'' \exp(-s_m \log p_m)$, *where* $\epsilon_s = \sqrt{\frac{s_m \log p_m}{n_k p_{-m}}}$, $C'' > 0$ *is a constant,* $C' > 0$ *is some constant depending on* C, C', *and* $C_{\mathbf{\Sigma}}$, *and* $\phi_{\min}^{\widehat{\mathbf{\Sigma}}_m}(s)$ *and* $\phi_{\max}^{\widehat{\mathbf{\Sigma}}_m}(s)$ *are the restricted eigenvalues of* $\widehat{\mathbf{\Sigma}}_m$ *such that*

$$\phi_{\min}^{\widehat{\mathbf{\Sigma}}_m}(s) = \min_{\|\mathbf{u}\|_0 \leq s, \mathbf{u} \neq \mathbf{0}} \frac{\mathbf{u}^T \widehat{\mathbf{\Sigma}}_m \mathbf{u}}{\mathbf{u}^T \mathbf{u}}, \quad \phi_{\max}^{\widehat{\mathbf{\Sigma}}_m}(s) = \max_{\|\mathbf{u}\|_0 \leq s, \mathbf{u} \neq \mathbf{0}} \frac{\mathbf{u}^T \widehat{\mathbf{\Sigma}}_m \mathbf{u}}{\mathbf{u}^T \mathbf{u}}.$$

If we further have $\sum_{j \in S^c} \|\Upsilon_{j\cdot}\|_2 \leq 2 \sum_{j \in S} \|\Upsilon_{j\cdot}\|_2$ *where* $\Upsilon = \widehat{\mathbf{D}} - \mathbf{D}$, *then*

$$tr\Big\{(\widehat{\mathbf{D}} - \mathbf{D})^T \widehat{\mathbf{\Sigma}} (\widehat{\mathbf{D}} - \mathbf{D})\Big\} \gtrsim \|\widehat{\mathbf{D}} - \mathbf{D}\|_F^2. \tag{S7.132}$$

Lemma 23 (Corollary of Lemma A.1 in Min and Mai 2022)**.** *For the problem in* (S7.127), *let* $s = |S|$ *and assume that* $\sqrt{srp \log p/n} \leq C$ *for some positive constant* C. *Choose* $\lambda \asymp \sqrt{rp \log p/n}$. *If* $\widehat{\mathbf{\Sigma}}$ *and* $\widehat{\mathbf{a}}_l$ *satisfy*

(i) $\|\widehat{\mathbf{a}}_l - \mathbf{a}_l\|_\infty \leq \frac{C_1}{\eta} \sqrt{\frac{p \log p}{n}}$;

(ii) $\|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{d}_l\|_\infty \leq C_2 \sqrt{\frac{\log p}{n}}$;

(iii) $\|\widehat{\mathbf{D}} - \mathbf{D}\|_F^2 \leq C_3 tr\Big\{(\widehat{\mathbf{D}} - \mathbf{D})^T \widehat{\mathbf{\Sigma}} (\widehat{\mathbf{D}} - \mathbf{D})\Big\}$,

with probability at least $1 - O(p^{-1})$ *where* C_1, C_2, C_3 *are some positive constants, we have*

$$\|\widehat{\mathbf{D}} - \mathbf{D}\|_F \leq \frac{C}{\eta} \sqrt{\frac{srp \log p}{n}}. \tag{S7.133}$$

Proof. By Lemma 21, we know that when conditions (i) and (ii) hold with $\lambda \asymp \frac{1}{\eta} \sqrt{\frac{rp \log p}{n}}$,

$$\sum_{j \in S^c} \|\widehat{\mathbf{D}}_{j\cdot} - \mathbf{D}_{j\cdot}\|_2 = \sum_{j \in S^c} \|\Upsilon_{j\cdot}\|_2 \leq 2 \sum_{j \in S} \|\Upsilon_{j\cdot}\|_2 = 2 \sum_{j \in S} \|\widehat{\mathbf{D}}_{j\cdot} - \mathbf{D}_{j\cdot}\|_2,$$

$$\lambda\sum_{j=1}^{p}\left(\|\widehat{\mathbf{D}}_{j\cdot}\|_2-\|\mathbf{D}_{j\cdot}\|_2\right)\le\sum_{l=1}^{r}\left[-\frac{1}{2}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)^T\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)-\langle\widehat{\mathbf{\Sigma}}\mathbf{d}_l-\widehat{\mathbf{a}}_l,\widehat{\mathbf{d}}_l-\mathbf{d}_l\rangle\right].$$

It follows that

$$\begin{aligned}\sum_{l=1}^{r}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)^T\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)&\le 2\sum_{l=1}^{r}\left|\langle\widehat{\mathbf{\Sigma}}\mathbf{d}_l-\widehat{\mathbf{a}}_l,\widehat{\mathbf{d}}_l-\mathbf{d}_l\rangle\right|+2\lambda\sum_{j=1}^{p}\left|\|\widehat{\mathbf{D}}_{j\cdot}\|_2-\|\mathbf{D}_{j\cdot}\|_2\right|\\&\le 2\sum_{l=1}^{r}\left\|\widehat{\mathbf{\Sigma}}\mathbf{d}_l-\widehat{\mathbf{a}}_l\right\|_\infty\|\widehat{\mathbf{d}}_l-\mathbf{d}_l\|_1+2\lambda\sum_{j=1}^{p}\left|\|\widehat{\mathbf{D}}_{j\cdot}\|_2-\|\mathbf{D}_{j\cdot}\|_2\right|\\&\le\frac{C_1'}{\eta}\sqrt{\frac{p\log p}{n}}\sum_{l=1}^{r}\|\widehat{\mathbf{d}}_l-\mathbf{d}_l\|_1+\frac{C_4}{\eta}\sqrt{\frac{rp\log p}{n}}\sum_{j=1}^{p}\|\widehat{\mathbf{D}}_{j\cdot}-\mathbf{D}_{j\cdot}\|_2.\end{aligned}$$

Also, note that

$$\begin{aligned}\sum_{l=1}^{r}\|\widehat{\mathbf{d}}_l-\mathbf{d}_l\|_1&=\sum_{j=1}^{p}\|\widehat{\mathbf{D}}_{j\cdot}-\mathbf{D}_{j\cdot}\|_1\le\sqrt{r}\sum_{j=1}^{p}\|\widehat{\mathbf{D}}_{j\cdot}-\mathbf{D}_{j\cdot}\|_2\le 3\sqrt{r}\sum_{j\in S}\|\widehat{\mathbf{D}}_{j\cdot}-\mathbf{D}_{j\cdot}\|_2\\&\le 3\sqrt{sr\sum_{j\in S}\|\widehat{\mathbf{D}}_{j\cdot}-\mathbf{D}_{j\cdot}\|_2^2}\le 3\sqrt{sr}\|\widehat{\mathbf{D}}-\mathbf{D}\|_F.\end{aligned}$$

Hence,

$$\sum_{l=1}^{r}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)^T\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)\le\frac{C_5}{\eta}\sqrt{\frac{srp\log p}{n}}\|\widehat{\mathbf{D}}-\mathbf{D}\|_F.$$

Meanwhile, when condition (iii) holds,

$$\|\widehat{\mathbf{D}}-\mathbf{D}\|_F^2\le C_3\mathrm{tr}\left\{(\widehat{\mathbf{D}}-\mathbf{D})^T\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{D}}-\mathbf{D})\right\}=C_3\sum_{l=1}^{r}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)^T\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{d}}_l-\mathbf{d}_l)\le\frac{C}{\eta}\sqrt{\frac{srp\log p}{n}}\|\widehat{\mathbf{D}}-\mathbf{D}\|_F,$$

which implies that

$$\|\widehat{\mathbf{D}}-\mathbf{D}\|_F\le\frac{C}{\eta}\sqrt{\frac{srp\log p}{n}}.$$

□

Lemma 24 (Lemma S.1. Lyu et al. 2019)**.** *Assume i.i.d. data* $\mathbf{X}_1,...,\mathbf{X}_n\in\mathbb{R}^{p\times q}$ *follows the matrix-variate normal distribution such that* $vec(\mathbf{X}_i)\sim N(\mathbf{0},\mathbf{\Psi}^*\otimes\mathbf{\Sigma}^*)$ *with* $\mathbf{\Psi}^*\in\mathbb{R}^{q\times q}$ *and* $\mathbf{\Sigma}^*\in\mathbb{R}^{p\times p}$*. Assume that* $0<C_1\le\lambda_{\min}(\mathbf{\Sigma}^*)\le\lambda_{\max}(\mathbf{\Sigma}^*)\le 1/C_1<\infty$ *and*

$0 < C_2 \leq \lambda_{\min}(\mathbf{\Psi}^*) \leq \lambda_{\max}(\mathbf{\Psi}^*) \leq 1/C_2 < \infty$ *for some positive constants* C_1, C_2. *For any vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *such that* $\|\mathbf{a}\|_2 \leq C_3 \leq \infty$ *and* $\|\mathbf{b}\|_2 \leq C_4 \leq \infty$, *we have*

$$P\left\{\left|\frac{1}{np}\sum_{i=1}^{n}\mathbf{a}^T\mathbf{X}_i^T\mathbf{X}_i\mathbf{b} - \frac{1}{p}\mathrm{E}(\mathbf{a}^T\mathbf{X}_i^T\mathbf{X}_i\mathbf{b})\right| \geq x\|\mathbf{a}\|_2\|\mathbf{b}\|_2\right\} \leq c_1\exp\left\{-c_2npx^2\right\}. \quad \text{(S7.134)}$$

Lemma 25 (Lemma S.12. Lyu et al., 2019)**.** *Suppose that a p-dimensional Gaussian random vector* $\mathbf{y} \sim N(\mathbf{0}, \mathbf{Q})$. *Then, for any* $x > 2/\sqrt{p}$, *we have*

$$P\left\{\frac{1}{p}\left|\|\mathbf{y}\|_2^2 - \mathrm{E}\|\mathbf{y}\|_2^2\right| > 4x\|\mathbf{Q}\|\right\} \leq 2\exp\left\{-\frac{p(x-2/\sqrt{p})^2}{2}\right\} + 2\exp\left\{-p/2\right\}. \quad \text{(S7.135)}$$

For a positive integer $s < p$, let $\Gamma(s) = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_{\mathcal{S}^c}\|_1 \leq 2\|\mathbf{u}_{\mathcal{S}}\|_1$ for some $\mathcal{S} \subset [p]$ with $|\mathcal{S}| = s\}$, where $\mathbf{u}_{\mathcal{S}}$ denotes the subvector of $\mathbf{u}$ on the index set $\mathcal{S}$. Define the restricted norm $\|\mathbf{u}\|_{2,s} = \sup_{\|\mathbf{v}\|_2=1, \mathbf{v}\in\Gamma(s)} |\mathbf{u}^T\mathbf{v}|$.

Lemma 26 (Lemma A.2. Min and Mai 2022)**.** *With probability at least* $1 - O(p^{-1})$, *we have*

$$\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_{\max} \leq C\sqrt{\frac{\sum_{m=1}^{M}\log p_m}{n_k}}, \quad \text{(S7.136)}$$

where $C > 0$ *is a constant.*

Lemma 27 (Lemma B.5. Min and Mai 2022)**.** *If* $\mathbf{u} \in \Gamma(s)$ *and* $\|\mathbf{u}\|_2 = 1$, *then*

$$\|\mathbf{u}\|_1 \lesssim \sqrt{s}. \quad \text{(S7.137)}$$

Lemma 28 (Proposition A.1. Min and Mai 2022)**.** *For two vectors* $\boldsymbol{\gamma}$ *and* $\widehat{\boldsymbol{\gamma}}$, *if* $\|\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}\|_2 = o(1)$ *as* $n \to \infty$, *and* $\|\boldsymbol{\gamma}\|_2 \geq c$ *for some constant* $c > 0$, *then, when* $n \to \infty$,

$$\|\boldsymbol{\gamma}\|_2\|\widehat{\boldsymbol{\gamma}}\|_2 - \boldsymbol{\gamma}^T\widehat{\boldsymbol{\gamma}} \asymp \|\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}\|_2^2. \quad \text{(S7.138)}$$

For the solution to (S1.1), we further consider the following assumption:

(A1) $|\hat{\sigma}_{m,ij} - \sigma_{m,ij}| \leq \frac{\epsilon}{s_{\max}}$ for any $i = 1, \ldots, p_m$ and $j \in \mathcal{S}_m$, and $|(\hat{A}_{mj} - \hat{A}_{1j}) - (A_{mj} - A_{1j})| < \epsilon$ for any m and j, where $s_{\max} = \max\{s_1, \ldots, s_M\}$ and $\epsilon > 0$.

Lemma 29. *Under the TLC model, if conditions (C1) & (C2) and the assumption (A1) are all satisfied, there exist constants ψ_1, ψ_2 such that if $\psi_1 \epsilon < \lambda < \min\{\frac{D_{m,\min}}{8\phi}, \psi_2(1 - \kappa_m)\}$, the sparse factor matrix estimate* (S1.1) *achieves $\widehat{\mathcal{S}}_m = \mathcal{S}_m$.*

Proof of Lemma 29. The proof of Lemma 29] is similar to that of Theorem 1 in Mai et al. (2019b), and is thus omitted here. □

Lemma 30 (Theorem 5 in Luo et al. 2021)**.** *Suppose $\mathbf{B} = \mathbf{A} + \mathbf{Z}$ where $\mathbf{A}$ is an unknown rank-r matrix, $\mathbf{B}$ is the observation, and $\mathbf{Z}$ is the perturbation. Let $\widehat{\mathbf{A}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T$ be the best rank-r approximation of $\mathbf{B}$. Then,*

$$\max\left\{\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|, \|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|\right\} \leq \frac{2\|\mathbf{Z}\|}{\sigma_r(\mathbf{A})}. \tag{S7.139}$$

Bibliography

Almodóvar-Rivera, I. and R. Maitra (2019). Fast adaptive smoothing and thresholding for improved activation detection in low-signal fmri. *IEEE transactions on medical imaging 38*(12), 2821–2828.

Aston, J. A., D. Pigoli, and S. Tavakoli (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, 1431–1461.

Baranzini, S. E., P. Mousavi, J. Rio, S. J. Caillier, A. Stillman, P. Villoslada, M. M. Wyatt, M. Comabella, L. D. Greller, R. Somogyi, et al. (2004). Transcription-based prediction of response to ifnβ using supervised computational methods. *PLoS Biol 3*(1), e2.

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

Chen, E. Y. and J. Fan (2021). Statistical inference for high-dimensional matrix-variate factor models. *Journal of the American Statistical Association*, 1–18.

Chen, R., D. Yang, and C.-H. Zhang (2022). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association 117*(537), 94–116.

Chen, Y., Y. Chi, J. Fan, C. Ma, et al. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning 14*(5), 566–806.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2000a). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications 21*(4), 1253–1278.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2000b). On the best rank-1 and rank-(R_1, R_2,..., R_N) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications 21*(4), 1324–1342.

Fleming, T. R. and D. P. Harrington (2011). *Counting processes and survival analysis*, Volume 169. John Wiley & Sons.

Georghiades, A., P. Belhumeur, and D. Kriegman (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence 23*(6), 643–660.

Han, Y., R. Chen, and C.-H. Zhang (2022). Rank determination in tensor factor model. *Electronic Journal of Statistics 16*(1), 1726–1803.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and

optimal rates for noisy low-rank matrix completion. *The Annals of Statistics 39*(5), 2302–2329.

Li, Q. and D. Schonfeld (2014). Multilinear discriminant analysis for higher-order tensor data classification. *IEEE transactions on pattern analysis and machine intelligence 36*(12), 2524–2537.

Lu, W., Z. Zhu, and H. Lian (2020). High-dimensional quantile tensor regression. *Journal of Machine Learning Research 21*(250), 1–31.

Luo, Y., R. Han, and A. R. Zhang (2021). A schatten-q low-rank matrix perturbation analysis via perturbation projection error bound. *Linear Algebra and its Applications 630*, 225–240.

Lyu, T., E. F. Lock, and L. E. Eberly (2017). Discriminating sample groups with multiway data. *Biostatistics 18*(3), 434–450.

Lyu, X., W. W. Sun, Z. Wang, H. Liu, J. Yang, and G. Cheng (2019). Tensor graphical model: Non-convex optimization and statistical inference. *IEEE transactions on pattern analysis and machine intelligence 42*(8), 2024–2037.

Mai, Q., Y. Yang, and H. Zou (2019a). Multiclass sparse discriminant analysis. *Statistica Sinica 29*(1), 97–111.

Mai, Q., Y. Yang, and H. Zou (2019b). Multiclass sparse discriminant analysis. *Statistica Sinica 29*(1), 97–111.

Maitra, R. (2010). A re-defined and generalized percent-overlap-of-activation measure for studies of fmri reproducibility and its use in identifying outlier activation maps. *Neuroimage 50*(1), 124–135.

Maitra, R., S. R. Roys, and R. P. Gullapalli (2002). Test-retest reliability estimation of functional mri data. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 48*(1), 62–70.

Min, K. and Q. Mai (2022). Optimality in high-dimensional tensor discriminant analysis. *Manuscript*.

Molstad, A. J. and A. J. Rothman (2019). A penalized likelihood method for classification with matrix-valued predictors. *Journal of Computational and Graphical Statistics 28*(1), 11–22.

Pan, Y., Q. Mai, and X. Zhang (2019a). Covariate-adjusted tensor classification in high dimensions. *Journal of the American Statistical Association 114*(527), 1305–1319.

Pan, Y., Q. Mai, and X. Zhang (2019b). Covariate-adjusted tensor classification in high dimensions. *Journal of the American Statistical Association 114*(527), 1305–1319.

Rudelson, M. and R. Vershynin (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability 18*, 1–9.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association 81*(393), 142–149.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association 97*(460), 1167–1179.

Thompson, G. Z., R. Maitra, W. Q. Meeker, and A. F. Bastawros (2020). Classification with the matrix-variate-t distribution. *Journal of Computational and Graphical Statistics 29*(3), 668–674.

Tucker, L. R. (1963). Implications of factor analysis of three-way matrices for measurement of change. *Problems in measuring change 15*, 122–137.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika 31*(3), 279–311.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Wang, D., X. Liu, and R. Chen (2019). Factor models for matrix-valued high-dimensional time series. *Journal of econometrics 208*(1), 231–248.

Wang, M. and Y. Zeng (2019). Multiway clustering via tensor block models. *Advances in Neural Information Processing Systems 32 (NeurIPS) 32*.

Yu, L., Y. He, X. Kong, and X. Zhang (2022). Projected estimation for large-dimensional matrix factor models. *Journal of Econometrics 229*(1), 201–217.

Zhang, A. and D. Xia (2018). Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory 64*(11), 7311–7338.