Supplementary Materials for
"Parsimonious Tensor Discriminant Analysis"

S1 Adjusting for Covariates

S1.1 Model

In the situation when vectors of clinical covariates such as sex, age are available, integrating the information from both the tensor predictor and vector of covariates can help for a more accurate classification results (see Pan et al. (2019)). In this section, we show that our proposal can be easily extended in such scenario. To model the direct and indirect effect of covariates for classification, we employ the CATCH model by Pan et al. (2019), more specifically:

$$\mathbf{U} \mid (Y=k) \sim N(\boldsymbol{\phi}_k, \boldsymbol{\Phi}) \tag{S1}$$

$$\mathbf{X} \mid (\mathbf{U}=\mathbf{u}, Y=k) \sim TN(\boldsymbol{\mu}_k + \boldsymbol{\alpha}\bar{\times}_{(M+1)}\mathbf{u}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M). \tag{S2}$$

Here $\boldsymbol{\phi}_k \in \mathbb{R}^q$, $\boldsymbol{\Phi} \in \mathbb{R}^{q\times q}$, $\boldsymbol{\alpha} \in \mathbb{R}^{p_1\times\cdots\times p_M\times q}$ and $\mathbf{U} \in \mathbb{R}^q$ is a vector of covariates. From Proposition 1 of Pan et al. (2019), the Bayes' rule of CATCH model is $\phi(\mathbf{U}, \mathbf{X}) = \arg\max_{k=1,\cdots,K} P(Y = k \mid \mathbf{X}, \mathbf{U}) = \arg\max_{k=1,\cdots,K}\{c_k + \boldsymbol{\gamma}_k^T\mathbf{U} + \langle\, \mathbf{B}_k, \mathbf{X} - \boldsymbol{\alpha}\bar{\times}_{M+1}\mathbf{U}\rangle\}$, with $\boldsymbol{\gamma}_k = \boldsymbol{\Phi}^{-1}(\boldsymbol{\phi}_k - \boldsymbol{\phi}_1)$ and $c_k = \log(\pi_k/\pi_1) - \frac{1}{2}\boldsymbol{\gamma}_k^T(\boldsymbol{\phi}_k + \boldsymbol{\phi}_1) - \langle\, \mathbf{B}_k, \frac{\boldsymbol{\mu}_k+\boldsymbol{\mu}_1}{2}\rangle$. Denote $\mathbf{X}_{\text{adj}} = \mathbf{X} - \boldsymbol{\alpha}\bar{\times}_{M+1}\mathbf{U}$ as the tensor predictor after adjusting for the indirect effect of $\mathbf{U}$, then we have $\phi(\mathbf{X}, \mathbf{U}) = \phi(\langle\mathbf{B}, \mathbf{X}_{\text{adj}}\rangle, \mathbf{U})$. Thus similarly we aim for improved predictions through a more efficient estimate of discriminant coefficients $\mathbf{B}$ regarding $\mathbf{X}_{\text{adj}}$

In addition, the results in Proposition 2 are easily extended to CATCH. Recall that $\Pr(Y = k \mid \mathbf{X}, \mathbf{U}) = \Pr(Y = k \mid \mathbf{X}_{\text{adj}}, \mathbf{U})$, by assuming there exists a reducing subspace $\mathcal{S} = \mathcal{S}_M \otimes \cdots \otimes \mathcal{S}_1$, then the "generalized sparsity principle" under CATCH follows that,

$$\Pr(Y = k \mid \mathbf{X}_{\text{adj}}, \mathbf{U}) = \Pr(Y = k \mid \mathbb{P}(\mathbf{X}_{\text{adj}}), \mathbf{U}), \quad \mathbb{Q}(\mathbf{X}_{\text{adj}}) \perp\!\!\!\perp \mathbb{P}(\mathbf{X}_{\text{adj}}) \mid (Y = k, \mathbf{U} = \mathbf{u}) \tag{S3}$$

The first condition in (S3) suggests that the Bayes' rule under CATCH model does not depend on $\mathbb{Q}(\mathbf{X}_{\text{adj}})$. The second condition suggest that there is no information leak between $\mathbb{P}(\mathbf{X}_{\text{adj}})$ and $\mathbb{Q}(\mathbf{X}_{\text{adj}})$. Thus the statements in Proposition 2 satisfies under CATCH model by replacing $\mathbf{X}$ with $\mathbf{X}_{\text{adj}}$.

S1.2 Estimation

In the case when covariates $\mathbf{U}$ are available, as discussed above, we can directly apply the estimation procedure described in Section 3.3 on the covariate-adjusted tensor $\mathbf{X}_{\text{adj}}$ for an estimate of $\mathbf{B}$. Since $\mathbf{X}_{\text{adj}} = \mathbf{X} - \boldsymbol{\alpha}\bar{\times}_{M+1}\mathbf{U}$, then we only need an estimate of $\boldsymbol{\alpha}$. Moreover, to estimate the coefficients $\boldsymbol{\gamma}_k$ for the covariates $\mathbf{U}$, we directly apply the MLE $\widehat{\boldsymbol{\gamma}}_k$ from Pan et al. (2019). The following Lemma from Pan et al. (2019) provides the MLE results of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\gamma}}_k$,

Lemma S1. *Under the CATCH model* (S1),(S2)*, the MLEs for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are*

$$\widehat{\boldsymbol{\alpha}} = \widetilde{\mathbf{X}} \times_{(M+1)} \{(\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T)^{-1}\widetilde{\mathbf{U}}\}, \quad \widehat{\boldsymbol{\gamma}}_k = \widehat{\boldsymbol{\Phi}}^{-1}(\widehat{\boldsymbol{\phi}}_k - \widehat{\boldsymbol{\phi}}_1)), k = 1, \ldots, K \tag{S4}$$

where $\widetilde{\mathbf{X}}^i = \mathbf{X}^i - \overline{\mathbf{X}}_k$ and $\widetilde{\mathbf{U}}^i = \mathbf{U}^i - \overline{\mathbf{U}}_k$. $\widetilde{\mathbf{X}} \in \mathbb{R}^{p_1 \times \ldots p_M \times n}$ and $\widetilde{\mathbf{U}} \in \mathbb{R}^{q \times n}$ is the tensor and matrix that consists of $\widetilde{\mathbf{X}}^i$ and $\widetilde{\mathbf{U}}^i$. $\widehat{\boldsymbol{\phi}}_k = \overline{\mathbf{U}}_k$ and $\widehat{\boldsymbol{\Phi}} = \frac{1}{n}\sum_{k=1}^K \sum_{Y^i=k}(\mathbf{U}^i - \widehat{\boldsymbol{\phi}}_k)(\mathbf{U}^i - \widehat{\boldsymbol{\phi}}_k)^T$.

With estimator $\widehat{\mathbf{X}}_{\text{adj}}, \widehat{\boldsymbol{\gamma}}_k$ and $\widehat{\mathbf{B}}_k$, the classification follows by a vector LDA on new predictors $\widehat{\boldsymbol{\gamma}}_k^T\mathbf{U}$ and $\langle \widehat{\mathbf{B}}_k, \widehat{\mathbf{X}}_{\text{adj}}\rangle, k = 2, \ldots, K$.

S1.3 Simulation Examples

In this section, we present simulation results for model with covariates. Some competitors mentioned in Section 5 are included. We consider the CATCH model with envelope structure here. The model is generated as follows:

- (E1) The predictor tensor $\mathbf{X} \in \mathbb{R}^{20\times 30\times 40}$ with $\mathbf{U} \in \mathbb{R}^2$ and envelope dimension $u = (2, 3, 4)$ with multiclass $K = 3$ as in model (M12). The covariance matrices $\boldsymbol{\Sigma}_m, m = 1, 2, 3$ are generated according to (M12) and $\boldsymbol{\phi}_2 = 0.3, \boldsymbol{\phi}_3 = 0.6$. Let $n_1 = 0.2n, n_2 = 0.3n, n_3 = 0.5n, n = 300$ with testing sample size is 3000. The parameters $\boldsymbol{\alpha} = [\![\boldsymbol{\alpha}^*; \boldsymbol{\Sigma}_1^{1/2}, \ldots, \boldsymbol{\Sigma}_M^{1/2}, \mathbf{I}]\!]$, where $\boldsymbol{\alpha}^*_{ijst} = 0.5$ for $i, j, s = 1, \ldots, 10$ and $t = 1$ with other elements equals to 0 and $\boldsymbol{\Phi} = \mathbf{I}$.

- (E2) The same as Model (E1) except $\boldsymbol{\alpha}^*_{ijst} = 0.25$.

Model	Bayes	MLE		vec DLDA	Envelope
E1	10.54(0.05)	48.71(0.06)	$\mathbf{X}$	29.95(0.24)	23.63(0.24)
			$(\mathbf{X}, \mathbf{U})$	29.95(0.24)	15.37(0.25)
E2	10.54(0.05)	49.38(0.03)	$\mathbf{X}$	21.95(0.15)	17.85(0.27)
			$(\mathbf{X}, \mathbf{U})$	21.95(0.15)	15.37(0.25)
Model		CATCH	ℓ_1-GLM	ℓ_1-FDA	
E1	$\mathbf{X}$	36.70(0.54)	22.64(0.13)	29.95(0.24)	
	$(\mathbf{X}, \mathbf{U})$	24.54(0.26)	22.32(0.12)	29.95(0.24)	
E2	$\mathbf{X}$	28.83(0.40)	22.32(0.13)	21.95(0.15)	
	$(\mathbf{X}, \mathbf{U})$	24.55(0.26)	22.49(0.13)	21.95(0.15)	

Table S1: Average results over 100 data simulations. The reported $\mathbf{X}$ represents only use tensor predictor and $(\mathbf{X}, \mathbf{U})$ represents use both tensor predictor and covariates vector predictor. Tucker-Reg failed model (E1) since it does not directly apply to multiclass classification.

From Table S1, we see that if the model follows the envelope structure assumption, our method gives the best classification results compared with other methods and also improved the classification by adjusting the effects of covariates. We also apply our method on the ADHD dataset where covariates are available.

S1.4 ADHD Data

The attention deficit hyperactivity disorder (ADHD) dataset, is a three-way tensor valued data. It is public available on NITRC (http://fcon_1000.projects.nitrc.org/indi/adhd200, see Bellec et al. (2017)). We use the $30 \times 36 \times 30$ tensor data with 930 samples with categories: Typically Developing Children (TDC), ADHD Combined, ADHD Hyperactive and ADHD Inattentive. Followed by Pan et al. (2019), the dataset is stratified on male and female separately and a pooled classification error rates is reported. Also, we consider the three-class problem which combine the ADHD Hyperactive category with ADHD combined class and a binary classification problem by grouping the ADHD grouped with ADHD hyperactive together. We followed the same preprocessing procedure in Pan et al. (2019), where the zero tensor element is replaced by half of the minimum of the nonzero elements. Then a logarithm transformation is conducted on predictors to make the variables on the same scale and more normally distributed. For the binary classification, a 10-fold cross validation choose $u = (3, 3, 3)$ while on the multi-class case, $u = (4, 3, 2)$ is chosen. The testing error over 100 random training/testing splits (762 training, 168 testing) are reported. Since that ADHD data consists of other covariates(age, gender and handedness) other than the tensor predictors. To adjust these covariates, we use function `adjten` from R package `catch` (see Pan et al. (2019)) to remove the effects of covariate on the predictor tensor). We report the classification results of different methods as follows. The ℓ_1-FDA was not included since it is too sensitive to tuning parameters on this dataset.

ADHD	DATE	ℓ_1-GLM	CATCH
Binary (%)	23.77(0.18)	23.64(0.16)	23.69(0.20)
Multi-calss (%)	37.74(0.22)	35.42(0.22)	36.25(0.24)

Table S2: ADHD data. Reported are the mean and standard errors (in parentheses) of the classification error.

S2 Additional Implementation Details and Numerical Results

S2.1 One-step alpgorithm

Recall that the objective function (3.8) depends on all the other $\{\mathbf{\Gamma}_j\}_{j\neq m}$ intrinsically through $\widehat{\mathbf{\Sigma}}_{-m}$ and s^{ik}. Therefore, the fully iterative algorithm can be slow. To speed up the computation and reduce the iterations, we consider a fast one-step algorithm. In the one-step algorithm, we run the iteration of the MLE equations only once. In addition, the optimization for $\ell_m(\mathbf{\Gamma}_m)$ can be further simplified and sped up by the 1D envelope algorithm Cook and Zhang (2016) that approximately optimizes $\ell_m(\mathbf{\Gamma}_m)$ but with the same theoretical properties as fully optimization.

The one-step estimator is aiming at an approximate solution for the following objective function

$$f(\mathbf{\Gamma}_m) = \log|\mathbf{\Gamma}_m^T\widehat{\mathbf{M}}_m\mathbf{\Gamma}_m| + \log|\mathbf{\Gamma}_m^T(\widehat{\mathbf{N}}_m)^{-1}\mathbf{\Gamma}_m|, \tag{S5}$$

which allows for estimation of $\{\mathbf{\Gamma}_m\}_{m=1}^M$ separately. The constrained minimization of $f(\mathbf{\Gamma}_m)$ can be done through gradient descent on a Grassmann manifold. However direct optimization is computationally expensive and requires good initial values. Therefore for a faster and more stable optimization, we adopt the 1D algorithm by Cook and Zhang (2016) with the manifold optimization method proposed by Wen and Yin (2013) to solve for $\mathbf{\Gamma}_m$. Together, this algorithm result in a one-step estimation procedure for $\mathbf{B}$ and is summarized in the Algorithm S1.

Although we derive the one-step estimator from computational consideration, we also provide theoretical justifications of it. We make the statement that under mild moment conditions, the estimation $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_m^{\text{os}}}$ obtained by one-step algorithm is $\sqrt{n}$-consistent estimator of projection onto the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{B}_{(m)})$. Then the consistency of one-step envelope based estimation follows, where we relaxed the normality assumption in the TDA model (3.1).

Theorem S1. *Assuming vec$(\mathbf{X}^i \mid Y^i = k)$ are i.i.d with finite fourth moments and the covariance estimator $\widehat{\boldsymbol{\Sigma}}_m$ is $\sqrt{n}$-consistent, $m = 1, \ldots, M$, then $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_m^{os}}$ is $\sqrt{n}$-consistent estimator for the projection onto the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{B}_{(m)})$. Moreover, the one-step estimator $\widehat{\mathbf{B}}^{os}$ is also a $\sqrt{n}$-consistent estimator for $\mathbf{B}$.*

This theorem shows that the one-step envelope estimator is indeed flexible and widely applicable from both numerical and theoretical aspects.

Algorithm S1 One-Step algorithm for estimation $\mathbf{B}$

Calculate the MLE for $\widehat{\boldsymbol{\Sigma}}_m$ as in equation (3.7).
For $m = 1, \ldots, M$, minimizing $f(\boldsymbol{\Gamma}_m)$.

- Set $\mathbf{g}_m^0 = \mathbf{G}_m^0 = 0$ and $\mathbf{G}_m^l = (\mathbf{g}_m^1, \ldots, \mathbf{g}_m^l)$ if $l \geq 1$ with $(\mathbf{G}_m^l, \mathbf{G}_{0m}^l)$ as an orthogonal matrix. For $l = 0, \ldots, u_m - 1$,
  - Solve the following objective function with constraint $\mathbf{w}^T\mathbf{w} = 1$.
  $$\mathbf{w}_{l+1} = \arg\min_{\mathbf{w}} \log[\mathbf{w}^T\{(\mathbf{G}_{0m}^l)^T\widehat{\boldsymbol{\Sigma}}_m\mathbf{G}_{0m}^l\}\mathbf{w}] + \log[\mathbf{w}^T\{(\mathbf{G}_{0m}^l)^T\widehat{\mathbf{N}}_m\mathbf{G}_{0m}^l\}^{-1}\mathbf{w}]$$
  - Set $\mathbf{g}_m^{l+1} = \mathbf{G}_{0m}^l\mathbf{w}_{l+1}$.
- Set $\widehat{\boldsymbol{\Gamma}}_m^{\text{os}} = \mathbf{G}_m^{u_m}$, and the standard MLE $\widehat{\mathbf{B}}_k^{\text{MLE}} = [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1; \widehat{\boldsymbol{\Sigma}}_1, \ldots, \widehat{\boldsymbol{\Sigma}}_M]\!]$ for $k = 2, \ldots, K$.

Set the one-step estimator $\widehat{\mathbf{B}}_k^{\text{os}} = [\![\widehat{\mathbf{B}}_k^{\text{MLE}}, \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_1^{\text{os}}}, \ldots, \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_m^{\text{os}}}]\!]$.

S2.2 Dimension Selection

As discussed in the main paper, we apply the cross-validation with grid search to choose u_m that minimizes the classification error. We also remark that Bayesian Information Criterion (BIC) is frequently considered for envelope dimension selection. However, the consistency relies on a large sample size Shao (1997); Yang (2005), which requires a large amount of samples in high-dimensional settings. Aiming at the best prediction, we use the cross-validation approach for envelope dimension selection. When estimation is the main focus, one may also directly apply the 1D-BIC approach Zhang and Mai (2018) that has good theoretical properties and also computational feasible (much faster and more stable than standard BIC). For all simulation models under the DATE model assumptions, we summarize the envelope dimension selection accuracy of DATE-L in Table S3. The envelope dimension selection is based on the five-fold cross-validation. When the sample size is sufficiently large, the dimension is accurately selected. However, cross-validation tends to under-estimate the true envelope dimension when sample sizes are small. This is not surprising since the focus is on classification error, which is a predictive criterion than a parameter estimation criterion. Arguably, in discriminant analysis, the correct specification of envelope dimension is not as crucial as in regression analysis and inference, because the focus here is on classification accuracy while the dimensions can be treated as tuning parameters.

	M1				M2		
$\{n_k\}$	(C1)	(C2)	(C3)	$\{n_k\}$	(C1)	(C2)	(C3)
$\{100, 100\}$	71	71	58	$\{150, 150\}$	76	54	76
$\{250, 250\}$	92	90	69	$\{300, 300\}$	94	93	91
	M3				M4		
$\{n_k\}$	(C1)	(C2)	(C3)	$\{n_k\}$	(C1)	(C2)	(C3)
$\{75, 75, 75, 75\}$	47	57	47	$\{60, 90, 150\}$	58	43	46
$\{150, 150, 150, 150\}$	73	70	50	$\{180, 270, 450\}$	100	100	99

Table S3: Summary of the dimension selection accuracy. For each model setup, we repeated 100 simulations and report the number of cases (out of 100) where the envelope dimension $\{u_m\}$ are all correctly selected.

S2.3 Visualization of parameter in M2

In simulation model M2 in the main paper, the parameter $\mathbf{B}$ in the Bayes has a cross shape. We show the visualization of the estimation of it by several methods. From Figure S1, it is easily seen DATE-L can recover the classification direction quite well in all cases. However, the Tucker method fails in all three cases because of the violation of model assumptions and failure on recovery of the covariance structure. RMR fails for (C2) and works for (C1) and (C3), of which the performance is worse than DATE-L. The less satisfying performance of RMR is caused by failing to use the information in the covariance.

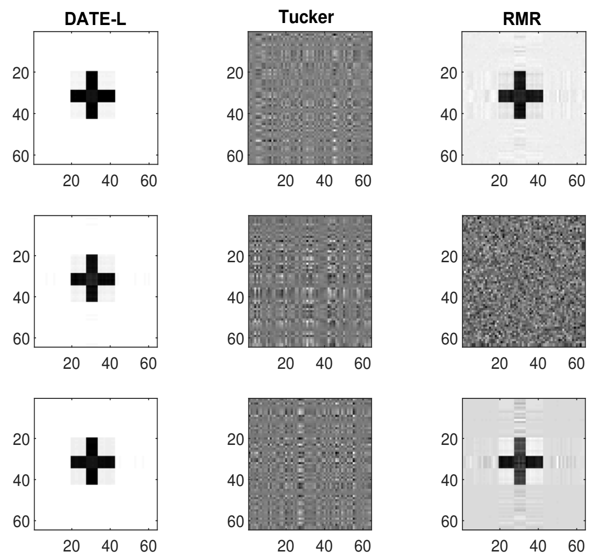

Figure S1: Visualization of the key parameter estimation for $\mathbf{B}$ under Model M2. The true coefficient is displayed on the left, which is a 64×64 matrix. The estimators are displayed on the right: from top to bottom are the three models, (C1)–(C3).

S2.4 Diagnosis procedure of DATE assumptions for Gene Time data

In our real data analysis, to see how realistic the DATE model assumptions are, we display the envelopes scores (a new concept developed in DATE-D) in Figure S2. We can see that for the first mode, the envelope scores are almost 0 except the first 10 directions (The envelope scores become small after the first 5 directions but are not very close to 0 until the 10th direction) and for the second mode, the envelope scores are almost 0 after the first direction. This result agrees with the envelope dimension selection based on the cross-validation for DATE-D, in which we select $u_1 = 7$, and $u_2 = 1$. The good classification result indicates that the subspace spanned by the eigenvectors identified by DATE-D can capture the characteristics of the data set successfully. Figure S2 and the classification result agree with Lemma 1: the subspace spanned by the eigenvectors with non-zero envelope scores is the envelope subspace, which contains most of the information for the discriminant analysis on this data set.

S3 Proofs for Lemmas and Propositions

S3.1 Proof for Lemma 1

Recall that $\mathbf{B}_{(m)} = \boldsymbol{\Sigma}_m^{-1}\{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(m)}, \cdots, (\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)_{(m)}\} \bigotimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1}$, where $\{(\boldsymbol{\mu}_2 -$

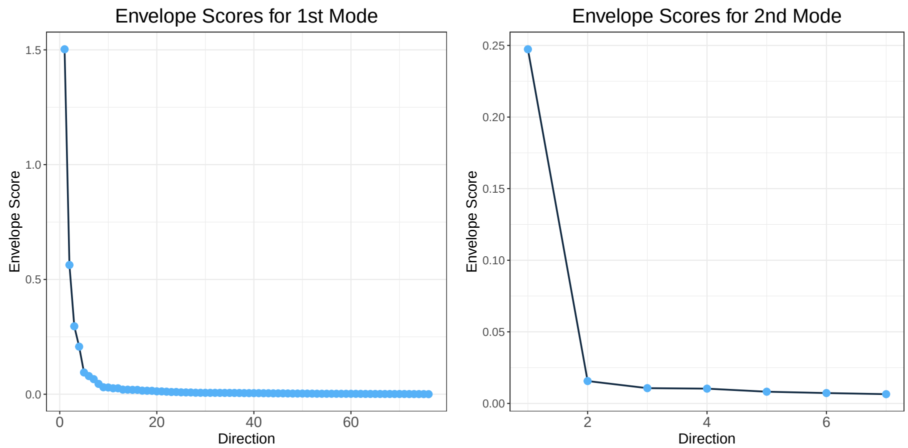

Figure S2: Envelope scores for GT datasets. The first picture shows envelope scores for the first mode that represents the genes and the second picture shows the envelope scores for the second mode that represents the time points.

$\boldsymbol{\mu}_1)_{(m)}, \cdots, (\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)_{(m)}\}$ is a matrix with dimension $p_m \times (K-1)p_{-m}$ constructed by stacking $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)_{(m)}$, $k = 2, \cdots, K$, on to its column. By Proposition 2.4 in Cook et al. (2010), we have $\mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{B}_{(m)}) = \mathcal{E}_{\boldsymbol{\Sigma}_m}(\{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(m)}, \cdots, (\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)_{(m)}\} \bigotimes_{m' \neq m} \boldsymbol{\Sigma}_{m'}^{-1})$. It follows that $\mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{B}_{(m)}) = \mathcal{E}_{\boldsymbol{\Sigma}_m}(\{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(m)}, \cdots, (\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)_{(m)}\})$. Obviously, $\mathrm{span}(\mathbf{U}_m) \subseteq \mathrm{span}(\{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(m)}, \cdots, (\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)_{(m)}\})$. For any $\boldsymbol{\nu} \in \mathrm{span}^{\perp}(\mathbf{U}_m)$, we have $\boldsymbol{\nu}^T(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})_{(m)} = 0$, for all $k = 1, \cdots, K$, which implies that $\boldsymbol{\nu}^T(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)_{(m)} = 0$ for all k. Thus, $\boldsymbol{\nu} \in \mathrm{span}^{\perp}(\{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(m)}, \cdots, (\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)_{(m)}\})$. So, we have $\mathrm{span}(\{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(m)}, \cdots, (\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)_{(m)}\}) \subseteq \mathrm{span}(\mathbf{U}_m)$. To sum up, we have $\mathrm{span}(\mathbf{U}_m) = \mathrm{span}(\{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(m)}, \cdots, (\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)_{(m)}\})$. Thus, $\mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{B}_{(m)}) = \mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{U}_m)$.

By Proposition 2.3 of Cook et al. (2010), we know that $\mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{U}_m)$ is the summation of the eigenspaces of $\boldsymbol{\Sigma}_m$ such that the projection of $\mathbf{U}_m$ onto the subspace is nonzero. By the assumption that the nonzero eigenvalues of $\mathbf{P}_{\boldsymbol{\Gamma}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\boldsymbol{\Gamma}_m}$ are distinct from those of $\mathbf{Q}_{\boldsymbol{\Gamma}_m}\boldsymbol{\Sigma}_m\mathbf{Q}_{\boldsymbol{\Gamma}_m}$, we know that a eigenspace of $\boldsymbol{\Sigma}_m$ belongs to $\mathrm{span}(\boldsymbol{\Gamma}_m)$ or $\mathrm{span}^{\perp}(\boldsymbol{\Gamma}_m)$. Thus, if an eigenvector $\boldsymbol{\nu}_j$ of $\boldsymbol{\Sigma}_m$ satisfies $\boldsymbol{\nu}_j^T\mathbf{U}_m\boldsymbol{\nu}_j \neq 0$, then $\boldsymbol{\nu}_j \in \mathrm{span}(\boldsymbol{\Gamma}_m)$. It follows that $\mathrm{span}(\boldsymbol{\Gamma}_m) = \sum_{(\mathbf{v}_i^{(m)})^T\mathbf{U}_m\mathbf{v}_i^{(m)} \neq 0} \mathrm{span}(\mathbf{v}_i^{(m)})$.

S3.2 Proof for Proposition 1

The proof can be found in Supplemental Material of Li and Zhang (2017). We omit the details here.

S3.3 Proof for Proposition 2

We first prove $\mathrm{span}(\mathbf{B}_{(M+1)}^T) \subseteq \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_M$. Recall for the Bayes' rule since $P(Y = k \mid \mathbf{X}) = P(Y = k \mid \mathbf{X} \times_1 \mathbf{P}_1 \times_2 \cdots \times_M \mathbf{P}_M)$ then we have $\phi(\mathbf{X}) = \phi(\mathbb{P}(\mathbf{X})) = \phi((\mathbf{P}_1 \otimes \cdots \otimes \mathbf{P}_m)\mathrm{vec}(\mathbf{X}))$. And $\phi(\mathbf{X}) = \phi(\mathbf{B}_{(M+1)}\mathrm{vec}(\mathbf{X}))$. If $\boldsymbol{\beta}_k = \mathrm{vec}(\mathbf{B}_k) \notin \mathcal{S}$, then the value of $\boldsymbol{\beta}_k^T\mathrm{vec}(\mathbb{Q}_{\mathcal{S}}\mathbf{X})$ will potentially change $\phi(\mathbf{X})$ which is a contradiction of $\phi(\mathbf{X}) = \phi(\mathbf{B}_{(M+1)}\mathrm{vec}(\mathbf{X}))$. Thus $\mathrm{span}(\mathbf{B}_{(M+1)}^T) \subseteq \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_M$. Then to prove $\boldsymbol{\Sigma}_m = \mathbf{P}_m\boldsymbol{\Sigma}_m\mathbf{P}_m + \mathbf{Q}_m\boldsymbol{\Sigma}_m\mathbf{Q}_m$, we only need to show $\mathbf{P}_m\boldsymbol{\Sigma}_m\mathbf{Q}_m = 0$. Since we have $(\mathbf{X} \mid Y = k) = \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}$, with $\mathrm{vec}(\boldsymbol{\varepsilon}) \sim N(0, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \ldots \otimes \boldsymbol{\Sigma}_1)$, and $(\mathbf{X} \times_m \mathbf{Q}_m \mid Y = k) \perp\!\!\!\perp (\mathbf{X} \times_m \mathbf{P}_m \mid Y = k)$, therefore $(\boldsymbol{\varepsilon} \times_m \mathbf{Q}_m) \perp\!\!\!\perp (\boldsymbol{\varepsilon} \times_m \mathbf{P}_m)$

follows. Equivalently for the vectorized case $\text{vec}(\boldsymbol{\varepsilon} \times_m \mathbf{Q}_m) \perp\!\!\!\perp \text{vec}(\boldsymbol{\varepsilon} \times_m \mathbf{P}_m)$ and $\text{cov}\{\text{vec}(\boldsymbol{\varepsilon} \times_m \mathbf{Q}_m), \text{vec}(\boldsymbol{\varepsilon} \times_m \mathbf{P}_m)\} = 0$. Thus, $\boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_{m+1} \otimes \mathbf{Q}_m \boldsymbol{\Sigma}_m \mathbf{P}_m \otimes \boldsymbol{\Sigma}_{m-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_1 = 0 \Leftrightarrow \mathbf{P}_m \boldsymbol{\Sigma}_m \mathbf{Q}_m = 0$.

The second statement is a direct conclusion from Proposition 1.

Finally, we need to prove that $\mathbf{P}_{\mathcal{S}} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{Q}_{\mathcal{S}} = 0$. By the definition (3.3), $\mathbb{P}(\mathbf{X}) \perp\!\!\!\perp \mathbb{Q}(\mathbf{X}) \mid (Y = k)$ for all $k = 1, \ldots, K$ implies that $\mathbb{P}(\mathbf{X}) \perp\!\!\!\perp \mathbb{Q}(\mathbf{X})$. Therefore $\text{cov}(\text{vec}(\mathbb{P}(\mathbf{X})), \text{vec}(\mathbb{Q}(\mathbf{X}))) = \mathbf{P}_{\mathcal{S}} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{Q}_{\mathcal{S}} = 0$. Since $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathbf{B}^T_{(M+1)})$ is the smallest reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$ that contains $\mathbf{B}^T_{(M+1)}$, therefore $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathbf{B}^T_{(M+1)}) \subseteq \mathcal{T}_{\boldsymbol{\Sigma}}(\mathbf{B})$ follows.

S3.4 Proofs for Proposition 3

Based on our likelihood function

$$
\begin{aligned}
\ell(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}) &= \log|\boldsymbol{\Sigma}| + n^{-1} \sum_{k=1}^{K} \sum_{Y^i = k} \{\text{vec}(\boldsymbol{X}^i - \boldsymbol{\mu}_k)^T\} \boldsymbol{\Sigma}^{-1} \{\text{vec}(\boldsymbol{X}^i - \boldsymbol{\mu}_k)\} \\
&= \log|\boldsymbol{\Sigma}| + n^{-1} \sum_{k=1}^{K} \sum_{Y^i = k} \{\text{vec}(\boldsymbol{X}^i - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}) - \overline{\boldsymbol{\mu}})\}^T \\
&\quad \boldsymbol{\Sigma}^{-1} \{\text{vec}(\boldsymbol{X}^i - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}) - \overline{\boldsymbol{\mu}})\}.
\end{aligned}
$$

- MLE for $\overline{\boldsymbol{\mu}}$. We take derivative for $\text{vec}(\overline{\boldsymbol{\mu}})$. Since $\text{vec}(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}) = (\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\text{vec}(\boldsymbol{\Theta}_k)$, we have

$$
0 = \sum_{k=1}^{K} \sum_{Y^i = k} \{\text{vec}(\mathbf{X}^i) - (\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\text{vec}(\boldsymbol{\Theta}_k) - \text{vec}(\overline{\boldsymbol{\mu}})\}.
$$

  therefore given $\{\widehat{\boldsymbol{\Gamma}}_m\}_{m=1}^M$, $\text{vec}(\overline{\boldsymbol{\mu}}) = \text{vec}(\overline{\mathbf{X}}) - (\widehat{\boldsymbol{\Gamma}}_M \otimes \cdots \otimes \widehat{\boldsymbol{\Gamma}}_1) \sum_{k=1}^K n_k \text{vec}(\boldsymbol{\Theta}_k)$, and $\text{vec}(\overline{\boldsymbol{\mu}}) = \text{vec}(\overline{\mathbf{X}})$ follows by the constraint $\sum_{k=1}^K \widehat{\pi}_k \text{vec}(\boldsymbol{\Theta}_k) = 0$, thus $\widehat{\overline{\boldsymbol{\mu}}} = \overline{\mathbf{X}}$.

- MLE for $\boldsymbol{\Theta}_k$. Take $\boldsymbol{\Theta}_1$ for example, we need to maxmizing

$$
-\sum_{Y^i=1} \{\text{vec}(\mathbf{X}^i) - (\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\text{vec}(\boldsymbol{\Theta}_1) - \text{vec}(\overline{\mathbf{X}})\}^T \boldsymbol{\Sigma}^{-1} \{\text{vec}(\mathbf{X}^i) - (\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\text{vec}(\boldsymbol{\Theta}_1) - \text{vec}(\overline{\mathbf{X}})\}
$$

  We decompose $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_{01}$, with

$$
\begin{aligned}
\boldsymbol{\Delta}_1 &= (\boldsymbol{\Gamma}_M \boldsymbol{\Omega}_M^{-1} \boldsymbol{\Gamma}_M^T + \boldsymbol{\Gamma}_{0M} \boldsymbol{\Omega}_{0M}^{-1} \boldsymbol{\Gamma}_{0M}) \otimes \cdots \otimes (\boldsymbol{\Gamma}_2 \boldsymbol{\Omega}_2^{-1} \boldsymbol{\Gamma}_2^T + \boldsymbol{\Gamma}_{02}^{-1} \boldsymbol{\Omega}_{02} \boldsymbol{\Gamma}_{02}^T) \otimes (\boldsymbol{\Gamma}_1 \boldsymbol{\Omega}_1^{-1} \boldsymbol{\Gamma}_1^T) \\
\boldsymbol{\Delta}_{01} &= (\boldsymbol{\Gamma}_M \boldsymbol{\Omega}_M^{-1} \boldsymbol{\Gamma}_{0M}^T + \boldsymbol{\Gamma}_{0M} \boldsymbol{\Omega}_{0M}^{-1} \boldsymbol{\Gamma}_{0M}) \otimes \cdots \otimes (\boldsymbol{\Gamma}_2 \boldsymbol{\Omega}_2^{-1} \boldsymbol{\Gamma}_2^T + \boldsymbol{\Gamma}_{02} \boldsymbol{\Omega}_{02} \boldsymbol{\Gamma}_{02}^T) \otimes (\boldsymbol{\Gamma}_{01} \boldsymbol{\Omega}_{01}^{-1} \boldsymbol{\Gamma}_{01}^T).
\end{aligned}
$$

  Then $\boldsymbol{\Delta}_{01}(\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1) = 0$ and $\boldsymbol{\Delta}_1 = \boldsymbol{\Delta}_{11} + \boldsymbol{\Delta}_{\text{res}}$, where $\boldsymbol{\Delta}_{11} = (\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1)(\boldsymbol{\Omega}_M^{-1} \otimes \cdots \otimes \boldsymbol{\Omega}_1^{-1})(\boldsymbol{\Gamma}_M^T \otimes \cdots \otimes \boldsymbol{\Gamma}_1^T)$ and $\boldsymbol{\Delta}_{\text{res}}(\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1) = 0$ by simple calculation. Therefore, we minimize over function $g(\text{vec}(\boldsymbol{\Theta}_1))$ with

$$
\begin{aligned}
g(\text{vec}(\boldsymbol{\Theta}_1)) &= \sum_{Y^i=1} \{\text{vec}(\mathbf{X}^i) - (\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\text{vec}(\boldsymbol{\Theta}_1) - \text{vec}(\overline{\mathbf{X}})\}^T \boldsymbol{\Delta}_{11} \\
&\quad \{\text{vec}(\mathbf{X}^i) - (\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\text{vec}(\boldsymbol{\Theta}_1) - \text{vec}(\overline{\mathbf{X}})\} \\
&= \sum_{Y^i=1} \{(\boldsymbol{\Gamma}_M^T \otimes \cdots \otimes \boldsymbol{\Gamma}_1^T)\text{vec}(\mathbf{X}_i - \overline{\mathbf{X}}) - \text{vec}(\boldsymbol{\Theta}_1)\}^T (\boldsymbol{\Omega}_m^{-1} \otimes \cdots \boldsymbol{\Omega}_1^{-1}) \\
&\quad \{(\boldsymbol{\Gamma}_M^T \otimes \cdots \otimes \boldsymbol{\Gamma}_1^T)\text{vec}(\mathbf{X}_i - \overline{\mathbf{X}}) - \text{vec}(\boldsymbol{\Theta}_1)\}
\end{aligned}
$$

under the constraint $\sum_{k=1}^{K}\hat{\pi}_k\boldsymbol{\Theta}_k = 0$. By ignoring the constraint, the maximizer of $g(\mathrm{vec}(\boldsymbol{\Theta}_1))$ would be $(\widehat{\boldsymbol{\Gamma}}_M^T \otimes \cdots \otimes \widehat{\boldsymbol{\Gamma}}_1^T)\mathrm{vec}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}})$, here $\overline{\mathbf{X}}_1 = \sum_{Y^i=1}\mathbf{X}^i/n_1$, thus $\widehat{\boldsymbol{\Theta}}_1 = [\![\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}; \widehat{\boldsymbol{\Gamma}}_1^T, \ldots, \widehat{\boldsymbol{\Gamma}}_m^T]\!]$. More generally, $\widehat{\boldsymbol{\Theta}}_k = [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}; \widehat{\boldsymbol{\Gamma}}_1^T, \ldots, \widehat{\boldsymbol{\Gamma}}_m^T]\!]$ for $k = 1, \ldots, K$, and this estimator automatically satisfy the constraint.

- MLE for $\boldsymbol{\Omega}_{0m}$ given $\widehat{\boldsymbol{\Gamma}}_m$ for $m = 1, \ldots, M$. First, we decompose $\log|\boldsymbol{\Sigma}|$ as

$$\log|\boldsymbol{\Sigma}| = \log|\boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1| = \sum_{\{}(\prod_{l\neq j} p_l)\log|\boldsymbol{\Sigma}_j|\} = \sum_{\{}(\prod_{l\neq j} p_l)(\log|\boldsymbol{\Omega}_j| + \log|\boldsymbol{\Omega}_{0j})|\}. \tag{S6}$$

Take $\boldsymbol{\Omega}_{01}$ as an example, we have

$$\begin{aligned}\ell(\boldsymbol{\Omega}_{01}|\{\boldsymbol{\Gamma}_j\}) &= (\prod_{l>1} p_l)\log|\boldsymbol{\Omega}_{01}| + n^{-1}\sum_{k=1}^{K}\sum_{Y^i=k}\{\mathrm{vec}(\mathbf{X}^i - \overline{\mathbf{X}} - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}))\}^T\boldsymbol{\Delta}_{01} \\ &\quad \{\mathrm{vec}(\mathbf{X}^i - \overline{\mathbf{X}} - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}))\}\end{aligned}$$

since $\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} = [\![\boldsymbol{\Theta}_k; \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}]\!]$, then $\boldsymbol{\Delta}_{01}\mathrm{vec}(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}) = 0$, and

$$\begin{aligned}&\mathrm{vec}(\mathbf{X}_i - \overline{\mathbf{X}})^T\boldsymbol{\Delta}_{01}\mathrm{vec}(\mathbf{X}_i - \overline{\mathbf{X}}) = \|\mathrm{vec}(\mathbf{X}_i - \overline{\mathbf{X}}) \times_1 \boldsymbol{\Omega}_{01}^{-1/2}\boldsymbol{\Gamma}_{01}^T \times_2 \boldsymbol{\Sigma}_2^{-1/2} \times \cdots \times_M \boldsymbol{\Sigma}_M^{-1/2}\|^2 \\ = \;& \|\mathrm{vec}(\mathbf{V}_i)\|^2 = \|(\mathbf{V}_{i(1)})\|^2.\end{aligned}$$

here $\mathbf{V}_{i(1)} = \boldsymbol{\Omega}_{01}^{-1/2}\boldsymbol{\Gamma}_{01}^T(\mathbf{X}_i - \overline{\mathbf{X}})_{(1)}(\boldsymbol{\Sigma}_M^{-1/2} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1/2})$. We have

$$\begin{aligned}\ell(\boldsymbol{\Omega}_{01}|\{\boldsymbol{\Gamma}_j\}) &= (\prod_{l>1} p_l)\log|\boldsymbol{\Omega}_{01}| + n^{-1}\mathrm{tr}(\boldsymbol{\Omega}_{01}^{-1}\sum_{k=1}^{K}\sum_{Y^i=k}\{\boldsymbol{\Gamma}_{01}^T(\mathbf{X}^i - \overline{\mathbf{X}})_{(1)}(\boldsymbol{\Sigma}_m^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1}) \\ &\quad (\mathbf{X}^i - \overline{\mathbf{X}})_{(1)}^T\boldsymbol{\Gamma}_{01}\})\end{aligned}$$

therefore we have

$$\boldsymbol{\Omega}_{01} = (n\prod_{l>1} p_l)^{-1}\sum_{k=1}^{K}\sum_{Y^i=k}\{\widehat{\boldsymbol{\Gamma}}_{01}^T(\mathbf{X}^i - \overline{\mathbf{X}})_{(1)}(\widehat{\boldsymbol{\Sigma}}_M^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_2^{-1})(\mathbf{X}^i - \overline{\mathbf{X}})_{(1)}^T\widehat{\boldsymbol{\Gamma}}_{01}\} \tag{S7}$$

In general,

$$\boldsymbol{\Omega}_{0m} = (n\prod_{l\neq m} p_m)^{-1}\sum_{i=1}^{n}\{\widehat{\boldsymbol{\Gamma}}_{0m}^T(\mathbf{X}^i - \overline{\mathbf{X}})_{(m)}(\widehat{\boldsymbol{\Sigma}}_M^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_{m+1}^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_{m-1}^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_1^{-1})(\mathbf{X}^i - \overline{\mathbf{X}})_{(m)}^T\widehat{\boldsymbol{\Gamma}}_{0m}\}. \tag{S8}$$

- MLE for $\boldsymbol{\Omega}_m$ given $\boldsymbol{\Gamma}_m$ for $m = 1, \ldots, M$. Take $\boldsymbol{\Omega}_1$ as an example, we have

$$\begin{aligned}\ell(\boldsymbol{\Omega}_1|\{\boldsymbol{\Gamma}_j\}) &= (\prod_{l>1})\log|\boldsymbol{\Omega}_1| + n^{-1}\sum_{k=1}^{K}\sum_{Y^i=k}\{\mathrm{vec}(\mathbf{X}^i - \overline{\mathbf{X}} - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}))\}^T\boldsymbol{\Delta}_1 \\ &\quad \{\mathrm{vec}(\mathbf{X}^i - \overline{\mathbf{X}} - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}))\}\end{aligned}$$

and

$$\begin{aligned} & \text{vec}(\mathbf{X}^i - \overline{\mathbf{X}} - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})) = \|\text{vec}(\mathbf{X}_i - \overline{\mathbf{X}} - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})) \times_1 \boldsymbol{\Omega}_1^{-1/2}\boldsymbol{\Gamma}_1^T \times_2 \boldsymbol{\Sigma}_2^{-1/2} \times \cdots \times_M \boldsymbol{\Sigma}_M^{-1/2}\|^2 \\ = & \|\text{vec}(\mathbf{U}_i)\|^2 = \|(\mathbf{U}_{i(1)})\|^2. \end{aligned}$$

here $\mathbf{U}_{i(1)} = \boldsymbol{\Omega}_1^{-1/2}\boldsymbol{\Gamma}_1^T(\mathbf{X}_i - \overline{\mathbf{X}} - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}))_{(1)}(\boldsymbol{\Sigma}_M^{-1/2} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1/2})$. We have

$$\ell(\boldsymbol{\Omega}_1|\{\boldsymbol{\Gamma}_j\}) = (\prod_{l>1} p_l) \log |\boldsymbol{\Omega}_1| + n^{-1}\text{tr}(\boldsymbol{\Omega}_1^{-1} \sum_{k=1}^{K} \sum_{Y^i=k} \{\boldsymbol{\Gamma}_1^T \mathbf{e}_{i(1)}(\boldsymbol{\Sigma}_m^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1})\mathbf{e}_{i(1)}\boldsymbol{\Gamma}_1\})$$

where $\mathbf{e}_i = (\mathbf{X}_i - \overline{\mathbf{X}} - (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}))$

$$\begin{aligned} \boldsymbol{\Gamma}_1^T \mathbf{e}_{i(1)} &= \boldsymbol{\Gamma}_1^T(\mathbf{X}_i - \overline{\mathbf{X}} - [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}; \mathbf{P}_{\boldsymbol{\Gamma}_1}, \cdots, \mathbf{P}_{\boldsymbol{\Gamma}_M}]\!])_{(1)} \\ &= \{(\mathbf{X}_i - \overline{\mathbf{X}}) \times_1 \boldsymbol{\Gamma}_1^T - (\overline{\mathbf{X}}_k - \overline{\mathbf{X}}) \times_1 \boldsymbol{\Gamma}_1^T \cdots \times_M \mathbf{P}_{\boldsymbol{\Gamma}_M}\}_{(1)} \\ &= \{[(\mathbf{X}_i - \overline{\mathbf{X}}) - (\overline{\mathbf{X}}_k - \overline{\mathbf{X}}) \times_1 \mathbf{I}_{p_1} \cdots \times_M \mathbf{P}_{\boldsymbol{\Gamma}_M}] \times_1 \boldsymbol{\Gamma}_1^T\}_{(1)} \\ &= \boldsymbol{\Gamma}_1^T \mathbf{s}_{(1)}^i \end{aligned}$$

Therefore $\widehat{\boldsymbol{\Omega}}_1 = (n\prod_{l>1} p_l)^{-1} \sum_{k=1}^K \sum_{Y^i=k} \{\widehat{\boldsymbol{\Gamma}}_1^T \mathbf{s}_{(1)}^i(\widehat{\boldsymbol{\Sigma}}_m^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_2^{-1})(\mathbf{s}_{(1)}^i)^T\widehat{\boldsymbol{\Gamma}}_1\}$. More generally, we have

$$\widehat{\boldsymbol{\Omega}}_m = (n\prod_{m\neq j} p_m)^{-1} \sum_{k=1}^{K} \sum_{Y^i=K} \{\widehat{\boldsymbol{\Gamma}}_m^T \mathbf{s}_{(m)}^{ik}(\widehat{\boldsymbol{\Sigma}}_m^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_{m+1}^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_{m-1}^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_1^{-1})(\mathbf{s}_{(m)}^{ik})^T\widehat{\boldsymbol{\Gamma}}_m\}$$

- MLE for $\boldsymbol{\Gamma}_m$ for $m = 1, \ldots, M$. The partially maximized log-likelihood satisfies

$$\ell(\boldsymbol{\Gamma}_m) = \log |\boldsymbol{\Gamma}_m^T \widehat{\mathbf{M}}_m \boldsymbol{\Gamma}_m| + \log |\boldsymbol{\Gamma}_m^T (\widehat{\mathbf{N}}_m)^{-1} \boldsymbol{\Gamma}_m| \tag{S9}$$

where $\boldsymbol{M}_m$ and $\boldsymbol{N}_m$ are defined as

$$\begin{aligned} \widehat{\boldsymbol{M}}_m &= (n\prod_{l\neq m} p_l)^{-1} \sum_{k=1}^{K} \sum_{Y^i=k} \{\mathbf{s}_{(m)}^i(\widehat{\boldsymbol{\Sigma}}_m^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_{m+1}^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_{m-1}^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_1^{-1})(\mathbf{s}_{(m)}^i)^T\}, \\ \widehat{\boldsymbol{N}}_m &= (n\prod_{l\neq m} p_l)^{-1} \sum_{k=1}^{K} \sum_{Y^i=k} \{(\mathbf{X}^i - \overline{\mathbf{X}})_{(m)}(\widehat{\boldsymbol{\Sigma}}_M^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_{m+1}^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_{m-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_1)(\mathbf{X}^i - \overline{\mathbf{X}})_{(m)}^T\} \\ &= \frac{1}{n\prod_{l\neq j} p_l} \sum_{n=1}^{n} (\mathbf{X}^i - \overline{\mathbf{X}})_{(m)}\{\widehat{\boldsymbol{\Sigma}}_m^{-1} \otimes \cdots \widehat{\boldsymbol{\Sigma}}_{(m+1)}^{-1} \otimes \widehat{\boldsymbol{\Sigma}}_{(m-1)}^{-1} \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_1^{-1}\}(\mathbf{X}^i - \overline{\mathbf{X}})_{(m)}^T. \end{aligned}$$

The by plugging in all the envelope estimator, we have

$$\begin{aligned} \boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} &= [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}; \widehat{\mathbf{P}}_1, \ldots, \widehat{\mathbf{P}}_M]\!] \\ \widehat{\boldsymbol{B}}_k &= [\![\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1; \widehat{\boldsymbol{\Sigma}}_1^{-1}, \ldots, \widehat{\boldsymbol{\Sigma}}_m^{-1}]\!] \\ &= [\![[\![(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1); \widehat{\mathbf{P}}_1, \ldots, \widehat{\mathbf{P}}_M]\!]; \widehat{\boldsymbol{\Gamma}}_1\widehat{\boldsymbol{\Omega}}_1^{-1}\widehat{\boldsymbol{\Gamma}}_1^T + \widehat{\boldsymbol{\Gamma}}_{01}\widehat{\boldsymbol{\Omega}}_{01}^{-1}\widehat{\boldsymbol{\Gamma}}_{01}^T, \cdots, \widehat{\boldsymbol{\Gamma}}_m\widehat{\boldsymbol{\Omega}}_m^{-1}\widehat{\boldsymbol{\Gamma}}_m^T + \widehat{\boldsymbol{\Gamma}}_{0m}\widehat{\boldsymbol{\Omega}}_{0m}^{-1}\widehat{\boldsymbol{\Gamma}}_{0m}^T]\!] \\ &= [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1; \widehat{\boldsymbol{\Gamma}}_1\widehat{\boldsymbol{\Omega}}_1^{-1}\widehat{\boldsymbol{\Gamma}}_1^T, \ldots, \widehat{\boldsymbol{\Gamma}}_M\widehat{\boldsymbol{\Omega}}_M^{-1}\widehat{\boldsymbol{\Gamma}}_M^T]\!]. \end{aligned}$$

S4 Proofs for Theorems

S4.1 Proof for Theorem S1

Firstly, we need the $\sqrt{n}$-consistency of the estimator $\widehat{\boldsymbol{\Sigma}}_m$ used in the one-step algorithm. Assume $(\mathbf{X}^i, Y^i), i = 1, \ldots, n$ are independent and identically distributed according to TDA model in section 3.1, we prove that the estimator $\widehat{\boldsymbol{\Sigma}}_m$ in (3.7) is $\sqrt{n}$-consistent estimate for $\boldsymbol{\Sigma}_m, m = 1, \ldots, M$.

First we note that given $Y = k$, the mode-m matricization $\widetilde{\mathbf{X}}^i_{(m)} \equiv (\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)} \in \mathbb{R}^{p_m \times (\prod_{l \neq m} p_l)}, i = 1, \ldots, n$ are i.i.d. from a matrix normal distribution with mean 0 and covariances $\boldsymbol{\Sigma}_m$ (row covariance matrix) and $\boldsymbol{\Delta}_m \equiv \boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_{m+1} \otimes \boldsymbol{\Sigma}_{m-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_1$ (the column covariance matrix). Then follows Gupta and Nagar (2018), we have the following second-order moments:

$$E(\widetilde{\mathbf{X}}^i_{(m)}\{\widetilde{\mathbf{X}}^i_{(m)}\}^T | Y = k) = \boldsymbol{\Sigma}_m \mathrm{tr}(\boldsymbol{\Delta}_m)$$

where $\mathrm{tr}(\cdot)$ is the trace operator of matrices. Next, we note that the following starting value is a $\sqrt{n}$-consistent for $\boldsymbol{\Sigma}_m$ up to a scaler difference due to $\mathrm{tr}(\boldsymbol{\Delta}_m)$,

$$\widehat{\boldsymbol{\Sigma}}_m = \frac{1}{n \prod_{l \neq m} p_l} \sum_{k=1}^{K} \sum_{Y^i = k} (\mathbf{X}^i - \overline{\mathbf{X}}_k)_{(m)} (\mathbf{X}^i - \overline{\mathbf{X}}_k)^T_{(m)}.$$

The scalar difference can be resolved after normalizing $\widehat{\boldsymbol{\Sigma}} = \tau \widehat{\boldsymbol{\Sigma}}_m \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_1$ according to the scaler

$$\tau = \frac{1}{n \prod_{l \neq m} p_l} \sum_{k=1}^{K} \sum_{Y^i = k} \mathrm{vec}(\widetilde{\mathbf{X}}^i) \widehat{\boldsymbol{\Sigma}}^{-1} \mathrm{vec}(\widetilde{\mathbf{X}}^i)^T$$

at the end of iteration. Therefore, $\widehat{\boldsymbol{\Sigma}}_m$ in (3.7) is a $\sqrt{n}$-consistent estimator for $\boldsymbol{\Sigma}_m$.

From Cook and Zhang (2016) (Proposition 6) we know that if $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are $\sqrt{n}$-consistent estimators for $\mathbf{M} > 0$ and $\mathbf{U} \geq 0$, the the 1D algorithm minimizer $\widehat{\mathbf{G}} = \underset{\mathbf{G}^T\mathbf{G} = \mathbf{I}_u}{\arg\min} \log |\mathbf{G}^T \widehat{\mathbf{M}} \mathbf{G}| + \log |\mathbf{G}^T (\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1} \mathbf{G}|$ produce $\sqrt{n}$-consistent estimator for the projection onto the envelope $\mathcal{E}_M(\mathbf{U})$. If we look at the objective function (3.8), $\widehat{\boldsymbol{\Sigma}}_m$ is a $\sqrt{n}$-consistent estimator for $\boldsymbol{\Sigma}_m$ and $\widehat{\boldsymbol{N}}_m = (n \prod_{j \neq m} p_j)^{-1} \sum_{i=1}^{n} (\mathbf{X}^i - \overline{\mathbf{X}})_{(m)} \widehat{\boldsymbol{\Sigma}}^{-1}_{-m} (\mathbf{X}^i - \overline{\mathbf{X}})^T_{(m)}$ is analogous to $(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})$. Then we have

$$\widehat{\boldsymbol{N}}_m - \widehat{\boldsymbol{\Sigma}}_m = (\prod_{j \neq m} p_j)^{-1} \sum_{k=1}^{K} \widehat{\pi}_k (\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)} \widehat{\boldsymbol{\Sigma}}^{-1}_{-m} (\overline{\mathbf{X}}_k - \overline{\mathbf{X}})^T_{(m)} \tag{S10}$$

which is analogous to $\widehat{\mathbf{U}}$. If we define $\widetilde{\boldsymbol{\mu}} = \{\overline{\boldsymbol{\mu}}_1 - \overline{\boldsymbol{\mu}}, \ldots, \overline{\boldsymbol{\mu}}_K - \overline{\boldsymbol{\mu}}\} \in \mathbb{R}^{p_1 \times \cdots \times p_M \times K}$ and $\mathbf{D} = \mathrm{diag}(\pi_1, \ldots, \pi_K)$, $\mathbf{C}^m = [\![\widetilde{\boldsymbol{\mu}}; \boldsymbol{\Sigma}_1^{-1/2}, \ldots, \boldsymbol{\Sigma}_{m-1}^{-1/2}, \mathbf{I}_{p_m}, \boldsymbol{\Sigma}_{m+1}^{-1/2}, \ldots, \boldsymbol{\Sigma}_M^{-1/2}, \mathbf{D}^{1/2}]\!]$, then $\widehat{\boldsymbol{N}}_m - \widehat{\boldsymbol{\Sigma}}_m = (\prod_{j \neq m} p_j)^{-1} \{\widehat{\mathbf{C}}^m\}_{(m)} \{\widehat{\mathbf{C}}^m\}^T_{(m)}$, where

$$\widehat{\mathbf{C}}^m = [\![\widetilde{\mathbf{X}}; \widehat{\boldsymbol{\Sigma}}_1^{-1/2}, \ldots, \widehat{\boldsymbol{\Sigma}}_{m-1}^{-1/2}, \mathbf{I}_{p_m}, \widehat{\boldsymbol{\Sigma}}_{m+1}^{-1/2}, \ldots, \widehat{\boldsymbol{\Sigma}}_M^{-1/2}, \widehat{\mathbf{D}}^{1/2}]\!].$$

with $\widetilde{\mathbf{X}} = \{\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}, \ldots, \overline{\mathbf{X}}_K - \overline{\mathbf{X}}\} \in \mathbb{R}^{p_1 \times \cdots \times p_M \times (K-1)}$. We claim that $\widehat{\boldsymbol{N}}_m - \widehat{\boldsymbol{\Sigma}}_m$ is a $\sqrt{n}$-consistent estimators for $\{\mathbf{C}^m\}_{(m)} \{\mathbf{C}^m\}^T_{(m)}$ up to an scaling constant $(\prod_{j \neq m} p_j)^{-1}$. Moreover, denote $\widetilde{\boldsymbol{\mu}}^* = \{\overline{\boldsymbol{\mu}}_2 - \overline{\boldsymbol{\mu}}_1, \dots, \overline{\boldsymbol{\mu}}_K - \overline{\boldsymbol{\mu}}_1\} \in \mathbb{R}^{p_1 \times \cdots \times p_M \times (K-1)}$ and $\mathbf{D}^* = \mathrm{diag}(\pi_2/\pi_1, \dots, \pi_K/\pi_1)$, we have $\mathrm{span}(\{\mathbf{C}^m\}_{(m)}\{\mathbf{C}^m\}_{(m)}^T) = \mathrm{span}(\{\mathbf{B}^m\}_{(m)}\{\mathbf{B}^m\}_{(m)}^T)$, where

$$\{\mathbf{B}^m\} = [\![\widetilde{\boldsymbol{\mu}}^*; \boldsymbol{\Sigma}_1^{-1/2}, \dots, \boldsymbol{\Sigma}_{m-1}^{-1/2}, \mathbf{I}_{p_m}, \boldsymbol{\Sigma}_{m+1}^{-1/2}, \dots, \boldsymbol{\Sigma}_M^{-1/2}, \{\mathbf{D}^*\}^{1/2}]\!].$$

Therefore, by Proposition 6 in Cook and Zhang (2016), $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_m^{\text{os}}}$ is a $\sqrt{n}$-consistent estimator for the projection onto the envelope $\mathcal{E}_{\widehat{\boldsymbol{\Sigma}}_m}(\{\mathbf{C}^m\}_{(m)}\{\mathbf{C}^m\}_{(m)}^T) = \mathcal{E}_{\boldsymbol{\Sigma}_m}(\{\mathbf{B}^m\}_{(m)}\{\mathbf{B}^m\}_{(m)}^T)$. By the definition of $\{\mathbf{B}_k^m\}_{(m)}$ and the property of Tucker operator, we have

$$\{\mathbf{B}^m\}_{(m)} = \{\mathbf{B}\}_{(m)}(\boldsymbol{\Sigma}_1^{-1/2} \otimes \cdots \otimes \boldsymbol{\Sigma}_{m-1}^{1/2} \otimes \boldsymbol{\Sigma}_{m+1}^{1/2} \otimes \cdots \otimes \boldsymbol{\Sigma}_M^{1/2} \otimes \mathbf{I}_{K-1}) \tag{S11}$$

which implies $\mathrm{span}(\{\mathbf{B}^m\}_{(m)}) = \mathrm{span}(\mathbf{B}_{(m)})$ and hence $\mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathrm{span}(\{\mathbf{B}^m\}_{(m)})) = \mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathrm{span}(\mathbf{B}_{(m)})$. Since $\widehat{\mathbf{B}}^{\text{os}} = [\![\widehat{\mathbf{B}}_{\text{MLE}}, \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_1^{\text{os}}}, \dots, \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_m^{\text{os}}}]\!]$, therefore $\widehat{\mathbf{B}}^{\text{os}}$ is also a $\sqrt{n}$-consistent estimator for $\mathbf{B}$.

S4.2 Proofs for Theorem 1

Since we have

$$\mathbf{h} = \begin{pmatrix} \{\boldsymbol{\beta}_k\}_{k=2}^K \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix}, \quad \boldsymbol{\phi} = \begin{pmatrix} \{\boldsymbol{\beta}_k\}_{k=2}^K \\ \{\mathrm{vech}(\boldsymbol{\Sigma}_m)\}_{m=1}^M \end{pmatrix}, \quad \boldsymbol{\xi} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Gamma}_m)_{m=1}^M \\ \mathrm{vec}(\boldsymbol{\eta}_k)_{k=2}^K \\ \mathrm{vech}(\boldsymbol{\Omega}_m)_{m=1}^M \\ \mathrm{vech}(\boldsymbol{\Omega}_{0m})_{m=1}^M \end{pmatrix}. \tag{S12}$$

$\mathbf{h}_{\text{MLE}} = \mathbf{h}(\widehat{\boldsymbol{\phi}})$ contains the MLE estimator of $\boldsymbol{\mu}_k$ and the Kronecker covariance estimator from Manceur and Dutilleul (2013). Thus $\widehat{\mathbf{h}}_{\text{MLE}} = \mathbf{h}(\widehat{\boldsymbol{\phi}})$ and $\widehat{\mathbf{h}}_{\text{env}} = \mathbf{h}(\widehat{\boldsymbol{\xi}})$ is the MLE under the envelope assumption. Since $\mathbf{h} = \mathbf{h}(\boldsymbol{\phi}) = \mathbf{h}(\boldsymbol{\epsilon})$ is overparameterized, by the conclusion of Shapiro (1986), $\sqrt{n}(\widehat{\mathbf{h}}_{\text{MLE}} - \mathbf{h}_{\text{True}}) \to N(0, \mathbf{U})$ with $\mathbf{U} = \mathbf{H}(\mathbf{H}^T\mathbf{J}_h\mathbf{H})^{\dagger}\mathbf{H}^T$ and $\sqrt{n}(\widehat{\mathbf{h}}_{\text{env}} - \mathbf{h}_{\text{True}}) \to N(0, \mathbf{V})$ with $\mathbf{V} = \mathbf{K}(\mathbf{K}^T\mathbf{J}_h\mathbf{K})^{\dagger}\mathbf{K}^T$. Matrix $\mathbf{J}_h$ is the inverse of covariance matrix of the standard MLE of TDA model we denote as $\widehat{\mathbf{h}}^* = ((\widehat{\boldsymbol{\beta}}_2^*)^T, \dots, (\widehat{\boldsymbol{\beta}}_K^*)^T, \mathrm{vech}(\mathbf{S})^T)$, where $\mathbf{S}$ is MLE under vectorized tensor LDA model and $\widehat{\boldsymbol{\beta}}_2^* = \mathbf{S}^{-1}\mathrm{vec}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$. The asymptotic covariance matrix of $\widetilde{\boldsymbol{\psi}}_1 = (\mathrm{vec}(\overline{\mathbf{X}}_1)^T, \dots, \mathrm{vec}(\overline{\mathbf{X}}_K)^T, \mathrm{vech}(\mathbf{S})^T)^T$ is $\mathbf{J}_1^{-1}$, with $\mathbf{J}_1 = \mathrm{diag}\{\pi_1\boldsymbol{\Sigma}^{-1}, \dots, \pi_K\boldsymbol{\Sigma}^{-1}, \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\}$, here $p = \prod_{m=1}^M p_m$ and $\mathbf{E}_p \in \mathbb{R}^{p^2 \times p(p+1)/2}$ is the expansion matrix such that $\mathrm{vec}(\boldsymbol{\Sigma}) = \mathbf{E}_p\mathrm{vech}(\boldsymbol{\Sigma})$. Then the asymptotic covariance $\mathbf{J}_2^{-1}$ of $\widetilde{\boldsymbol{\psi}}_2 = (\mathrm{vec}(\overline{\mathbf{X}}_2 - \overline{\mathbf{X}}_1)^T, \dots, \mathrm{vec}(\overline{\mathbf{X}}_K - \overline{\mathbf{X}}_1)^T, \mathrm{vech}(\mathbf{S})^T)^T$ is $\mathbf{J}_2^{-1} = \mathbf{L}_1\mathbf{J}_1^{-1}\mathbf{L}_1^T$, where the $\mathbf{L}_1 = \mathrm{diag}\{(-\mathbf{1}_{K-1}, \mathbf{I}_{K-1}) \times \mathbf{I}_p, \mathbf{I}_{p(p+1)/2}\}$ is the linear map between the estimators $\widetilde{\boldsymbol{\psi}}_2 = \mathbf{L}_1 \times \widetilde{\boldsymbol{\psi}}_1$. Thus $\mathbf{J}_2^{-1} = \mathrm{diag}\left[\mathbf{A} \otimes \boldsymbol{\Sigma}, \{\frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\}^{-1}\right]$, where $\mathbf{A} \in \mathbb{R}^{(K-1)\times(K-1)}$ is a symmetric matrix has diagonal elements $(\pi_k^{-1} + \pi_1^{-1})$ and off-diagonal π_1^{-1}, and $\mathbf{A}^{-1}$ has diagonal elements $\pi_i(1-\pi_i)$ and $-\pi_i\pi_j$ for off-diagonal elements. Then the asymptotic covariance $\widehat{\mathbf{h}}^*$ can be calculated from one to one mapping from $\boldsymbol{\psi}_2$ to $\mathbf{h}$. Since $\mathrm{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) = \boldsymbol{\Sigma}\boldsymbol{\beta}_k$, then

$$\frac{\partial \mathrm{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)}{\partial \boldsymbol{\beta}_k} = \boldsymbol{\Sigma}, \quad \frac{\partial \mathrm{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)}{\partial(\mathrm{vech}(\boldsymbol{\Sigma}))} = (\boldsymbol{\beta}_k^T \otimes \mathbf{I}_p)\mathbf{E}_p \tag{S13}$$

Thus $\mathbf{J}_h = \mathbf{G}\mathbf{J}_2\mathbf{G}$, where $\mathbf{G} = \frac{\partial \psi_2}{\partial \mathbf{h}} = \begin{pmatrix} \mathbf{I}_{(K-1)p} \otimes \boldsymbol{\Sigma} & (\boldsymbol{\beta}^T \otimes \mathbf{I}_p)\mathbf{E}_p \\ 0 & \mathbf{I}_{p(p+1)/2} \end{pmatrix}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p\times(K-1)}$. A straight forward calculation shows that

$$\mathbf{J}_h = \begin{pmatrix} \mathbf{A}^{-1} \otimes \boldsymbol{\Sigma} & (\mathbf{A}^{-1}\boldsymbol{\beta}^T \otimes \mathbf{I}_p)\mathbf{E}_p \\ \mathbf{E}_p^T(\boldsymbol{\beta}\mathbf{A}^{-1} \otimes \mathbf{I}_p) & \mathbf{E}_p^T(\boldsymbol{\beta}\mathbf{A}^{-1}\boldsymbol{\beta}^T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p + \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \end{pmatrix} \tag{S14}$$

Therefore the asymptotic covariance $\widehat{\mathbf{h}}^*$ is $\mathbf{J}_h^{-1} = \mathbf{W}$.

Moreover $\mathbf{H} = \frac{\partial \mathbf{h}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \begin{pmatrix} \mathbf{I}_{(K-1)p} & 0 \\ 0 & (\frac{\partial \text{vech}(\boldsymbol{\Sigma})}{\partial \text{vech}(\boldsymbol{\Sigma}_1)}, \cdots, \frac{\partial \text{vech}(\boldsymbol{\Sigma})}{\partial \text{vech}(\boldsymbol{\Sigma}_m)}) \end{pmatrix}$. And $\mathbf{K} = \frac{\partial \mathbf{h}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\epsilon}}$, for the element of $\mathbf{K}$, we have

$$
\begin{aligned}
\frac{\partial \boldsymbol{\beta}_k}{\partial \text{vec}(\boldsymbol{\eta}_k)} &= (\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1) && \text{(S15)} \\
\frac{\partial \boldsymbol{\beta}_k}{\partial \text{vec}(\boldsymbol{\Gamma}_m)} &= \frac{\partial (\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\text{vec}(\boldsymbol{\eta}_k)}{\partial \text{vec}(\boldsymbol{\Gamma}_m)} && \text{(S16)} \\
\frac{\partial \boldsymbol{\beta}_k}{\partial \boldsymbol{\Omega}_m} &= \frac{\partial \boldsymbol{\beta}_k}{\partial \boldsymbol{\Omega}_{0m}} = 0, && \text{(S17)} \\
\frac{\partial \text{vech}(\boldsymbol{\Sigma})}{\partial \text{vech}(\boldsymbol{\Sigma}_m)} &= \mathbf{C}_p \frac{\partial \text{vec}(\boldsymbol{\Sigma})}{\partial \text{vec}(\boldsymbol{\Sigma}_m)} \mathbf{E}_{p_m}, && \text{(S18)}
\end{aligned}
$$

where $\mathbf{C}_p \in \mathbb{R}^{p(p+1)/2 \times p^2}$ is the contraction matrix such that $\text{vech}(\boldsymbol{\Sigma}) = \mathbf{C}_p \text{vec}(\boldsymbol{\Sigma})$. Notice in (S18). For $m = 1$ and $m = M$, we can use the formulas from Fackler (2005) for the derivatives. For $1 \leq m \leq M-1$, there is only elementwise derivatives.

We have $\mathbf{H}(\mathbf{H}^T \mathbf{J}_h \mathbf{H})^{\dagger} \mathbf{H}^T = \mathbf{J}_h^{-1/2} \mathbf{H}(\mathbf{H}^T \mathbf{J}_h^{1/2} \mathbf{J}_h^{1/2} \mathbf{H})^{\dagger} \mathbf{H}^T \mathbf{J}_h^{1/2} \mathbf{J}_h^{-1/2} = \mathbf{J}_h^{-1/2} \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}} \mathbf{J}_h^{-1/2}$ and similarly $\mathbf{K}(\mathbf{K}^T \mathbf{J}_h \mathbf{K})^{\dagger} \mathbf{K}^T = \mathbf{J}_h^{-1/2} \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}} \mathbf{J}_h^{-1/2}$, and $\mathbf{K} = \partial \mathbf{h}(\boldsymbol{\epsilon})/\partial \boldsymbol{\epsilon} = \partial \mathbf{h}(\boldsymbol{\phi})/\partial \boldsymbol{\phi} \times \partial \boldsymbol{\phi}(\boldsymbol{\epsilon}) \partial \boldsymbol{\epsilon} = \mathbf{H} \times \partial \boldsymbol{\phi}(\boldsymbol{\epsilon}) \partial \boldsymbol{\epsilon}$. Thus $\text{span}(\mathbf{J}_h^{1/2}\mathbf{K}) \subseteq \text{span}(\mathbf{J}_h^{1/2}\mathbf{H})$, with $\mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}} = \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}} \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}} = \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{K}} \mathbf{P}_{\mathbf{J}_h^{1/2}\mathbf{H}}$.

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

Therefore, since $\text{vec}(\boldsymbol{\beta})$ is the upper left block of $\mathbf{h}$, the result of Theorem 1 follows.

S4.3 Proofs for Theorem 2

Given fixed $\boldsymbol{\Gamma}$, we have $\boldsymbol{\beta}_k = (\boldsymbol{\Gamma}_M \boldsymbol{\Omega}_M^{-1} \boldsymbol{\Gamma}_M^T \otimes \cdots \otimes \boldsymbol{\Gamma}_1 \boldsymbol{\Omega}_1^{-1} \boldsymbol{\Gamma}_1^T) \text{vec}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$. Since $(\text{vec}(\overline{\mathbf{X}}_2 - \overline{\mathbf{X}}_1)^T, \ldots, \text{vec}(\overline{\mathbf{X}}_K - \overline{\mathbf{X}}_1)^T)^T \sim N(0, \mathbf{A} \otimes \boldsymbol{\Sigma})$, therefore $\text{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ has covariance matrix $\mathbf{A} \otimes [(\boldsymbol{\Gamma}_M \boldsymbol{\Omega}_M^{-1} \boldsymbol{\Gamma}_M) \otimes \cdots \otimes (\boldsymbol{\Gamma}_1 \boldsymbol{\Omega}_1^{-1} \boldsymbol{\Gamma}_1)]$. The conclusion follows.

S4.4 Proofs for Theorem 3

Recall that $\ell(\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k) = \log|\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{\Sigma}^{-1} \frac{\sum_{k=1}^K \sum_{Y^i = k} \text{vec}(\mathbf{X}^i - \boldsymbol{\mu}_k) \text{vec}(\mathbf{X}^i - \boldsymbol{\mu}_k)^T}{n})$. We define a minimum discrepancy function,

$$
F(\mathbf{h}, \widehat{\mathbf{h}}^*) = \log|\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}) - \log|\mathbf{S}| - \text{tr}(\mathbf{S}^{-1}\mathbf{S}), \tag{S19}
$$

where $\mathbf{S}$ is the MLE under vectorized tensor LDA model without separable covariance structure assumption for $\boldsymbol{\Sigma}$. Also $\mathbf{h} = \mathbf{h}(\boldsymbol{\phi}) = \mathbf{h}(\boldsymbol{\xi})$, and $\widehat{\mathbf{h}}^*$ the MLE under the vectorized tensor LDA model without separable covariance structure assumption on $\boldsymbol{\Sigma}$. Since $\widehat{\mathbf{h}}^*$ is a smooth function of sample estimators, and the sample estimators are asymptotically normal by Central Limit Theorem. Therefore by delta method, we have $\mathrm{avar}(\sqrt{n}\widehat{\mathbf{h}}^*) = \boldsymbol{\Xi}$. Since (i) $F(\mathbf{h}, \widehat{\mathbf{h}}^*) \geq 0$ for all $\mathbf{h}$ and $\widehat{\mathbf{h}}^*$; (ii) $F(\mathbf{h}, \widehat{\mathbf{h}}^*) = 0$ if and only if $\mathbf{h} = \widehat{\mathbf{h}}^*$; (iii) $F(\mathbf{h}, \widehat{\mathbf{h}}^*)$ is twice continuously differentiable in $\mathbf{h}$ and $\widehat{\mathbf{h}}^*$. Then by Proposition 4.1 of Shapiro (1986), we have consistency results of $\widehat{\mathbf{h}}_{\text{MLE}}$ and $\widehat{\mathbf{h}}_{\text{env}}$. Moreover, we have $\mathbf{W} = \mathrm{avar}(\sqrt{n}\widehat{\mathbf{h}}_{\text{MLE}}) = \mathbf{H}(\mathbf{H}^T\mathbf{J}_\mathbf{h}\mathbf{H})^\dagger\mathbf{H}^T\mathbf{J}_\mathbf{h}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}\mathbf{H}(\mathbf{H}^T\mathbf{J}_\mathbf{h}\mathbf{H})^\dagger\mathbf{H}^T$ and $\mathbf{U} = \mathrm{avar}(\sqrt{n}\widehat{\mathbf{h}}_{\text{env}}) = \mathbf{K}(\mathbf{K}^T\mathbf{J}_\mathbf{h}\mathbf{K})^\dagger\mathbf{K}^T\mathbf{J}_\mathbf{h}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}\mathbf{K}(\mathbf{K}^T\mathbf{J}_\mathbf{h}\mathbf{K})^\dagger\mathbf{K}^T$.

We have

$$\begin{aligned}
\boldsymbol{\Xi} - \mathbf{W} &= \boldsymbol{\Xi} - \mathbf{H}(\mathbf{H}^T\mathbf{J}_\mathbf{h}\mathbf{H})^\dagger\mathbf{H}^T\mathbf{J}_\mathbf{h}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}\mathbf{H}(\mathbf{H}^T\mathbf{J}_\mathbf{h}\mathbf{H})^\dagger\mathbf{H}^T \\
&= \mathbf{J}_\mathbf{h}^{-1/2}\left[\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2} - \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}}\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2}\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}}\right]\mathbf{J}_h^{-1/2} \\
\mathbf{W} - \mathbf{U} &= \mathbf{J}_\mathbf{h}^{-1/2}\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}}\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2}\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}}\mathbf{J}_h^{-1/2} - \mathbf{J}_\mathbf{h}^{-1/2}\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}}\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2}\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}}\mathbf{J}_\mathbf{h}^{-1/2} \\
&= \mathbf{J}_\mathbf{h}^{-1/2}(\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}}\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2}\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}} - \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}}\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2}\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}})\mathbf{J}_\mathbf{h}^{-1/2} \\
&= \mathbf{J}_\mathbf{h}^{-1/2}\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}}(\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2} - \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}}\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2}\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}})\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}}\mathbf{J}_\mathbf{h}^{-1/2}.
\end{aligned}$$

The conclusion follows.

S5 Proof for Theorem 4

S5.1 Auxiliary Lemmas

Some auxiliary lemmas are presented in this section as the preparation for proving Theorem 4. We first review the technical assumptions presented in Section 4.2 of the main paper.

- (A1) The eigenvalues of $\boldsymbol{\Sigma}_m$, $m = 1, \cdots, M$, are all bounded between positive constants c_1 and c_2.
- (A2) The smallest non-zero eigenvalue of $\mathbf{U}_m$ is bounded below by c_3.
- (A3) $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|_F \leq c_4$.
- (A4) The differences of any eigenvalue of $\mathbf{P}_{\boldsymbol{\Gamma}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\boldsymbol{\Gamma}_m}$ against all those of $\mathbf{Q}_{\boldsymbol{\Gamma}_m}\boldsymbol{\Sigma}_m\mathbf{Q}_{\boldsymbol{\Gamma}_m}$ are greater than c_5.
- (A5) $c_6/K \leq n_k/n \leq c_7/K$, for $k = 1, \cdots, K$.

Lemma S2 (Vershynin (2018), Section 2.8, Bernstein inequality)**.** *Let $X_1, \cdots, X_n$ be independent, mean zero, sub-exponential random variables, and $\mathbf{a} = (a_1, \cdots, a_n) \in \mathbb{R}^n$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\big(|\sum_{i=1}^n a_i X_i| \geq t\big) \leq 2\exp\{-c\min\big(\frac{t^2}{K^2\|\mathbf{a}\|_2^2}, \frac{t}{K\|\mathbf{a}\|_\infty}\big)\},$$

where $K = \max_i \|X_i\|_{\psi_1}$.

Lemma S3 (Vershynin (2018), Lemma 5.2)**.** *The ϵ-covering number $\mathcal{N}(S^{n-1}, \epsilon)$ of the unit Euclidean sphere S^{n-1} equipped with the Euclidean metric satisfies for every $\epsilon > 0$ that*

$$\mathcal{N}(S^{n-1}, \epsilon) \leq (1 + 2/\epsilon)^n.$$

Lemma S4 (Vershynin (2018), Lemma 5.4)**.** *Let $\mathbf{A}$ be a symmetric $n \times n$ matrix, and let $\mathcal{N}_\epsilon$ be an ϵ-net of S^{n-1}. When $\epsilon < 1/2$,*

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x} \in S^{n-1}} |\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle| \leq (1 - 2\epsilon)^{-1} \sup_{\mathbf{x} \in \mathcal{N}_\epsilon} |\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle|.$$

Lemma S5 (Yu et al. (2015), Theorem 2)**.** *Let $\mathbf{\Sigma}$, $\widehat{\mathbf{\Sigma}}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$, respectively. Fix $1 \leq r \leq s \leq p$ and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) > 0$, where $\lambda_0 = \infty$ and $\lambda_{p+1} = -\infty$. Let $d = s - r + 1$, and let $\mathbf{V} = (\mathbf{v}_r, \cdots, \mathbf{v}_s) \in \mathbb{R}^{p \times d}$ and $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_r, \cdots, \widehat{\mathbf{v}}_s) \in \mathbb{R}^{p \times d}$ have orthogonal columns satisfying $\mathbf{\Sigma}\mathbf{v}_j = \lambda_j \mathbf{v}_j$ and $\widehat{\mathbf{\Sigma}}\widehat{\mathbf{v}}_j = \widehat{\lambda}_j \widehat{\mathbf{v}}_j$ for $j = r, \cdots, s$. Then*

$$\|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F \leq \frac{2d\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})},$$

where $\Theta(\widehat{\mathbf{V}}, \mathbf{V})$ is a $d \times d$ diagonal matrix whose j-th diagonal entry is the j-th principal angle and $\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})$ is defined entrywise.

Lemma S6. *Let $\mathbf{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ follow $\mathrm{TN}(0, \mathbf{I}_{p_1}, \cdots, \mathbf{I}_{p_M})$ and $\mathbf{A} \in \mathbb{R}^{p_{-m} \times p_{-m}}$ be a positive and symmetric definite matrix whose eigenvalues are bounded above by a constant c. Then for any fixed vector $\mathbf{x} \in \mathcal{S}^{p_m - 1} = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^{p_m}, \|\mathbf{x}\| = 1\}$and any $\epsilon \geq 0$, we have*

$$\mathbb{P}(|\frac{1}{p_{-m}}\mathbf{x}^T\mathbf{Z}_{(m)}\mathbf{A}\mathbf{Z}_{(m)}^T\mathbf{x} - \frac{1}{p_{-m}}\mathrm{tr}(\mathbf{A})| \leq \epsilon) \geq 2\exp\{-Cp_{-m}\min(\frac{\epsilon^2}{16c^2}, \frac{\epsilon}{4c})\}.$$

Proof. Let $\mathbf{P}\mathbf{\Delta}\mathbf{P}^T$ be the eigenvalue decomposition of $\mathbf{A}$, where $\mathbf{\Delta}$ is a $p_{-m} \times p_{-m}$ diagonal matrix whose i-th diagonal elements is δ_i. Let $\mathbf{W}_{(m)} = \mathbf{Z}_{(m)}\mathbf{P}$. We know that $\mathrm{vec}(\mathbf{W}_{(m)}) \sim N(0, \mathbf{I}_p)$. Then we have

$$\begin{aligned}
\frac{1}{p_{-m}}\mathbf{x}^T\mathbf{Z}_{(m)}\mathbf{A}\mathbf{Z}_{(m)}^T\mathbf{x} &= \frac{1}{p_{-m}}\mathbf{x}^T\mathbf{W}_{(m)}\mathbf{\Delta}\mathbf{W}_{(m)}^T\mathbf{x} \\
&= \frac{1}{p_{-m}}\mathbf{x}^T\sum_{l=1}^{p_{-m}}\delta_l\mathbf{W}_{(m),l}\mathbf{W}_{(m),l}^T\mathbf{x} \\
&= \frac{1}{p_{-m}}\sum_{l=1}^{p_{-m}}\delta_l\mathbf{x}^T\mathbf{W}_{(m),l}\mathbf{W}_{(m),l}^T\mathbf{x},
\end{aligned}$$

where $\mathbf{W}_{(m),l}$ is the l-th column of $\mathbf{W}_{(m)}$. Note that $\mathbf{x}^T\mathbf{W}_{(m),l} \sim N(0,1)$ and $\mathbf{x}^T\mathbf{W}_{(m),l}$ are independent for $l = 1, \cdots, p_{-m}$. Taking expectation on both sides of last equation, we have

$$\mathrm{E}(\frac{1}{p_{-m}}\mathbf{x}^T\mathbf{Z}_{(m)}\mathbf{A}\mathbf{Z}_{(m)}^T\mathbf{x}) = \mathrm{E}(\frac{1}{p_{-m}}\sum_{l=1}^{p_{-m}}\delta_l\mathbf{x}^T\mathbf{W}_{(m),l}\mathbf{W}_{(m),l}^T\mathbf{x}) = \frac{1}{p_{-m}}\sum_{l=1}^{p_{-m}}\delta_l = \frac{\mathrm{tr}(\mathbf{A})}{p_{-m}}.$$

By Bernstein's inequality (Lemma S2), we have

$$\mathbb{P}(|\frac{1}{p_{-m}}\sum_{l=1}^{p_{-m}}\delta_l\mathbf{x}^T\mathbf{W}_{(m),l}\mathbf{W}_{(m),l}^T\mathbf{x}-\frac{1}{p_{-m}}\mathrm{tr}(\mathbf{A})|\geq\epsilon)$$
$$\leq 2\exp\{-Cp_{-m}\min\big(\frac{\epsilon^2}{16\sum_{l=1}^{p_{-m}}\delta_l^2/p_{-m}},\frac{\epsilon}{4\max_l\delta_l}\big)\}.$$

By the assumption that the largest eigenvalues of $\mathbf{A}$ is bounded above by c, we have $\max_l\delta_l\leq c$. Then using the fact $K=\max_l\|\mathbf{x}^T\mathbf{W}_{(m),l}\|_{\psi_1}\leq 4$, we get the desired conclusion. □

Lemma S7. *Let $\mathbf{Z}_i\sim\mathrm{TN}(0,\mathbf{I}_{p_1},\cdots,\mathbf{I}_{p_M})$, for $i=1,\cdots,n$, independently, and*

$$\mathbf{L}=\mathbf{\Sigma}_m^{1/2}\{\frac{p}{np_{-m}\mathrm{tr}(\mathbf{\Sigma})}\sum_{i=1}^n(\mathbf{Z}_i)_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})(\mathbf{Z}_i)_{(m)}^T\}\mathbf{\Sigma}_m^{1/2}.$$

There exists a constant ϵ_0, such that when $\epsilon\leq\epsilon_0$, we have

$$\mathbb{P}(\|\mathbf{L}-\frac{p_m}{\mathrm{tr}(\mathbf{\Sigma}_m)}\mathbf{\Sigma}_m\|_2\geq\epsilon)\leq\exp(C_1p_m-C_2np_{-m}\epsilon^2),$$

where $p=\prod_{m=1}^M p_m$ and $p_{-m}=\prod_{j\neq m}p_j$, for some constant C_1 and C_2.

Proof. By Assumption (A1), we know that the largest eigenvalue of $\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'}$ is bounded above by some constant C. By the same techniques used in Lemma S6, we have

$$\begin{aligned}&\mathbb{P}(|\mathbf{x}\frac{p}{np_{-m}\mathrm{tr}(\mathbf{\Sigma})}\sum_{i=1}^n(\mathbf{Z}_i)_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})(\mathbf{Z}_i)_{(m)}^T\mathbf{x}^T-\frac{p_m}{\mathrm{tr}(\mathbf{\Sigma}_m)}|\geq\epsilon)\\&\leq 2\exp\{-Cnp_{-m}\frac{\epsilon^2(\mathrm{tr}(\mathbf{\Sigma})/p)^2}{16}\}.\end{aligned}\tag{S20}$$

for any fixed vector $\mathbf{x}\in\mathcal{S}^{p_m-1}=\{\mathbf{x}:\mathbf{x}\in\mathbb{R}^{p_m},\|\mathbf{x}\|=1\}$ when ϵ is smaller than some constant ϵ_0.

Let $\mathcal{N}_{1/4}$ be a $1/4$-net of S^{p_m-1}. By Lemma S4, we have

$$\|\frac{p}{np_{-m}\mathrm{tr}(\mathbf{\Sigma})}\sum_{i=1}^n(\mathbf{Z}_i)_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})(\mathbf{Z}_i)_{(m)}^T-\frac{p_m}{\mathrm{tr}(\mathbf{\Sigma}_m)}\|_2$$
$$\leq 2\max_{\mathbf{x}\in\mathcal{N}_{1/4}}|\langle(\frac{p}{np_{-m}\mathrm{tr}(\mathbf{\Sigma})}\sum_{i=1}^n(\mathbf{Z}_i)_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})(\mathbf{Z}_i)_{(m)}^T-\frac{p_m}{\mathrm{tr}(\mathbf{\Sigma}_m)})\mathbf{x},\mathbf{x}\rangle|$$
$$=2\max_{\mathbf{x}\in\mathcal{N}_{1/4}}|\mathbf{x}\frac{p}{np_{-m}\mathrm{tr}(\mathbf{\Sigma})}\sum_{i=1}^n(\mathbf{Z}_i)_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})(\mathbf{Z}_i)_{(m)}^T\mathbf{x}^T-\frac{p_m}{\mathrm{tr}(\mathbf{\Sigma}_m)}|.$$

Then by Lemma S3, we know that the cardinal of $\mathcal{N}_{1/4}$ is smaller or equal to 9^{p_m}. The union bound and (S20) tell us

$$\mathbb{P}(\max_{\mathbf{x}\in\mathcal{N}_{1/4}}|\mathbf{x}\frac{p}{np_{-m}\mathrm{tr}(\mathbf{\Sigma})}\sum_{i=1}^n(\mathbf{Z}_i)_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})(\mathbf{Z}_i)_{(m)}^T\mathbf{x}^T-\frac{p_m}{\mathrm{tr}(\mathbf{\Sigma}_m)}|\geq\epsilon)$$
$$\leq 9^{p_m}\cdot 2\exp\{-Cnp_{-m}\frac{\epsilon^2(\mathrm{tr}(\mathbf{\Sigma})/p)^2}{16}\}.$$

It follows that

$$\mathbb{P}(\|\frac{p}{np_{-m}\mathrm{tr}(\mathbf{\Sigma})}\sum_{i=1}^{n}(\mathbf{Z}_i)_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})(\mathbf{Z}_i)_{(m)}^T-\frac{p_m}{\mathrm{tr}(\mathbf{\Sigma}_m)}\|_2\geq\epsilon)\leq 9^{p_m}\cdot 2\exp\{-Cnp_{-m}\frac{\epsilon^2(\mathrm{tr}(\mathbf{\Sigma})/p)^2}{64}\},$$

which implies

$$\mathbb{P}(\|\mathbf{L}-\frac{p_m}{\mathrm{tr}(\mathbf{\Sigma}_m)}\mathbf{\Sigma}_m\|_2\geq\|\mathbf{\Sigma}_m\|_2\epsilon)\leq 9^{p_m}\cdot 2\exp\{-Cnp_{-m}\frac{\epsilon^2(\mathrm{tr}(\mathbf{\Sigma})/p)^2}{64}\}.$$

Since $\|\mathbf{\Sigma}_m\|_2$ is assumed to be bounded by c_1 and c_2, we have

$$\mathbb{P}(\|\mathbf{L}-\frac{p_m}{\mathrm{tr}(\mathbf{\Sigma}_m)}\mathbf{\Sigma}_m\|_2\geq\epsilon)\leq\exp(C_1p_m-C_2np_{-m}\epsilon^2),$$

for some constant C_1 and C_2. □

Lemma S8. *Let* $\mathbf{Z}\sim\mathrm{TN}(0,\mathbf{I}_{p_1},\cdots,\mathbf{I}_{p_M})$ *and*

$$\mathbf{L}=\mathbf{\Sigma}_m^{1/2}\{\frac{1}{p_{-m}}\mathbf{Z}_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})\mathbf{Z}_{(m)}^T\}\mathbf{\Sigma}_m^{1/2}.$$

There exists a constant ϵ_0*, such that when* $\epsilon\leq\epsilon_0$*, we have*

$$\mathbb{P}(\|\mathbf{L}-\frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2\geq\epsilon)\leq\exp(C_1p_m-C_2p_{-m}\epsilon^2),$$

and when $\epsilon>\epsilon_0$*,*

$$\mathbb{P}(\|\mathbf{L}-\frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2\geq\epsilon)\leq\exp(C_1p_m-C_2p_{-m}\epsilon),$$

for some constant C_1 *and* C_2*.*

The proof for this lemma is similar to that for Lemma S7. We omit it here.

Lemma S9. *If* $p_m/(np_{-m})=o(1)$*, we have* $\|\widetilde{\mathbf{\Sigma}}_m-\frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2=O(\sqrt{\frac{p_m}{np_{-m}}}+\frac{1}{n})$ *with probability at least* $1-2\exp\{-C_1p_m(C_3-1)\}$ *for some* $C_3>1$*.*

Proof. Recall that

$$\widetilde{\mathbf{\Sigma}}_m=\frac{1}{np_{-m}}\sum_{k=1}^{K}\sum_{Y^i=k}(\mathbf{X}^i-\overline{\mathbf{X}}_k)_{(m)}(\mathbf{X}^i-\overline{\mathbf{X}}_k)_{(m)}^T.$$

We decompose $\widetilde{\mathbf{\Sigma}}_m$ in the following ways,

$$\begin{aligned}\widetilde{\mathbf{\Sigma}}_m&=\frac{1}{np_{-m}}\sum_{k=1}^{K}\sum_{Y^i=k}(\mathbf{X}^i-\boldsymbol{\mu}_k)_{(m)}(\mathbf{X}^i-\boldsymbol{\mu}_k)_{(m)}^T+\frac{1}{np_{-m}}\sum_{k=1}^{K}\sum_{Y^i=k}(\mathbf{X}^i-\boldsymbol{\mu}_k)_{(m)}(\boldsymbol{\mu}_k-\overline{\mathbf{X}}_k)_{(m)}^T\\&+\frac{1}{np_{-m}}\sum_{k=1}^{K}\sum_{Y^i=k}(\boldsymbol{\mu}_k-\overline{\mathbf{X}}_k)_{(m)}(\mathbf{X}^i-\boldsymbol{\mu}_k)_{(m)}^T+\frac{1}{np_{-m}}\sum_{k=1}^{K}\sum_{Y^i=k}(\boldsymbol{\mu}_k-\overline{\mathbf{X}}_k)_{(m)}(\boldsymbol{\mu}_k-\overline{\mathbf{X}}_k)_{(m)}^T\\&=\mathbf{L}_1+\mathbf{L}_2+\mathbf{L}_3+\mathbf{L}_4.\end{aligned}$$

Note that when $Y^i = k$, $\mathrm{vec}(\mathbf{X}^i - \overline{\mathbf{X}}_k) = \mathbf{\Sigma}^{1/2}\mathbf{Z}_i$, where $\mathbf{\Sigma} = \otimes_{m=M}^{1}\mathbf{\Sigma}_m$, and $\mathbf{Z}_i \sim \mathrm{TN}(0, \mathbf{I}_{p_1}, \cdots, \mathbf{I}_{p_M})$. We have

$$\mathbf{L}_1 = \mathbf{\Sigma}_m^{1/2}\{\frac{1}{np_{-m}}\sum_{i=1}^{n}(\mathbf{Z}_i)_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})(\mathbf{Z}_i)_{(m)}^T\}\mathbf{\Sigma}_m^{1/2}.$$

By Lemma S7, we know that

$$\mathbb{P}(\|\mathbf{L}_1 - \frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2 \geq \epsilon) \leq \exp(C_1 p_m - C_2 n p_{-m}\epsilon^2),$$

for some sufficient small ϵ. Let $\epsilon^2 = C_1 p_m C_3/(np_{-m}C_2)$, where $C_3 > 1$, we have

$$\mathbb{P}(\|\mathbf{L}_1 - \frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2 \geq \epsilon) \leq \exp(-C_1 p_m(C_3 - 1)).$$

Then we consider $\mathbf{L}_2$, note that

$$\mathbf{L}_2 = \frac{1}{np_{-m}}\sum_{k=1}^{K}\big[\{\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}\}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}^T\big].$$

We have

$$\begin{aligned}\|\mathbf{L}_2\|_2 &\leq \sum_{k=1}^{K}\{\frac{1}{n_k}\|\frac{1}{\sqrt{n_k p_{-m}}}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}\|_2\|\frac{\sqrt{n_k}}{\sqrt{p_{-m}}}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}\|_2\} \\ &= \sum_{k=1}^{K}\frac{1}{n_k}\sqrt{\|\frac{1}{n_k p_{-m}}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}^T\|_2\|\frac{n_k}{p_{-m}}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}^T\|_2}.\end{aligned} \tag{S21}$$

Since $\sum_{Y^i=k}\mathrm{vec}(\mathbf{X}^i - \boldsymbol{\mu}_k)/\sqrt{n_k} \sim N(0, \mathbf{\Sigma})$, we have

$$\|\frac{1}{n_k p_{-m}}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}^T\|_2 = \|\frac{1}{p_{-m}}\mathbf{\Sigma}_m^{1/2}\mathbf{Z}_{(m)}(\bigotimes_{m'\neq m}\mathbf{\Sigma}_{m'})\mathbf{Z}_{(m)}^T\mathbf{\Sigma}_m^{1/2}\|_2,$$

for $\mathbf{Z} \sim N(0, \mathbf{I}_{p_1}, \cdots, \mathbf{I}_{p_M})$. By Lemma S8, and let $\epsilon = C_1 p_m C_3/(p_{-m}C_2)$ (We assume that $C_1 p_m C_3/(p_{-m}C_2) \geq \epsilon_0$ here. If not, we can let $\epsilon^2 = C_1 p_m C_3/(p_{-m}C_2)$ and will reach the same results.), where $C_3 > 1$, we have

$$\mathbb{P}(\|\frac{1}{n_k p_{-m}}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}^T - \frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2 \geq \epsilon) \leq \exp\{C_1 p_m(C_3 - 1)\}.$$

Hence

$$\|\frac{1}{n_k p_{-m}}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\mu}_k)_{(m)}^T\|_2 \leq \|\frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2 + \epsilon \leq C_4 + \epsilon,$$

with probability at least $1 - \exp\{C_1 p_m(C_3 - 1)\}$.

Then note that $\mathrm{vec}(\overline{\mathbf{X}} - \boldsymbol{\mu}_k) \sim N(0, \boldsymbol{\Sigma}/n_k)$, by Lemma S8 and let $\epsilon = C_1 p_m C_3/(p_{-m} C_2)$, where $C_3 > 1$, we have

$$\mathbb{P}(\|\frac{n_k}{p_{-m}}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}^T - \frac{\prod_{m'\neq m}\mathrm{tr}(\boldsymbol{\Sigma}_{m'})}{p_{-m}}\boldsymbol{\Sigma}_m\|_2 \geq \epsilon) \leq \exp\{C_1 p_m (C_3 - 1)\}. \quad \text{(S22)}$$

Hence, we have

$$\|\frac{n_k}{p_{-m}}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}^T\|_2 \leq C_5 + \epsilon.$$

with probability at least $1 - \exp\{C_1 p_m (C_3 - 1)\}$. By (S21) and the assumption $c_6 K \leq n_k/n \leq c_7 K$, we have

$$\|\mathbf{L}_2\|_2 = O(\sqrt{\frac{p_m}{np_{-m}}} + \frac{1}{n}),$$

with probability at least $1 - 2\exp\{C_1 p_m (C_3 - 1)\}$. Since $\mathbf{L}_3 = \mathbf{L}_2^T$, we have the same conclusion for $\mathbf{L}_3$.

Finally, note that

$$\|\mathbf{L}_4\|_2 \leq \sum_{k=1}^K \frac{1}{n_k}\|\frac{n_k}{p_{-m}}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}(\boldsymbol{\mu}_k - \overline{\mathbf{X}}_k)_{(m)}^T\|_2.$$

By (S22), we have $\|\mathbf{L}_4\|_2 = O(\sqrt{\frac{p_m}{np_{-m}}} + \frac{1}{n})$ with probability at least $1 - 2\exp\{C_1 p_m (C_3 - 1)\}$.

To sum up, $\|\widetilde{\boldsymbol{\Sigma}}_m - \frac{\prod_{m'\neq m}\mathrm{tr}(\boldsymbol{\Sigma}_{m'})}{p_{-m}}\boldsymbol{\Sigma}_m\|_2 = O(\sqrt{\frac{p_m}{np_{-m}}} + \frac{1}{n})$ with probability at least $1 - 2\exp\{-C_1 p_m (C_3 - 1)\}$ for some $C_3 > 1$.

□

Note that $\widetilde{\boldsymbol{\Sigma}}_m$ is consistency for $\frac{\prod_{m'\neq m}\mathrm{tr}(\boldsymbol{\Sigma}_{m'})}{p_{-m}}$ times $\boldsymbol{\Sigma}_m$. The constant $\frac{\prod_{m'\neq m}\mathrm{tr}(\boldsymbol{\Sigma}_{m'})}{p_{-m}}$ will not affect the estimation of the envelope basis because we only need the eigenvalues of it. But the we need a unbiased estimation in the parameter $\mathbf{B}_k$. Note that $\mathrm{E}(\bigotimes_{m=M}^1 \widetilde{\boldsymbol{\Sigma}}_m) = (\mathrm{tr}(\boldsymbol{\Sigma})/p)^{M-1}\boldsymbol{\Sigma}$. Thus, we need to eliminate the effect of $(\mathrm{tr}(\boldsymbol{\Sigma})/p)^{M-1}$ in the estimation of $\boldsymbol{\Sigma}$. Let $\alpha = \mathrm{tr}(\boldsymbol{\Sigma})/p$, by Lemma S9, we have $|\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_m)/p_m - \alpha| = O(\sqrt{\frac{p_m}{np_{-m}}} + \frac{1}{n})$. Then we let $\widetilde{\alpha} = \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_m)/p_m$ and $\widetilde{\boldsymbol{\Sigma}} = 1/\widetilde{\alpha}^{M-1}\bigotimes_{m=M}^1 \widetilde{\boldsymbol{\Sigma}}_m$. The following lemma shows that $\widetilde{\boldsymbol{\Sigma}}$ is consistent for $\boldsymbol{\Sigma}$.

Lemma S10. *For a constant $C_3 > 1$, $\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 = O(\max_m \sqrt{\frac{p_m}{np_{-m}}} + \frac{1}{n})$ with probability at least $1 - KM\exp\{-C_1 p_m (C_3 - 1)\}$.*

Proof. Firstly, note that

$$\begin{aligned}
\|\bigotimes_{m=M}^1 \widetilde{\boldsymbol{\Sigma}}_m - \alpha^{M-1}\boldsymbol{\Sigma}\|_2 \leq{}& \|(\otimes_{m=M}^2 \widetilde{\boldsymbol{\Sigma}}_m) \otimes (\widetilde{\boldsymbol{\Sigma}}_1 - \frac{\prod_{m'\neq 1}\mathrm{tr}(\boldsymbol{\Sigma}_{m'})}{p_{-1}}\boldsymbol{\Sigma}_1)\|_2 \\
&+ \sum_{j=2}^{M-1}\|(\otimes_{m=M}^{j+1}\widetilde{\boldsymbol{\Sigma}}_j \otimes (\widetilde{\boldsymbol{\Sigma}}_j - \frac{\prod_{m'\neq j}\mathrm{tr}(\boldsymbol{\Sigma}_{m'})}{p_{-j}}\boldsymbol{\Sigma}_j)(\otimes_{m=j-1}^1 \frac{\prod_{m'\neq m}\mathrm{tr}(\boldsymbol{\Sigma}_{m'})}{p_{-m}}\boldsymbol{\Sigma}_m)\|_2 \\
&+ \|(\widetilde{\boldsymbol{\Sigma}}_M - \frac{\prod_{m'\neq M}\mathrm{tr}(\boldsymbol{\Sigma}_{m'})}{p_{-M}}\boldsymbol{\Sigma}_M) \otimes (\otimes_{m=M-1}^1 \frac{\prod_{m'\neq m}\mathrm{tr}(\boldsymbol{\Sigma}_{m'})}{p_{-m}}\boldsymbol{\Sigma}_m)\|_2,
\end{aligned}$$

and $\|\widetilde{\mathbf{\Sigma}}_m\|_2 \le \|\widetilde{\mathbf{\Sigma}}_m - \frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2 + \frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\|\mathbf{\Sigma}_m\|_2$. By Lemma S9 and Assumption (A1), we have

$$\|\bigotimes_{m=M}^{1}\widetilde{\mathbf{\Sigma}}_m - \alpha^{M-1}\mathbf{\Sigma}\|_2 = O(\max_m \sqrt{p_m/(np_{-m})} + 1/n).$$

with probability at least $1 - KM\exp\{-C_1 p_m(C_3 - 1)\}$. Then note that

$$\begin{aligned}
\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 &= \|1/\widetilde{\alpha}^{M-1}\bigotimes_{m=M}^{1}\widetilde{\mathbf{\Sigma}}_m - \mathbf{\Sigma}\|_2 \\
&= \|1/\widetilde{\alpha}^{M-1}\bigotimes_{m=M}^{1}\widetilde{\mathbf{\Sigma}}_m - \alpha^{M-1}/\widetilde{\alpha}^{M-1}\mathbf{\Sigma} + (\alpha^{M-1}/\widetilde{\alpha}^{M-1} - 1)\mathbf{\Sigma}\|_2 \\
&\le 1/\widetilde{\alpha}^{M-1}\|\bigotimes_{m=M}^{1}\widetilde{\mathbf{\Sigma}}_m - \alpha^{M-1}\mathbf{\Sigma}\|_2 + |(\alpha^{M-1}/\widetilde{\alpha}^{M-1} - 1)|\|\mathbf{\Sigma}\|_2,
\end{aligned}$$

we have $\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = O(\max_m \sqrt{\frac{p_m}{np_{-m}}} + \frac{1}{n})$ with probability at least $1 - KM\exp\{-C_1 p_m(C_3 - 1)\}$. $\square$

Lemma S11. *Suppose that* $p_m/(np_{-m}) = o(1)$. *For a constant* $C_3 > 1$, $\|\widehat{\mathbf{U}}_m - \mathbf{U}_m\|_2 = O(\sqrt{\frac{p_m}{np_{-m}}} + \sqrt{\frac{1}{n}})$ *with probability at least* $1 - K\exp\{-C_1 p_m(C_3 - 1)\}$.

Proof. Recall that

$$\widehat{\mathbf{U}}_m = \frac{1}{p_{-m}}\sum_{k=1}^{K}\frac{n_k}{n}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)}^T.$$

Note that $\mathrm{vec}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}) \sim N(\boldsymbol{\mu}_k - \boldsymbol{\mu}, (1/n_k + 1/n)\mathbf{\Sigma})$. By Lemma S8, letting $\epsilon = C_1 p_m C_3/(p_{-m}C_2)$, we have

$$\mathbb{P}(\|\frac{n}{p_{-m}}\{(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)} - (\boldsymbol{\mu}_k - \boldsymbol{\mu})_{(m)}\}\{(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)} - (\boldsymbol{\mu}_k - \boldsymbol{\mu})_{(m)}\}^T - \frac{c_1\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2 \ge \epsilon)$$
$$\le \exp\{-C_1 p_m(C_3 - 1)\}.$$

It follows that

$$\|\frac{n}{p_{-m}}\{(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)} - (\boldsymbol{\mu}_k - \boldsymbol{\mu})_{(m)}\}\{(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)} - (\boldsymbol{\mu}_k - \boldsymbol{\mu})_{(m)}\}^T\|_2 \le C_4 + \epsilon.$$

Thus, $\|\frac{1}{\sqrt{p_{-m}}}\{(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)} - (\boldsymbol{\mu}_k - \boldsymbol{\mu})_{(m)}\}\|_2 = O(\sqrt{\frac{p_m}{np_{-m}}} + \sqrt{\frac{1}{n}})$ with probability at least $1 - \exp\{-C_1 p_m(C_3 - 1)\}$.

Then we have

$$\begin{aligned}
&\|\frac{1}{p_{-m}}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)}^T - \frac{1}{p_{-m}}(\boldsymbol{\mu}_k - \boldsymbol{\mu})_{(m)}(\boldsymbol{\mu}_k - \boldsymbol{\mu})_{(m)}^T\|_2 \\
&\le \|\frac{1}{p_{-m}}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}} - \boldsymbol{\mu}_k + \boldsymbol{\mu})_{(m)}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}} - \boldsymbol{\mu}_k + \boldsymbol{\mu})_{(m)}^T\|_2 + 2\|\frac{1}{p_{-m}}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}} - \boldsymbol{\mu}_k + \boldsymbol{\mu})_{(m)}(\boldsymbol{\mu}_k - \boldsymbol{\mu})_{(m)}^T\|_2 \\
&= O(\frac{p_m}{np_{-m}} + \frac{1}{n} + \sqrt{\frac{p_m}{np_{-m}}} + \sqrt{\frac{1}{n}}) = O(\sqrt{\frac{p_m}{np_{-m}}} + \sqrt{\frac{1}{n}}),
\end{aligned}$$

with probability at least $1 - \exp\{-C_1 p_m (C_3 - 1)\}$.

Note that the last inequality is equivalent to

$$\|\widehat{\mathbf{U}}_m - \mathbf{U}_m\|_2 = O(\sqrt{\frac{p_m}{np_{-m}}} + \sqrt{\frac{1}{n}}),$$

with probability at least $1 - K\exp\{-C_1 p_m (C_3 - 1)\}$. □

Next we will show two lemmas about the consistency of the estimated envelope score. We first review the definitions of the envelope scores and introduce some additional notations. Let $\mathbf{V}^m = (\mathbf{v}_1^m, \cdots, \mathbf{v}_{p_m}^m)$ be the eigenvalues of $\mathbf{\Sigma}_m$ with the corresponding eigenvalues satisfies that $\lambda_1^m \geq \cdots \geq \lambda_{p_m}^m$, and $\widehat{\mathbf{V}}^m = (\widehat{\mathbf{v}}_1^m, \cdots, \widehat{\mathbf{v}}_{p_m}^m)$ be the eigenvalues of $\widetilde{\mathbf{\Sigma}}_m$ with the corresponding eigenvalues satisfies that $\widehat{\lambda}_1^m \geq \cdots \geq \widehat{\lambda}_{p_m}^m$. Define $\mathcal{A}_m = \{l \mid \phi_l^m \neq 0\}$, then we have $\mathcal{S}_m = \mathrm{span}(\{\mathbf{v}_l \mid l \in \mathcal{A}_m\})$. We organize the envelope scores in the descending order $\phi_{(1)}^m \geq \phi_{(2)}^m \geq \cdots \geq \phi_{(p_m)}^m$. Since the dimension of the envelope subspace is d_m, we know that when $j \geq d_m + 1$, $\phi_{(j)}^m = 0$. Define $\mathbf{P}_{\mathcal{A}_m}$ be the projection matrix onto the subspace spanned by $\{\mathbf{v}_l \mid l \in \mathcal{A}_m\}$, and $\mathbf{P}_{\widehat{\mathcal{A}}_m}$ be the projection matrix onto the subspace spanned by $\{\widehat{\mathbf{v}}_l \mid l \in \mathcal{A}_m\}$. By Assumption (A2), the smallest non-zero eigenvalue of $\mathbf{U}_m$ is bounded below by c_3. Thus $\phi_{(d_m)}^m \geq c_3$. Also, by Assumption (A3), we have $\|\mathbf{U}_m\|_2 \leq c_4$. We define the estimated envelope score as $\widehat{\phi}_l^m = (\widehat{\mathbf{v}}_l^m)^T \widehat{\mathbf{U}}_{(m)} \widehat{\mathbf{v}}_l^m$.

We first consider the case where the eigenvalues λ_l^m, $l \in \mathcal{A}_m$ are all distinct and are different from λ_j^m, $j \in \mathcal{A}_m^c$. The distinction between eigenvalues λ_j^m, $j \in \mathcal{A}_m^c$ are not required. Define $\Delta = \min_{l \in \mathcal{A}_m, j \neq l}\{|\lambda_l^m - \lambda_j^m|\}$. By Assumption (A5), $\Delta \geq c_5$.

Lemma S12. *The estimated envelope score satisfies that* $|\widehat{\phi}_l^m - \phi_l^m| = O(\sqrt{p_m/(np_{-m})} + \sqrt{1/n})$ *with probability at least* $1 - C_2 K \exp\{-C_1 p_m (C_3 - 1)\}$.

Proof. By Lemma S5, we know that when $l \in \mathcal{A}_m$,

$$\sin\Theta(\mathbf{v}_l^m, \widehat{\mathbf{v}}_l^m) \leq \frac{2\|\widetilde{\mathbf{\Sigma}}_m - \frac{\prod_{m' \neq m} \mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}} \mathbf{\Sigma}_m\|_2}{\Delta}.$$

It follows that

$$\|\widehat{\mathbf{v}}_l^m - \mathbf{v}_l^m\|_2 \leq \frac{2\sqrt{2}\|\widetilde{\mathbf{\Sigma}}_m - \frac{\prod_{m' \neq m} \mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}} \mathbf{\Sigma}_m\|_2}{\Delta}. \tag{S23}$$

Then note that

$$\begin{aligned}
|\widehat{\phi}_l^m - \phi_l^m| &\leq |(\widehat{\mathbf{v}}_l^m)^T \widehat{\mathbf{U}}_{(m)} \widehat{\mathbf{v}}_l^m - (\mathbf{v}_l^m)^T \mathbf{U}_m \mathbf{v}_l^m| \\
&\leq |(\widehat{\mathbf{v}}_l^m)^T (\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_m) \widehat{\mathbf{v}}_l^m| + |(\widehat{\mathbf{v}}_l^m - \mathbf{v}_l^m)^T \mathbf{U}_m (\widehat{\mathbf{v}}_l^m - \mathbf{v}_l^m)| + 2|(\mathbf{v}_l^m)^T \mathbf{U}_m (\widehat{\mathbf{v}}_l^m - \mathbf{v}_l^m)| \\
&\leq \|(\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_m)\|_2 + 3\|\mathbf{U}_m\|_2 \|(\widehat{\mathbf{v}}_l^m - \mathbf{v}_l^m)\|_2.
\end{aligned}$$

Combined with Lemmas S9 and S11, we know that when $l \in \mathcal{A}$, $|\widehat{\phi}_l^m - \phi_l^m| = O(\sqrt{p_m/(np_{-m})} + \sqrt{1/n})$ with probability at least $1 - K\exp\{-C_1 p_m (C_3 - 1)\}$.

Then we prove that when $j \in \mathcal{A}^c$, $\widehat{\phi}_j^m = O(u_m p_m/(np_{-m}) + \sqrt{1/n})$. A direct results follows from (S23) is

$$\|\mathbf{P}_{\widehat{\mathcal{A}}_m} - \mathbf{P}_{\mathcal{A}_m}\|_F \le \frac{2\sqrt{2u_m}\|\widetilde{\mathbf{\Sigma}}_m - \frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2}{\Delta}.$$

Note that $\mathbf{P}_{\widehat{\mathcal{A}}_m}\widehat{\mathbf{v}}_j^m = 0$, we have

$$\|\mathbf{P}_{\mathcal{A}_m}\widehat{\mathbf{v}}_j^m\|_2 \le \|(\mathbf{P}_{\mathcal{A}_m} - \mathbf{P}_{\widehat{\mathcal{A}}_m})\widehat{\mathbf{v}}_j^m\|_2 + \|\mathbf{P}_{\widehat{\mathcal{A}}_m}\widehat{\mathbf{v}}_j^m\|_2 \le \frac{2\sqrt{2u_m}\|\widetilde{\mathbf{\Sigma}}_m - \frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2}{\Delta}.$$

Then note that

$$\begin{aligned}
(\widehat{\mathbf{v}}_j^m)^T\widehat{\mathbf{U}}_m\widehat{\mathbf{v}}_j^m &\le (\widehat{\mathbf{v}}_j^m)^T\mathbf{U}_m\widehat{\mathbf{v}}_j^m + \|\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_m\|_2\\
&= (\widehat{\mathbf{v}}_j^m)^T(\mathbf{P}_{\mathcal{A}_m} + \mathbf{Q}_{\mathcal{A}_m})\mathbf{U}_m(\mathbf{P}_{\mathcal{A}_m} + \mathbf{Q}_{\mathcal{A}_m})\widehat{\mathbf{v}}_j^m + \|\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_{(m)}\|_2\\
&= (\widehat{\mathbf{v}}_j^m)^T\mathbf{P}_{\mathcal{A}_m}\mathbf{U}_m\mathbf{P}_{\mathcal{A}_m}\widehat{\mathbf{v}}_j^m + \|\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_m\|_2\\
&\le \|\mathbf{P}_{\mathcal{A}_m}\widehat{\mathbf{v}}_j^m\|_2^2\|\mathbf{U}_m\|_2 + \|\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_m\|_2.
\end{aligned}$$

By Lemmas S9 and S11, we know that

$$(\widehat{\mathbf{v}}_j^m)^T\widehat{\mathbf{U}}_{(m)}\widehat{\mathbf{v}}_j^m = O(u_m p_m/(np_{-m}) + \sqrt{1/n})$$

with probability at least $1 - K\exp\{-C_1 p_m(C_3-1)\}$, for $j \in \mathcal{A}_m^c$.

To sum up, we have $|\widehat{\phi}_l^m - \phi_l^m| = O(\sqrt{p_m/(np_{-m})} + \sqrt{1/n})$ with probability at least $1 - C_2K\exp\{-C_1p_m(C_3-1)\}$. □

Next, we first consider the case where the eigenvalues λ_l^m, $l \in \mathcal{A}_m$ are not all distinct but are different from λ_j^m, $j \in \mathcal{A}_m^c$. Define $\Delta = \min_{l\in\mathcal{A}_m, j\in\mathcal{A}_m^c}\{|\lambda_l^m - \lambda_j^m|\}$. Again, by Assumption (A5), we have $\Delta \ge c_5$.

Lemma S13. *When $l \in \mathcal{A}_m$, $\widehat{\phi}_l^m$ are bounded below by some constant C, and when $l \in \mathcal{A}_m^c$, $\widehat{\phi}_l^m = O(\sqrt{u_m p_m/(np_{-m})} + \sqrt{1/n})$ with probability at least $1 - C_2K\exp\{-C_1p_m(C_3-1)\}$.*

Proof. By Lemma S5, we know that when $l \in \mathcal{A}_m$,

$$\|\sin\Theta(\mathbf{v}_{\widehat{\mathcal{A}}_m}^m, \mathbf{v}_{\mathcal{A}_m}^m)\|_F \le \frac{2\sqrt{u_m}\|\widetilde{\mathbf{\Sigma}}_m - \frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2}{\Delta}.$$

It follows that

$$\|\mathbf{P}_{\widehat{\mathcal{A}}_m} - \mathbf{P}_{\mathcal{A}_m}\|_2 \le \frac{2\sqrt{2u_m}\|\widetilde{\mathbf{\Sigma}}_m - \frac{\prod_{m'\neq m}\mathrm{tr}(\mathbf{\Sigma}_{m'})}{p_{-m}}\mathbf{\Sigma}_m\|_2}{\Delta}. \tag{S24}$$

For any $l \in \mathcal{A}_m$, we have

$$\begin{aligned}
(\widehat{\mathbf{v}}_l^m)^T\widehat{\mathbf{U}}_{(m)}\widehat{\mathbf{v}}_l^m &\ge (\widehat{\mathbf{v}}_l^m)^T\widehat{\mathbf{U}}_{(m)}\widehat{\mathbf{v}}_l^m - \|\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_m\|_2\\
&= (\widehat{\mathbf{v}}_l^m)^T(\mathbf{P}_{\mathcal{A}_m} + \mathbf{Q}_{\mathcal{A}_m})\widehat{\mathbf{U}}_{(m)}(\mathbf{P}_{\mathcal{A}_m} + \mathbf{Q}_{\mathcal{A}_m})\widehat{\mathbf{v}}_l^m - \|\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_m\|_2\\
&= (\widehat{\mathbf{v}}_l^m)^T\mathbf{P}_{\mathcal{A}_m}\widehat{\mathbf{U}}_{(m)}\mathbf{P}_{\mathcal{A}_m}\widehat{\mathbf{v}}_l^m - \|\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_m\|_2\\
&\ge \|\widehat{\mathbf{v}}_l^m\mathbf{P}_{\mathcal{A}_m}\|_2^2\lambda_{min}(\mathbf{U}_m) - \|\widehat{\mathbf{U}}_{(m)} - \mathbf{U}_m\|_2.
\end{aligned}$$

Also, note that

$$\|\widehat{\mathbf{v}}_l^m \mathbf{P}_{\mathcal{A}_m}\|_2 \geq \|\widehat{\mathbf{v}}_l^m \mathbf{P}_{\widehat{\mathcal{A}}_m}\|_2 - \|\mathbf{P}_{\widehat{\mathcal{A}}_m} - \mathbf{P}_{\mathcal{A}_m}\|_2 = 1 - \|\mathbf{P}_{\widehat{\mathcal{A}}_m} - \mathbf{P}_{\mathcal{A}_m}\|_2.$$

Combined with Lemmas S9 and S11, we know that $\widehat{\phi}_l^m = (\widehat{\mathbf{v}}_l^m)^T \widehat{\mathbf{U}}_{(m)} \widehat{\mathbf{v}}_l^m \geq C$ with probability at least $1 - K\exp\{-C_1 p_m(C_3 - 1)\}$.

When $l \in \mathcal{A}_m^c$, the proof for $\widehat{\phi}_l^m = O(u_m p_m/(np_{-m}) + \sqrt{1/n})$ with probability at least $1 - K\exp\{-C_1 p_m(C_3 - 1)\}$ is analogus to that in Lemma S12. We omit the details. Thus, when c_3 in Assumption (A2) satisfies that $c_3 \gg u_m p_m/(np_{-m}) + \sqrt{1/n}$, we can correctly select the envelope dimensions. $\square$

Lemma S14. *For some constant $C_3 > 1$, $|\widehat{c}_k - c_k| = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})$, with probability at least $1 - C_4 KM\exp\{-C_1 p_m(C_3 - 1)\}$.*

Proof. Recall that $c_k = \log(\pi_k/\pi_1) - \langle \mathbf{B}_k, \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_1}{2}\rangle = \log(\pi_k/\pi_1) - \mathrm{vec}(\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}\mathrm{vec}(\boldsymbol{\mu}_k)/2 + \mathrm{vec}(\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}\mathrm{vec}(\boldsymbol{\mu}_1)/2$.

Since $I(Y^i = k)$ is sub-Gaussian, we have

$$\mathbb{P}(|\widehat{\pi}_k - \pi_k| \geq t) \leq \exp\{-Cnt^2\}.$$

Let $t = M/\sqrt{n}$, we have

$$|\widehat{\pi}_k - \pi_k| = O(\frac{1}{\sqrt{n}}),$$

with probability at least $1 - \exp(-C_M)$ for arbitrary constant C_M. Then note that

$$2|\log(\widehat{\pi}_k/\widehat{\pi}_1) - \log(\pi_k/\pi_1)| \leq |\widehat{\pi}_k - \pi_k| + |\widehat{\pi}_1 - \pi_1|.$$

We have

$$|\log(\widehat{\pi}_k/\widehat{\pi}_1) - \log(\pi_k/\pi_1)| = O(\frac{1}{\sqrt{n}}),$$

with probability at least $1 - 2\exp(-C_M)$.

Define $\Delta_k = \mathrm{vec}(\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}\mathrm{vec}(\boldsymbol{\mu}_k)$, and $\widehat{\Delta}_k = \mathrm{vec}(\overline{\mathbf{X}}_k)^T \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} \widetilde{\boldsymbol{\Sigma}}^{-1} \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} \mathrm{vec}(\overline{\mathbf{X}}_k)$. As is shown in the proof of Theorem S3, we have

$$\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathbf{P}_{\boldsymbol{\Gamma}}\|_2 = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{1/n}),$$

and

$$\|\widetilde{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2 = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{1/n}).$$

with probability at least $1 - KM\exp\{-C_1 p_m(C_3 - 1)\}$.

Note that

$$\begin{aligned}|\widehat{\Delta}_k - \Delta_k| \leq{}& \|\mathrm{vec}(\overline{\mathbf{X}}_k)^T \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} \widetilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathbf{P}_{\boldsymbol{\Gamma}}\mathrm{vec}(\boldsymbol{\mu}_k))\|_2 + \|\mathrm{vec}(\overline{\mathbf{X}}_k)^T \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}(\widetilde{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{P}_{\boldsymbol{\Gamma}}\mathrm{vec}(\boldsymbol{\mu}_k)\|_2 \\ &+ \|(\mathrm{vec}(\overline{\mathbf{X}}_k^T)\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathrm{vec}(\boldsymbol{\mu}_k^T)\mathbf{P}_{\boldsymbol{\Gamma}})\boldsymbol{\Sigma}^{-1}\mathbf{P}_{\boldsymbol{\Gamma}}\mathrm{vec}(\boldsymbol{\mu}_k)\|_2.\end{aligned}$$

For $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathbf{P}_{\boldsymbol{\Gamma}}\mathrm{vec}(\boldsymbol{\mu}_k)$, we have

$$\begin{aligned}\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathbf{P}_{\boldsymbol{\Gamma}}\mathrm{vec}(\boldsymbol{\mu}_k)\|_2 &\leq \|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\boldsymbol{\mu}_k))\|_2 + \|(\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathbf{P}_{\boldsymbol{\Gamma}})\mathrm{vec}(\boldsymbol{\mu}_k)\|_2 \\ &\leq \|\widehat{\boldsymbol{\Gamma}}^T \mathrm{vec}(\overline{\mathbf{X}}_k - \boldsymbol{\mu}_k)\|_2 + \|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathbf{P}_{\boldsymbol{\Gamma}}\|_2 \|\mathrm{vec}(\boldsymbol{\mu}_k)\|_2.\end{aligned}$$

Since $\mathbf{X}$ are normally distributed, we know that $\sum_{Y^i=k}(\mathbf{X}_i - \overline{\mathbf{X}}_k)$ are independent with $\overline{\mathbf{X}}_k$, which implies that $\widetilde{\boldsymbol{\Sigma}}_m$, $m = 1, \cdots, M$ are independent with $\overline{\boldsymbol{\mu}}_k$, $k = 1, \cdots, K$. Therefore, $\widehat{\boldsymbol{\Gamma}}$ are independent with $\overline{\mathbf{X}}_k$. Then we know that given $\widehat{\boldsymbol{\Gamma}}$, $\widehat{\boldsymbol{\Gamma}}^T\mathrm{vec}(\overline{\mathbf{X}}_k - \boldsymbol{\mu}_k) \sim N(0, \widehat{\boldsymbol{\Gamma}}^T\boldsymbol{\Sigma}\widehat{\boldsymbol{\Gamma}}/n_k)$. So $\|\widehat{\boldsymbol{\Gamma}}^T\mathrm{vec}(\overline{\mathbf{X}}_k - \boldsymbol{\mu}_k)\|_2 = O(\sqrt{u/n})$.

Then we know that

$$\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathbf{P}_{\boldsymbol{\Gamma}}\mathrm{vec}(\boldsymbol{\mu}_k)\|_2 = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})$$

with probability at least $1 - C_4KM\exp\{-C_1p_m(C_3-1)\}$. It follows that

$$|\widehat{\Delta}_k - \Delta_k| = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})$$

with probability at least $1 - C_4KM\exp\{-C_1p_m(C_3-1)\}$.

To sum up, we have $|\widehat{c}_k - c_k| = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})$, with probability at least $1 - C_4KM\exp\{-C_1p_m(C_3-1)\}$. □

S5.2 Main Proofs for Theorem 4

We divide the three conclusions in Theorem 4 into the following three theorems and prove them one by one.

Theorem S2. *Under Assumptions (A1)-(A5), for a constant $C_3 > 1$, we have*

$$\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_m} - \mathbf{P}_{\boldsymbol{\Gamma}_m}\|_F = O(\sqrt{u_m p_m/(np_{-m})} + \sqrt{1/n}),$$

with probability at least $1 - 2K\exp\{-C_1p_m(C_3-1)\}$.

Proof. In Lemma S13, showed that we can select the envelope dimensions and the envelope basics correctly with probability at least $1 - K\exp\{-C_1p_m(C_3-1)\}$. Then by (S24) and Lemma S9, we have $\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_m} - \mathbf{P}_{\boldsymbol{\Gamma}_m}\|_F = O(\sqrt{u_m p_m/(np_{-m})} + \sqrt{1/n})$ with probability at least $1 - 2K\exp\{-C_1p_m(C_3-1)\}$. □

Theorem S3. *Under Assumptions (A1)-(A5), for a constant $C_3 > 1$, we have*

$$\|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2 = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n}),$$

with probability at least $1 - C_4KM\exp\{-C_1p_m(C_3-1)\}$.

Proof. Let $\boldsymbol{\delta} = \mathrm{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$. We have

$$\mathrm{vec}(\mathbf{B}_k) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}^{-1}\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\delta}.$$

Next we will show that $\|\widetilde{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2 = O(\max_m \sqrt{p_m/(np_{-m})} + \sqrt{1/n})$, $\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathbf{P}_{\boldsymbol{\Gamma}}\| = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{1/n})$, and $\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\widehat{\boldsymbol{\delta}} - \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\delta}\| = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{1/n})$, with high probability.

Firstly, we have

$$\begin{aligned}\|\widetilde{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2 &= \|\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\|_2 \\ &\leq (\|\widetilde{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_2 + \|\boldsymbol{\Sigma}^{-1}\|_2)\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2\|\boldsymbol{\Sigma}\|_2.\end{aligned}$$

By Lemma S10, we know that when $\max_m \sqrt{p_m/(np_{-m})} + \sqrt{1/n} \to 0$, we have

$$\|\widetilde{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_2 = O(\max_m \sqrt{p_m/(np_{-m})} + \sqrt{1/n}) \tag{S25}$$

with probability at least $1 - KM\exp\{-C_1 p_m(C_3 - 1)\}$.

By the same technique in S10, we have

$$\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}}\|_2 \le \|(\otimes_{m=M}^{2}\mathbf{P}_{\widehat{\mathbf{\Gamma}}_m}) \otimes (\mathbf{P}_{\widehat{\mathbf{\Gamma}}_1} - \mathbf{P}_{\mathbf{\Gamma}_1})\|_2 + \sum_{j=2}^{M-1} \|(\otimes_{m=M}^{j+1}\mathbf{P}_{\widehat{\mathbf{\Gamma}}_m}) \otimes (\mathbf{P}_{\widehat{\mathbf{\Gamma}}_j} - \mathbf{P}_{\mathbf{\Gamma}_j})(\otimes_{m=j-1}^{1}\mathbf{P}_{\mathbf{\Gamma}_j})\|_2$$
$$+ \|(\mathbf{P}_{\widehat{\mathbf{\Gamma}}_M} - \mathbf{P}_{\mathbf{\Gamma}_M}) \otimes (\otimes_{m=M-1}^{1}\mathbf{P}_{\mathbf{\Gamma}_m})\|_2.$$

Note that $\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}_m}\|_2 \le \|\mathbf{P}_{\widehat{\mathbf{\Gamma}}_m} - \mathbf{P}_{\mathbf{\Gamma}_m}\|_2 + \|\mathbf{P}_{\mathbf{\Gamma}_m}\|_2$. By Theorem S2, we have

$$\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}}\|_2 = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{1/n}). \tag{S26}$$

with probability at least $1 - KM\exp\{-C_1 p_m(C_3 - 1)\}$.

For $\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widehat{\boldsymbol{\delta}} - \mathbf{P}_{\widehat{\mathbf{\Gamma}}}\boldsymbol{\delta}$, we have

$$\begin{aligned}\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widehat{\boldsymbol{\delta}} - \mathbf{P}_{\widehat{\mathbf{\Gamma}}}\boldsymbol{\delta}\|_2 &\le \|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|_2 + \|(\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}})\boldsymbol{\delta}\|_2 \\ &\le \|\widehat{\mathbf{\Gamma}}^T(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|_2 + \|\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}}\|_2\|\boldsymbol{\delta}\|_2.\end{aligned}$$

Since $\mathbf{X}$ are normally distributed, we know that $\sum_{Y^i=k}(\mathbf{X}_i - \overline{\mathbf{X}}_k)$ are independent with $\overline{\mathbf{X}}_k$, which implies that $\widetilde{\mathbf{\Sigma}}_m$, $m = 1, \cdots, M$, are independent with $\overline{\boldsymbol{\mu}}_k$, for $k = 1, \cdots, K$. Therefore, $\widehat{\mathbf{\Gamma}}$ are independent with $\widehat{\boldsymbol{\delta}}$. Then we know that given $\widehat{\mathbf{\Gamma}}$, $\widehat{\mathbf{\Gamma}}^T(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \sim N(0, (1/n_k + 1/n_1)\widehat{\mathbf{\Gamma}}^T\mathbf{\Sigma}\widehat{\mathbf{\Gamma}})$. It follows that $\|\widehat{\mathbf{\Gamma}}^T(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|_2 = O(\sqrt{u/n})$ with probability at least $1 - 2\exp(-c_M)$ for some constant c_M. By Assumption (A3), we have $\|\boldsymbol{\delta}\|_2 \le c$ for some constant c. Then by (S26) and $\|\boldsymbol{\delta}\|_2 \le c$ for some constant c, we know that

$$\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widehat{\boldsymbol{\delta}} - \mathbf{P}_{\mathbf{\Gamma}}\boldsymbol{\delta}\|_2 = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n}) \tag{S27}$$

with probability at least $1 - C_4KM\exp\{-C_1 p_m(C_3 - 1)\}$. Finally, note that

$$\begin{aligned}\|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2 &= \|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widehat{\boldsymbol{\delta}} - \mathbf{P}_{\widehat{\mathbf{\Gamma}}}\mathbf{\Sigma}^{-1}\mathbf{P}_{\mathbf{\Gamma}}\boldsymbol{\delta}\|_2 \\ &\le \|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widetilde{\mathbf{\Sigma}}^{-1}\|_2\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widehat{\boldsymbol{\delta}} - \mathbf{P}_{\mathbf{\Gamma}}\boldsymbol{\delta}\|_2 + \|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widetilde{\mathbf{\Sigma}}^{-1} - \mathbf{P}_{\mathbf{\Gamma}}\mathbf{\Sigma}^{-1}\|_2\|\mathbf{P}_{\mathbf{\Gamma}}\boldsymbol{\delta}\|_2 \\ &\le \|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\|_2\|\widetilde{\mathbf{\Sigma}}^{-1}\|_2\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\widehat{\boldsymbol{\delta}} - \mathbf{P}_{\mathbf{\Gamma}}\boldsymbol{\delta}\|_2 + (\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}}\|_2\|\widetilde{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_2 + \|\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}}\|_2\|\mathbf{\Sigma}^{-1}\|_2)\|\boldsymbol{\delta}\|_2.\end{aligned}$$

By the condition that $\|\boldsymbol{\delta}\|_2 \le c$, (S25), and (S27), we know that

$$\|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2 = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})$$

with probability at least $1 - C_4KM\exp\{-C_1 p_m(C_3 - 1)\}$.

□

Theorem S4. *Under Assumptions (A1)-(A5), for some constant* $C_3 > 1$*, we have*

$$|\widehat{R} - R| = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})$$

with probability at least $1 - C_4 KM \exp\{-C_1 p_m (C_3 - 1)\}$.

Proof. Define $l_k = c_k + \langle \mathbf{B}_k, \mathbf{X} \rangle$, $\widehat{l}_k = \widehat{c}_k + \langle \widehat{\mathbf{B}}_k, \mathbf{X} \rangle$, and events

$$\mathcal{B}_1 = \{|\widehat{c}_k - c_k| \le C_1^* (\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})\},$$
$$\mathcal{B}_2 = \{\|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2 \le C_2^* (\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})\}.$$

Assume that the event $\mathcal{B}_1 \cap \mathcal{B}_2$ has happened. By Theorem S3 and Lemma S14, we have

$$\mathbb{P}(\mathcal{B}_1 \cap \mathcal{B}_2) \ge 1 - C_4 KM \exp\{-C_1 p_m (C_3 - 1)\}.$$

For any $\epsilon > 0$,

$$\begin{aligned}
|\widehat{R} - R| &\le \mathbb{P}(\widehat{Y}(\widehat{\mathbf{B}}_k, \widehat{\pi}_k, k = 1, \cdots, K) \ne \widehat{Y}(\mathbf{B}_k, \pi_k, k = 1, \cdots, K)) \\
&\le 1 - \mathbb{P}(|\widehat{l}_k - l_k| < \epsilon/2, |l_k - l_{k'}| > \epsilon/2, \text{for any } k, k') \\
&\le \mathbb{P}(|\widehat{l}_k - l_k| \ge \epsilon/2 \text{ for some } k) + \mathbb{P}(|l_k - l_{k'}| \le \epsilon \text{ for some } k, k').
\end{aligned}$$

For $\mathbf{X}$ in each class, $l_k - l_{k'}$ is normally distributed with variance $\mathrm{vec}(\mathbf{B}_k - \mathbf{B}_{k'})^T \boldsymbol{\Sigma} \mathrm{vec}(\mathbf{B}_k - \mathbf{B}_{k'})$. Hence, by Markov inequality,

$$\begin{aligned}
\mathbb{P}(|l_k - l_{k'}| \le \epsilon \text{ for some } k, k') &\le \sum_{k''} \mathbb{P}(|l_k - l_{k'}| \le \epsilon \mid Y = k'') \pi_{k''} \\
&\le \sum_{k, k', k''} \pi_{k''} \frac{C\epsilon}{\{\mathrm{vec}(\mathbf{B}_k - \mathbf{B}_{k'})^T \boldsymbol{\Sigma} \mathrm{vec}(\mathbf{B}_k - \mathbf{B}_{k'})\}^{1/2}} \\
&\le CK^2 \epsilon.
\end{aligned}$$

Also, notice that conditional on training data, $\widehat{l}_k - l_k$ is normally distributed with mean $\mu(k, k') = \widehat{c}_k - c_k + \mathrm{vec}(\boldsymbol{\mu}_{k'})^T (\mathrm{vec}(\widehat{\mathbf{B}}_k - \mathbf{B}_k))$ and variance $\mathrm{vec}(\widehat{\mathbf{B}}_k - \mathbf{B}_k)^T \boldsymbol{\Sigma} \mathrm{vec}(\widehat{\mathbf{B}}_k - \mathbf{B}_k)$. By Chernoff bound ,we have

$$\begin{aligned}
\mathbb{P}(|\widehat{l}_k - l_k| \ge \epsilon/2 \text{ for some } k) &= \sum_{k'} \pi_{k'} \mathbb{P}(|\widehat{l}_k - l_k| \ge \epsilon/2 \mid Y = k') \\
&\le 2c_1 K \exp\{-\frac{c_2 (\epsilon/2 - \mu(k, k'))^2}{\max_k \mathrm{vec}(\widehat{\mathbf{B}}_k - \mathbf{B}_k)^T \boldsymbol{\Sigma} \mathrm{vec}(\widehat{\mathbf{B}}_k - \mathbf{B}_k)}\}.
\end{aligned}$$

Let $\epsilon = C_3^* (\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})$, where $(C_3^*/2 - C_1^*)/(\lambda_{max}(\boldsymbol{\Sigma}) C_2^*) = c_M$, then on event $\mathcal{B}_1 \cap \mathcal{B}_2$, we have

$$|\widehat{R} - R| \le CC_3^* K^2 (\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n}) + \exp(-c_M).$$

Since last equation holds for any c_M, we know that $|\widehat{R} - R| = O(\max_m \sqrt{u_m p_m/(np_{-m})} + \sqrt{u/n})$ on event $\mathcal{B}_1 \cap \mathcal{B}_2$.

□

Combine Theorems S2, S3, and S4 together, we have Theorem 4 in the main paper.

References

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D. S., and Craddock, R. C. (2017). The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286.

Cook, R., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20(3):927–960.

Cook, R. D. and Zhang, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300.

Fackler, P. L. (2005). Notes on matrix calculus. *Technical Report*.

Gupta, A. K. and Nagar, D. K. (2018). *Matrix variate distributions*. Chapman and Hall/CRC.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146.

Manceur, A. M. and Dutilleul, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239:37–49.

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

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, pages 221–242.

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

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

Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434.

Yang, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Zhang, X. and Mai, Q. (2018). Model-free envelope dimension selection. *Electronic Journal of Statistics*, 12(2):2193–2216.