# Supplementary Materials for "Covariate-Adjusted Tensor Classification in High Dimensions"

Yuqing Pan, Qing Mai, and Xin Zhang

## A Additional Simulation Results

Recall that we include various popular and state-of-the-art classification methods as competitors in Sections 6 & 7. We used the following R packages that implement our competing methods in numerical studies: *penalizedLDA* for $\ell_1$-FDA, *sparseLDA* for SOS, *glmnet* for $\ell_1$-GLM, *randomForest* for Random Forest and $penalizedSVM$ for $\ell_1$-SVM. For other competitors, we downloaded R code from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4714801/#SD1` for MDA and PMDA, MATLAB code from `https://hua-zhou.github.io/TensorReg/` for CP-GLM, and MATLAB code from `http://www.scholat.com/portalPaperInfo.html?paperID=25520&Entry=laizhihui` for STDA. The author of (Li and Schonfeld, 2014) kindly provides the code for CMDA and DGTDA.

For methods with only one tuning parameter, we select the tuning parameter based on the error rates on validation set. For methods with multiple tuning parameters, we select them as follows. For random forest, we tuned the number of variables randomly selected as candidate in each split and used 400 ensembled trees. For STDA, we did a grid search on both the reduced tensor dimension and penalty parameter. For CMDA and DGTDA, we let the reduced tensor have the same dimension on each mode in simulation and ADHD, and we fixed the reduced dimension to be $10 \times 2$ in the CSA data.

## A.1 Variable Selection for results for Models (C1–C3i)

In Table A.1 we compare the variable selection results of CATCH, $\ell_1$-FDA, SOS as well as $\ell_1$-GLM and $\ell_1$-SVM. For each methods, two tasks are considered: one with only tensor predictor $\mathbf{X}$ (or $\mathrm{vec}(\mathbf{X})$), and one with both tensor $\mathbf{X}$ and covariates $\mathbf{U}$. It turns out that CATCH results in higher TPR and relatively lower FPR than others. Comparing the performance of methods with and without covariates, we found that CATCH with covariates has much higher TPR compared to CATCH without covariates, except in (C3a), where there is no dependence relationship between the covariates and the tensor. Meanwhile, in (C3b), the covariates have not influence on the classification by themselves, but help with the variable selection in $\mathbf{X}$ nonetheless.

## A.2 Higher dimension tensors

To demonstrate the applicability of CATCH in higher dimensional tensor data, we further considered variants of models (C1)–(C3) where we increase the tensor dimensions from $30\times36\times30$ to $80 \times 80 \times 80$, that is $512,000$ voxels in total. Parallel to models (C1)–(C3) in the paper, we now have three new models (C1H)–(C3H) where the sample size are also increased slightly from $n_k = 75$ to $n_k = 100$ in each class. Many methods (CMDA, DGTDA, RF, $\ell_1$-SVM) become practically inapplicable because of the prohibitive computational costs. Therefore, we only compare CATCH with $\ell_1$-FDA, $\ell_1$-logistic regression, GP-GLM and DGTDA. The classification errors are summarized in Table A.2, and the variable selection in Table A.3. Not surprisingly, CATCH continues to achieve better accuracy and variable selection than the competitors.

# B The Autism Spectrum disorder (ASD) dataset

In this section we study a real data example similar to the ADHD data in the paper, where we have a three-way tensor predictor from structural magnetic resonance imaging (MRI) and additional covariates (age and sex). The original data is from the Autism Brain Imaging Data

| Rate(%) | C1 | | C2 | | C3 | | C3a | | C3b | | C3i | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR |
| CATCH | 49.8 | 0.0 | 53.4 | 0.0 | 60.6 | 0.1 | 69.8 | 0.1 | 64.3 | 0.1 | 58.1 | 0.1 |
| without **U** | (2.0) | (0.0) | (1.9) | (0.0) | (1.4) | (0.0) | (1.2) | (0.0) | (1.0) | (0.0) | (1.4) | (0.0) |
| **CATCH** | 92.1 | 0.1 | 66.9 | 0.1 | 69.4 | 0.1 | 61.2 | 0.1 | 73.2 | 0.1 | 70.0 | 0.1 |
| | (0.7) | (0.0) | (0.1) | (0.0) | (1.0) | (0.0) | (1.0) | (0.0) | (1.2) | (0.0) | (1.1) | (0.0) |
| $\ell_1$-FDA | 96.0 | 33.9 | 100 | 25.8 | 100 | 24.0 | 100 | 21.8 | 100 | 20.9 | 97.9 | 28.8 |
| without **U** | (2.0) | (2.0) | (0) | (1.5) | (0) | (1.5) | (0) | (0.8) | (0) | (0.9) | (1.4) | (2.5) |
| $\ell_1$-FDA | 96.0 | 34.7 | 100 | 25.4 | 100 | 24.0 | 100 | 21.7 | 100 | 20.9 | 97.9 | 28.8 |
| with **U** | (2.0) | (2.1) | (0) | (1.5) | (0) | (1.5) | (0) | (0.8) | (0) | (0.9) | (1.4) | (2.5) |
| SOS | 26.8 | 0.0 | 55.8 | 0.1 | 53.4 | 0.1 | 61.3 | 0.1 | 54.8 | 0.1 | 58.0 | 0.1 |
| without **U** | (1.0) | (0.0) | (1.0) | (0.0) | (1.1) | (0.0) | (1.1) | (0.0) | (1.0) | (0.0) | (1.1) | (0.0) |
| SOS | 26.9 | 0.0 | 55.9 | 0.1 | 53.4 | 0.1 | 60.4 | 0.1 | 54.8 | 0.0 | 57.9 | 0.1 |
| with **U** | (1.0) | (0.0) | (1.0) | (0.0) | (1.1) | (0.0) | (1.1) | (0.0) | (1.0) | (0.0) | (1.1) | (0.0) |
| $\ell_1$ logistic | 20.8 | 0.0 | 54.1 | 0.0 | 51.1 | 0.0 | 57.1 | 0.0 | 52.1 | 0.1 | 58.9 | 0.1 |
| without **U** | (0.1) | (0.0) | (1.7) | (0.0) | (1.2) | (0.0) | (1.1) | (0.1) | (1.1) | (0.1) | (1.1) | (0.0) |
| $\ell_1$ logistic | 21.4 | 0.1 | 53.9 | 0.0 | 50.6 | 0.0 | 59.5 | 0.1 | 50.1 | 0.1 | 58.8 | 0.1 |
| with **U** | (0.9) | (0.0) | (1.1) | (0.0) | (1.2) | (0.0) | (1.0) | (0.0) | (1.1) | (0.1) | (1.1) | (0.0) |
| $\ell_1$ SVM | 43.9 | 0.1 | 55.9 | 0.1 | 31.3 | 0.1 | 56.1 | 0.3 | 43.6 | 0.1 | 56.5 | 0.12 |
| without **U** | (0.2) | (0.0) | (0.3) | (0.0) | (0.0) | (0.0) | (0.1) | (0.0) | (0.1) | (0.0) | (0.25) | (0.00) |
| $\ell_1$ SVM | 81.0 | 0.5 | 55.9 | 0.1 | 31.4 | 0 | 68.7 | 0.4 | 74.7 | 0.2 | 56.5 | 0.2 |
| with **U** | (0.3) | (0.0) | (0.3) | (0.0) | (0.1) | (0.0) | (0.1) | (0.0) | (0.3) | (0.0) | (0.25) | (0.0) |

Table A.1: Variable selection results for Models (C1–C3) & (C3a), (C3b), (C3i). True Positive Rate (TPR) and False Positive Rate (FPR) are defined in 6.1. Means and standard errors (in parentheses) are based on 100 replicates.

Exchange (ABIDE) (available at `http://fcon_1000.projects.nitrc.org/indi/abide/`), where the goal is to study the characterizations of Autism Spectrum disorder (ASD) among 23 laboratories. We use the dataset from New York University that contains $n = 182$ observations of 78 ASD subjects and 104 normal controls. For each subject, a $91 \times 109 \times 91$ tensor predictor is provided. We remove the fibers with all zero variables or limited nonzero variables, which lie on the outer shell of the tensor, so that the (nonzero) tensor predictor has dimension $71 \times 89 \times 71$. We record the 5-fold and 10-fold cross validation errors in Table B.1. Given the large dimension and slow speed of SOS, we complete 100 replicates for CATCH but only 20 replicates for SOS.

| Error rate (%) | C1H | | C2H | | C3H | |
|---|---|---|---|---|---|---|
| | $(\mathbf{X}, \mathbf{U})$ | $\mathbf{X}$ | $(\mathbf{X}, \mathbf{U})$ | $\mathbf{X}$ | $(\mathbf{X}, \mathbf{U})$ | $\mathbf{X}$ |
| Bayes | 5.58 | NA | 11.22 | NA | 8.34 | NA |
| | (0.22) | NA | (0.32) | NA | (0.30) | NA |
| **CATCH** | 10.03 | 30.33 | 16.11 | 20.01 | 10.58 | 17.12 |
| | (0.36) | (0.44) | (0.37) | (0.65) | (0.35) | (0.62) |
| CP-GLM | 43.33 | 43.84 | 29.23 | 29.02 | 16.71 | 17.22 |
| | (1.01) | (1.20) | (1.15) | (1.28) | 0.57 | 0.73 |
| DGTDA | 49.59 | 49.51 | 50.39 | 50.58 | 46.44 | 46.47 |
| | (0.50) | (0.48) | (0.52) | (0.52) | (0.45) | (0.45) |
| $\ell_1$-FDA | 43.75 | 43.72 | 34.59 | 34.59 | 35.6 | 35.57 |
| | (0.47) | (0.47) | (0.51) | (0.51) | (0.52) | (0.51) |
| $\ell_1$-GLM | 33.75 | 33.75 | 16.63 | 16.63 | 13.62 | 13.61 |
| | (0.53) | (0.53) | (0.38) | (0.38) | (0.35) | (0.35) |

Table A.2: Higher dimension simulations. We apply these methods on using only tensor and both tensor and covariates two scenarios. Due to memory limitation, results are recorded from 100 replicates and each replicate has 100 test observations.

| $\mathbf{X}$ | C1H | | C2H | | C3H | |
|---|---|---|---|---|---|---|
| | TPR | FPR | TPR | FPR | TPR | FPR |
| **CATCH** | 69.5 | 0.01 | 56.63 | 0.00 | 57.8 | 0.00 |
| | (2.82) | (0.00) | (1.97) | (0.00) | (1.93) | (0.00) |
| $\ell_1$-FDA | 98 | 51.18 | 100 | 24.46 | 100 | 45.61 |
| | (1.41) | (3.21) | (0.00) | (1.25) | (0.00) | (3.01) |
| $\ell_1$-GLM | 19.5 | 0.00 | 58.63 | 0.00 | 55.81 | 0.00 |
| | (0.90) | (0.00) | (1.04) | (0.00) | (0.98) | (0.00) |
| $(\mathbf{X}, \mathbf{U})$ | C1H | | C2H | | C3H | |
| | TPR | FPR | TPR | FPR | TPR | FPR |
| **CATCH** | 92.69 | 0.01 | 70.00 | 0.01 | 75.38 | 0.01 |
| | (0.73) | (0.00) | (1.12) | (0.00) | (1.12) | (0.00) |
| $\ell_1$-FDA | 98 | 51.81 | 100 | 24.46 | 100 | 46.22 |
| | (1.41) | (3.22) | (0.00) | (1.25) | (0.00) | (3.02) |
| $\ell_1$-logistic | 19.5 | 0.00 | 58.63 | 0.00 | 55.88 | 0.00 |
| | (0.90) | (0.00) | (1.04) | (0.00) | (0.98) | (0.00) |

Table A.3: Variable selection results.

Classification of ASD in this dataset is more difficult than ADHD study, given the subtle relationship between brain anatomy and autism. Classification of ASD has error rates over 40%

| Error rate (%) | CATCH | SOS | Random guess |
|---|---|---|---|
| 5 fold | 43.89 | 48.57 | 49.53 |
| | (0.31) | (0.29) | (0.04) |
| 10 fold | 43.77 | 48.91 | 49.06 |
| | (0.29) | (0.31) | (0.04) |

Table B.1: ASD data analysis. Error rates and standard error (in parentheses) of 5-fold and 10-fold cross validation.

across literature. It is shown that the error rate of SOS is very close to random guessing rate, while CATCH performs much better.

# C Technical Details for Example 1 and some related results

First, we prove the following Proposition C.1. Although Proposition C.1 is a well-known result in the LDA literature, we still present its proof for completeness.

**Proposition C.1.** *For $Y \in \{1, 2\}, \mathbf{W} \in \mathbb{R}^p$, assume that $(\mathbf{W}, Y)$ follows a linear discriminant analysis model in the sense that $\mathbf{W} \mid Y = k \sim N(\boldsymbol{\zeta}_k, \boldsymbol{\Omega})$. Further assume that $\pi_1 = \pi_2 = 1/2$. Then the lowest classification error rate possible is*

$$1 - \Phi(\frac{\sqrt{(\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1)^\intercal \boldsymbol{\Omega}^{-1} (\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1)}}{2}) \tag{C.1}$$

*where $\Phi$ is the cumulative distribution function of the standard normal random variable.*

**Proof of Proposition C.1.** It is well-known that the best classifier under this model classifies an observation to Class 2 if and only if $(\mathbf{W} - \frac{\boldsymbol{\zeta}_1 + \boldsymbol{\zeta}_2}{2})^\intercal \boldsymbol{\Omega}^{-1} (\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1) > 0$. It is easy to show that

$$\begin{aligned} & \Pr((\mathbf{W} - \frac{\boldsymbol{\zeta}_1 + \boldsymbol{\zeta}_2}{2})^\intercal \boldsymbol{\Omega}^{-1} (\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1) > 0 \mid Y = 1) && \text{(C.2)} \\ = \; & \Pr((\mathbf{W} - \frac{\boldsymbol{\zeta}_1 + \boldsymbol{\zeta}_2}{2})^\intercal \boldsymbol{\Omega}^{-1} (\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1) < 0 \mid Y = 2) && \text{(C.3)} \\ = \; & 1 - \Phi(\frac{\sqrt{(\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1)^\intercal \boldsymbol{\Omega}^{-1} (\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1)}}{2}) && \text{(C.4)} \end{aligned}$$

Hence, the classification error rate is

$$\Pr((\mathbf{W} - \frac{\boldsymbol{\zeta}_1 + \boldsymbol{\zeta}_2}{2})^\intercal \boldsymbol{\Omega}^{-1}(\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1) > 0 \mid Y = 1) \Pr(Y = 1) \tag{C.5}$$

$$+ \quad \Pr((\mathbf{W} - \frac{\boldsymbol{\zeta}_1 + \boldsymbol{\zeta}_2}{2})^\intercal \boldsymbol{\Omega}^{-1}(\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1) < 0 \mid Y = 2) \Pr(Y = 2) \tag{C.6}$$

$$= \quad 1 - \Phi(\frac{\sqrt{(\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1)^\intercal \boldsymbol{\Omega}^{-1}(\boldsymbol{\zeta}_2 - \boldsymbol{\zeta}_1)}}{2}) \tag{C.7}$$

□

Define $R(\mathbf{U}, \mathbf{X})$ as the lowest classification error rate possible if we build the classifier based on $\mathbf{X}$ and $\mathbf{U}$, and similarly, $R(\mathbf{X})$ as that based only on $\mathbf{X}$. For simplicity, we consider the binary classification case with equal class probability of $0.5$, and assume that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ because otherwise $(\mathbf{U}, \mathbf{X})$ are jointly independent of $Y$ and the best classifier is random guessing.

**Lemma C.1.** *For $K = 2$, $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$, $\pi_1 = \pi_2$, $\boldsymbol{\alpha} \neq 0$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, we have $R(\mathbf{U}, \mathbf{X}) < R(\mathbf{X})$.*

**Proof of Lemma C.1.** Straightforward calculation shows that

$$\begin{pmatrix} \mathbf{U} \\ \mathrm{vec}(\mathbf{X}) \end{pmatrix} \mid (Y = k) \sim N\left( \begin{pmatrix} \boldsymbol{\phi}_k \\ \boldsymbol{\mu}_k + \boldsymbol{\alpha} \bar{\times}_{(M+1)} \boldsymbol{\phi}_k \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} \right), \tag{C.8}$$

where $\boldsymbol{\Omega}_{11} = \boldsymbol{\Psi}, \boldsymbol{\Omega}_{12} = \boldsymbol{\Omega}_{21}^\intercal = \boldsymbol{\Psi}\boldsymbol{\alpha}_{(M+1)}, \boldsymbol{\Omega}_{22} = \otimes_{m=M}^{m=1} \boldsymbol{\Sigma}_m + \boldsymbol{\alpha}_{(M+1)}^\intercal \boldsymbol{\Psi} \boldsymbol{\alpha}_{(M+1)}$.

By Proposition C.1 and applying the block inverse formula, we have

$$R(\mathbf{U}, \mathbf{X}) \quad = \quad 1 - \Phi\left( \frac{\sqrt{\Delta_1 + \Delta_2}}{2} \right), \tag{C.9}$$

where we have used the fact that $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$, and,

$$\Delta_1 \quad = \quad (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\intercal \boldsymbol{\Omega}_{22}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \tag{C.10}$$

$$\Delta_2 \quad = \quad (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\intercal \{\boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\Omega}_{21} (\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\Omega}_{21})^{-1} \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-1}\} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \tag{C.11}$$

It is easy to see that $\Delta_1 > 0, \Delta_2 > 0$ under our assumptions.

On the other hand, by (C.8), $(\mathrm{vec}(\mathbf{X}), Y)$ also follows the LDA model. Hence, because of $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$,

$$R(\mathbf{X}) = 1 - \Phi(\frac{\sqrt{\Delta_1}}{2}) \tag{C.12}$$

where $\Delta_1$ is the same as defined in (C.10). Because $\Delta_2 > 0$, we must have $R(\mathbf{U}, \mathbf{X}) < R(\mathbf{X})$.

□

**Technical Details for Example 1.** From straightforward calculation, we have,

$$\mathbf{\Sigma}_a \equiv \mathrm{cov}\left(\begin{pmatrix} \mathrm{vec}(\mathbf{X}) \\ \mathbf{U} \end{pmatrix} \mid Y\right) = \begin{pmatrix} 1+\alpha^2 & \alpha^2 & 0 & 0 & \alpha \\ \alpha^2 & 1+\alpha^2 & 0 & 0 & \alpha \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \alpha & \alpha & 0 & 0 & 1 \end{pmatrix}.$$

Therefore, $\mathbf{\Sigma}_b \equiv \mathrm{cov}(\mathrm{vec}(\mathbf{X}) \mid Y)$ is the upper left $4 \times 4$ block of $\mathbf{\Sigma}_a$. It is thus easy to verify that,

$$\mathbf{\Sigma}_a^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & -\alpha \\ 0 & 1 & 0 & 0 & -\alpha \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -\alpha & -\alpha & 0 & 0 & 1+2\alpha^2 \end{pmatrix}, \quad \mathbf{\Sigma}_b^{-1} = \mathbf{I}_4 - (2+\alpha^{-2})^{-1} \cdot \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then from Proposition C.1 and Lemma C.1, we have, $R(\mathbf{X}, U) = R(X_{11}, U) = 1 - \Phi\left(\sqrt{(\mathbf{\Sigma}_a^{-1})_{11}}\right) = 1 - \Phi(1)$; $R(X_{11}) = 1 - \Phi\left(\mathrm{cov}^{-1}(X_{11} \mid Y)\right) = 1 - \Phi(1/\sqrt{1+\alpha^2})$; because $\mathbf{\Sigma}_b^{-1}(\mathrm{E}(\mathrm{vec}(\mathbf{X}) \mid Y = 2) - \mathrm{E}(\mathrm{vec}(\mathbf{X}) \mid Y = 1))$ contains two nonzero elements, $R(\mathbf{X}) = R(X_{11}, X_{21}) = 1 - \Phi\left(\sqrt{(\mathbf{\Sigma}_b^{-1})_{11}}\right) = 1 - \Phi(\sqrt{2+\alpha^2}/\sqrt{2+2\alpha^2})$. □

# D Proofs for all Lemmas and Propositions

**Proof of Proposition 1.** First, note that (2.5) is equivalent to

$$\widehat{Y} = \arg\max_{k=1,\ldots,K} \{\log(\pi_k/\pi_1) + \log \frac{f_k(\mathbf{U}, \mathbf{X})}{f_1(\mathbf{U}, \mathbf{X})}\} \tag{D.1}$$

Under the model in (2.1), we have that, within Class $k$, the probability density function for $(\mathbf{U}, \mathbf{X})$ is

$$f_k(\mathbf{U}, \mathbf{X}) = \frac{1}{\det(2\pi\mathbf{\Psi})^{1/2}} \cdot \exp(-\frac{1}{2}(\mathbf{U} - \boldsymbol{\phi}_k)^\intercal \mathbf{\Psi}^{-1}(\mathbf{U} - \boldsymbol{\phi}_k)) \cdot \frac{1}{\det(2\pi \otimes_{m=M}^{m=1} \mathbf{\Sigma}_m)^{1/2}}$$
$$\cdot \exp(-\frac{1}{2}\mathrm{vec}^\intercal(\mathbf{X} - \boldsymbol{\alpha}\bar{\times}_{(M+1)}\mathbf{U} - \boldsymbol{\mu}_k)(\otimes_{m=M}^{m=1}\mathbf{\Sigma}_m)^{-1}\mathrm{vec}(\mathbf{X} - \boldsymbol{\alpha}\bar{\times}_{(M+1)}\mathbf{U} - \boldsymbol{\mu}_k)) \tag{D.2}$$

We have the desired conclusion by substituting (D.2) into (D.1). □

**Proof of Lemma 1.** Under the CATCH model, the negative log-likelihood up to additive and scaling constants can be written as

$$\begin{aligned} & \ell(\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}) \\ \simeq \; & \log|\boldsymbol{\Sigma}| + n^{-1} \sum_{k=1}^{K} \sum_{Y^i=k} \operatorname{tr}\left\{ \boldsymbol{\Sigma}^{-1} \operatorname{vec}\left(\mathbf{X}^i - \boldsymbol{\mu}_k - \boldsymbol{\alpha} \bar{\times}_{(M+1)} \mathbf{U}^i\right) \operatorname{vec}^{\intercal}\left(\mathbf{X}^i - \boldsymbol{\mu}_k - \boldsymbol{\alpha} \bar{\times}_{(M+1)} \mathbf{U}^i\right) \right\}, \end{aligned}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_1$. Then we can first see that the MLE for $\boldsymbol{\mu}_k$ (since it only appears in the log-likelihood of class $k$) is

$$\widehat{\boldsymbol{\mu}}_k = \overline{\mathbf{X}}_k - \boldsymbol{\alpha} \bar{\times}_{(M+1)} \overline{\mathbf{U}}_k,$$

then the partially minimized $\ell(\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma})$ becomes

$$\begin{aligned} \ell(\boldsymbol{\alpha}, \boldsymbol{\Sigma}) \; \simeq \; & \log|\boldsymbol{\Sigma}| \\ + \; & n^{-1} \sum_{i=1}^{n} \operatorname{tr}\left\{ \boldsymbol{\Sigma}^{-1} \operatorname{vec}\left(\widetilde{\mathbf{X}}^i - \boldsymbol{\alpha} \bar{\times}_{(M+1)} \widetilde{\mathbf{U}}^i\right) \operatorname{vec}^{\intercal}\left(\widetilde{\mathbf{X}}^i - \boldsymbol{\alpha} \bar{\times}_{(M+1)} \widetilde{\mathbf{U}}^i\right) \right\}, \end{aligned}$$

where the second term no longer depends on $k$ explicitly and the centered data is defined as $\widetilde{\mathbf{X}}^i = \mathbf{X}^i - \overline{\mathbf{X}}_k$ and $\widetilde{\mathbf{U}}^i = \mathbf{U}^i - \overline{\mathbf{U}}_k$ for observation $i$ in class $k$. Also, notice that

$$\operatorname{vec}\left(\widetilde{\mathbf{X}}^i - \boldsymbol{\alpha} \bar{\times}_{(M+1)} \widetilde{\mathbf{U}}^i\right) = \operatorname{vec}(\widetilde{\mathbf{X}}^i) - \boldsymbol{\alpha}_{(M+1)}^{\intercal} \widetilde{\mathbf{U}}^i.$$

It is thus clear that the likelihood function under the CATCH model coincides with that of the tensor response regression model (Li and Zhang, 2017) with the modified $\widetilde{\mathbf{X}}$ as response and the modified $\widetilde{\mathbf{Z}}$ as predictor and with the same form of (partial) likelihood. Therefore, we have the desired conclusion. □

**Proof of Lemma 2.** Write $\mathbf{W} = [\![\mathbf{Z}; \boldsymbol{\Omega}_1^{1/2}, \ldots, \boldsymbol{\Omega}_m^{1/2}]\!]$. Then we have that all the entries in $\mathbf{Z}$ are independent of each other.

$$\mathrm{E}(\mathbf{W}_{(j)} \mathbf{W}_{(j)}^{\intercal}) = \boldsymbol{\Omega}_j^{1/2} \left\{ \mathrm{E}(\mathbf{Z}_{(j)} (\bigotimes_{l \neq j} \boldsymbol{\Omega}_l) \mathbf{Z}_{(j)}^{\intercal}) \right\} \boldsymbol{\Omega}_j^{1/2} \tag{D.3}$$

Define $\mathbf{V} = \bigotimes_{l \neq j} \mathbf{\Omega}_l$ and $\mathbf{H} = \mathrm{E}\left\{\mathbf{Z}_{(j)}(\bigotimes_{l \neq j} \mathbf{\Omega}_l)\mathbf{Z}_{(j)}^{\intercal}\right\}$. Then we have

$$h_{ij} = \mathrm{E}(\sum_{k,l} v_{kl} Z_{ik} Z_{jl}) = \begin{cases} \sum_i v_{ii} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{D.4}$$

It follows that $\mathbf{H} = \mathrm{tr}(\mathbf{V}) \cdot \mathbf{I}$. Then because of $\mathrm{tr}(\mathbf{V}) = \prod_{l \neq j} \mathrm{tr}(\mathbf{\Omega}_l)$, we have the desired conclusion.

□

**Lemma D.1.** *For each $j = 1, \ldots, p$, we can define a sequence $\{s_m, m = 2, \ldots, M+1\}$ as follows: $s_{M+1} = j$ and $s_m = s_{m+1} \bmod (\prod_{i=1}^{m-1} p_i)$ for $m = 2, \ldots, M$. Then*

$$(\bigotimes_{m=M}^{1} \mathbf{\Sigma}_m)_{\cdot j} = \mathrm{vec}(\bigotimes_{m=M}^{1} \mathbf{\Sigma}_{m,\cdot j_m}) \tag{D.5}$$

*where $j_1 = s_2 \bmod p_1$ and $j_m = \lceil \frac{s_{m+1}}{\prod_{i=1}^{m-1} p_i} \rceil$ for $m = 2, \ldots, M$.*

**Proof of Lemma D.1.** By the definition of Kronecker product, we have that, for two matrices $\mathbf{\Omega}_1 \in \mathbb{R}^{q_1}, \mathbf{\Omega}_2 \in \mathbb{R}^{q_2}$, we have

$$(\mathbf{\Omega}_2 \otimes \mathbf{\Omega}_1)_{\cdot j} = \mathrm{vec}(\mathbf{\Omega}_{2,\cdot \lceil \frac{j}{p_1} \rceil} \otimes \mathbf{\Omega}_{1,\cdot (j \bmod p_1)}). \tag{D.6}$$

where we restrict that $a \bmod b = b$ if $a$ is divisible by $b$. Hence, we have

$$\begin{aligned} (\mathbf{\Sigma}_M \otimes \cdots \otimes \mathbf{\Sigma}_1)_{\cdot j} &= \{\mathbf{\Sigma}_M \otimes (\mathbf{\Sigma}_{M-1} \cdots \otimes \mathbf{\Sigma}_1)\}_{\cdot j} && \text{(D.7)} \\ &= \mathbf{\Sigma}_{M,\cdot j_M} \otimes (\mathbf{\Sigma}_{M-1} \cdots \otimes \mathbf{\Sigma}_1)_{\cdot s_{M-1}} && \text{(D.8)} \\ &= \bigotimes_{m=M}^{m=1} \mathbf{\Sigma}_{m,\cdot j_m}, && \text{(D.9)} \end{aligned}$$

where the last equality is obtained by recursively applying (D.6). □

**Proof of Lemma 3.** For simplicity, we only present the proofs for no covariate case. Without loss of generality, assume that the observations are arranged such that $Y^i = 1$ for $i = 1, \ldots, n_1$; $Y^i = 2$ for $i = n_1 + 1, \ldots, n_1 + n_2$ and so on.

Define $\mathbf{X}_{(m)}$ as the $np_{-m} \times p_m$ matrix obtained by stacking the observations $\mathbf{X}^i_{(m)}, i = 1, \ldots, n$.

$$\mathbf{A} = \begin{pmatrix} (\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\intercal) \otimes \mathbf{I}_{p_{-m}} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & (\mathbf{I}_{n_1} - \frac{1}{n_K}\mathbf{1}_{n_K}\mathbf{1}_{n_K}^\intercal) \otimes \mathbf{I}_{p_{-m}} \end{pmatrix} \tag{D.10}$$

By (3.6), $\widehat{\mathbf{\Sigma}}_m \propto (\mathbf{X}^*_{(m)})^\intercal \mathbf{X}^*_{(m)}$, where

$$\mathbf{X}^*_{(m)} = \begin{pmatrix} \mathbf{X}^i_{(m)} - \widehat{\boldsymbol{\mu}}_1 \\ \cdots \\ \mathbf{X}^i_{(m)} - \hat{\boldsymbol{\mu}}_K \end{pmatrix} = \mathbf{A}\mathbf{X}_{(m)} \tag{D.11}$$

It follows that $\widehat{\mathbf{\Sigma}}_m$ is positive semi-definite, and $\text{rank}(\widehat{\mathbf{\Sigma}}) = \text{rank}(\mathbf{X}^*_{(m)})$. It leaves to show that $\text{rank}(\mathbf{X}^*_{(m)}) = p_m$.

Because $\text{rank}(\mathbf{A}) = (n-K)p_{-m}$, while $\mathbf{X}_{(m)}$ is randomly sampled from a continuous distribution, we have that, when $(n-K)p_{-m} > p_m$, with a probability of 1, $\mathbf{X}_{(m)^*}$ is column full-rank, i.e, $rank(\mathbf{X}^*_{(m)}) = p_m$. Consequently, with a probability of 1, $\widehat{\mathbf{\Sigma}}_m$ is full rank. □

**Proof of Lemma 4.** Set the partial derivative of $\mathcal{L}(\mathbf{C}_2, \ldots, \mathbf{C}_K)$ with respect to $\mathbf{C}_k$ to be zero and we have

$$\frac{\partial \mathcal{L}(\mathbf{C}_2, \ldots, \mathbf{C}_K)}{\partial \mathbf{C}_k} = 2[\![\mathbf{C}_k; \mathbf{\Sigma}_2, \ldots, \mathbf{\Sigma}_K]\!] - 2(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) = 0 \tag{D.12}$$

Solve this equation for $k = 2, \ldots, K$ and we have that $(\mathbf{B}_2, \cdots, \mathbf{B}_K)$ minimizes $\mathcal{L}(\mathbf{C}_2, \ldots, \mathbf{C}_K)$.

□

**Proof of Lemma 5.** By Lemma 1 in Mai et al. (2017), we have that, given $\boldsymbol{\beta}_{\cdot j'}, j' \neq j$, the solution of $\boldsymbol{\beta}_{\cdot j}$ to (4.1) is

$$\arg\min_{\boldsymbol{\beta}_{\cdot j}} \sum_{k=2}^{K} \frac{1}{2}(\beta_{kj} - \widetilde{\beta}_{kj})^2 + \frac{\lambda}{\widehat{\sigma}_{jj}} \|\boldsymbol{\beta}_{\cdot j}\| \tag{D.13}$$

where $\widetilde{\beta}_{kj} = \dfrac{(\widehat{\nu}_{kj} - \widehat{\nu}_{1j}) - \sum_{l \neq j} \widehat{\sigma}_{lj}\beta_{kl}}{\widehat{\sigma}_{jj}}$. Because $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_M \otimes \cdots \otimes \widehat{\mathbf{\Sigma}}_1$, by Lemma D.1 we have

$$\sum_{l \neq j} \widehat{\sigma}_{lj}\beta_{kl} = [\![\mathbf{B}^j_k; \widehat{\mathbf{\Sigma}}_{1,\cdot j_1}, \cdots, \widehat{\mathbf{\Sigma}}_{M,\cdot j_M}]\!], \widehat{\sigma}_{jj} = \prod_{m=1}^{M} \widehat{\sigma}_{m,j_m j_m} \tag{D.14}$$

$j_m = \lceil \frac{s_{m+1}}{\prod_{i=1}^{m-1} p_i} \rceil$ for $m = 2, \ldots, M$ and $j_1 = s_2 \bmod p_1$. And we have the first part of conclusion, (4.5, 4.6). By checking the subgradients we can easily obtain the second part of conclusion, (4.7). □

**Proof of Lemma 6.** We first consider the computational cost for (4.8). For each $j$, by Lemma D.1, we need $M-1$ operations to find $j_1, \ldots, j_M$. Additional $M-1$ operations are needed to find $\prod_{m=1}^{M} \widehat{\sigma}_{m,j_m j_m}$. It leaves to calculate $[\![\mathbf{B}_k^j; \widehat{\mathbf{\Sigma}}_{1,\cdot j_1}, \cdots, \widehat{\mathbf{\Sigma}}_{M,\cdot j_M}]\!]$, which is equal to $(\otimes_{m=M}^{1} \widehat{\mathbf{\Sigma}}_{m,\cdot j_m}^{\intercal})\mathrm{vec}(\mathbf{B}_k^j)$. Since $\mathbf{B}_j^k$ is sparse with at most $d_t$ nonzero entries, it suffices to find the $d_t$ corresponding elements in $(\otimes_{m=M}^{1} \widehat{\mathbf{\Sigma}}_{m,\cdot j_m}^{\intercal})$. For any $j'$, it takes $M-1$ operations to find $j'_m, m = 1, \ldots, M$ such that $\prod_{m=1}^{M} \widehat{\sigma}_{m,j' m j_m} = (\otimes_{m=M}^{1} \widehat{\mathbf{\Sigma}}_{m,\cdot j_m}^{\intercal})_{j'}$. The calculation of $\prod_{m=1}^{M} \widehat{\sigma}_{m,j' m j_m}$ is again $O(M-1)$. Hence, the calculation of $[\![\mathbf{B}_k^j; \widehat{\mathbf{\Sigma}}_{1,\cdot j_1}, \cdots, \widehat{\mathbf{\Sigma}}_{M,\cdot j_M}]\!], k = 2, \ldots, K$ is $O(MKd_t)$. The computational cost for soft-thresholding is lower. Hence, for each $j$, the computational cost is $O(MKd_t)$. The update for all $j$ has the computational cost of $O(NKd_t p)$. □

# E Proof for Theorem 1 (Convergence analysis for Algorithm 1)

In this section, we present the results in Theorem 1.

Our proofs reply on the following definitions and propositions in Tseng (2001). For a generic objective function

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_N) = f_0(\mathbf{x}_1, \ldots, \mathbf{x}_N) + \sum_{k=1}^{N} f_k(\mathbf{x}_k) \tag{E.1}$$

where $f_0 : \mathbb{R}^{n_1 + \cdots + n_N} \mapsto \mathbb{R} \cup \{\infty\}$ and some $f_k : \mathbb{R}^{n_k} \mapsto \mathbb{R} \cup \{\infty\}, k = 1, \ldots, N$.

We have the following definitions:

**Definition E.1** (Gâteaux-differentiable (Bertsekas, 2016))**.** *For a function $F : \mathbb{R}^p \mapsto \mathbb{R}$, its Gâteaux derivative is defined as*

$$F'(\mathbf{x}; \mathbf{y}) = \lim_{\lambda \searrow 0} \frac{F(\mathbf{x} + \lambda \mathbf{y}) - F(\mathbf{x})}{\lambda} \tag{E.2}$$

*If $F'(\mathbf{x}; \mathbf{y})$ exists for all $\mathbf{y}$ at $\mathbf{x}$, then $F$ is Gâteaux-differentiable at $\mathbf{x}$.*

**Definition E.2** (Stationary points (Tseng, 2001)). *A point* $\mathbf{z}$ *is stationary for function* $h$ *if* $g(\mathbf{z}; \mathbf{d}) \geq 0$ *for any* $\mathbf{d}$*, where*

$$g(\mathbf{z}; \mathbf{d}) = \liminf_{\lambda \searrow 0} \frac{h(\mathbf{x} + \lambda \mathbf{d}) - h(\mathbf{x})}{\lambda} \tag{E.3}$$

**Definition E.3** (Regular (Tseng, 2001)). *A function* $f$ *is regular at* $\mathbf{z}$ *if*

$$f'(\mathbf{z}; \mathbf{d}) \geq 0, \text{ for any } \mathbf{d} = (\mathbf{d}_1, \ldots, \mathbf{d}_N) \text{ such that } f'(\mathbf{z}; (0, \ldots, \mathbf{d}_k, \cdots, 0)) \geq 0, k = 1, \ldots, N. \tag{E.4}$$

We will also use the following propositions.

**Proposition E.1** (A simplified version of Lemma 3.1 in Tseng (2001)). *If* $f_0$ *is Gâteaux-differentiable.,* $f$ *is regular at each* $\mathbf{z}$*.*

**Proposition E.2** (A simplified version of Theorem 4.1 in Tseng (2001)). *Assume that the level set* $\mathbf{X}^0 = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$ *is compact, where* $\mathbf{x}^0$ *is the initial value of the algorithm, and that* $f$ *is continuous on* $\mathbf{X}^0$*. Further assume that* $f(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ *is pseudoconvex in* $(\mathbf{x}_k, \mathbf{x}_i)$ *for all* $i, k \in \{1, \ldots, N\}$*, and if* $f$ *is regular at every* $\mathbf{x}$*, then the solution generated by the cyclic coordinate descent method converges to a stationary point of* $f$*.*

We make use of these results to prove Theorem 1.

**Lemma E.1.** *If* $\widehat{\mathbf{\Sigma}}$ *is positive definite, the set* $\mathcal{B}^0 = \{\boldsymbol{\beta} : L(\boldsymbol{\beta}) \leq L(\boldsymbol{\beta}^0)\}$ *is compact for any* $\boldsymbol{\beta}^0$*, where*

$$L(\boldsymbol{\beta}) = \sum_{k=2}^{K} \boldsymbol{\beta}_k^{\intercal} \widehat{\mathbf{\Sigma}} \boldsymbol{\beta}_k - 2\boldsymbol{\beta}_k^{\intercal}(\boldsymbol{\nu}_k - \boldsymbol{\nu}_1) + \lambda \sum_{j=1}^{p} \|\boldsymbol{\beta}_{\cdot j}\| \tag{E.5}$$

*Proof.* Because $L(\boldsymbol{\beta})$ is continuous, $\mathcal{B}^0$ must be closed. It suffices to verify that $\mathcal{B}^0$ is bounded. When $\widehat{\mathbf{\Sigma}}$ is invertible, it is easy to see that, for any $\boldsymbol{\beta}_k \in \mathbb{R}^p$,

$$\boldsymbol{\beta}_k^{\intercal} \widehat{\mathbf{\Sigma}} \boldsymbol{\beta}_k - 2(\widehat{\boldsymbol{\nu}}_k - \widehat{\boldsymbol{\nu}}_1)^{\intercal} \boldsymbol{\beta}_k \geq -(\widehat{\boldsymbol{\nu}}_k - \widehat{\boldsymbol{\nu}}_1)^{\intercal} \widehat{\mathbf{\Sigma}}^{-1} (\widehat{\boldsymbol{\nu}}_k - \widehat{\boldsymbol{\nu}}_1) \tag{E.6}$$

It follows that, for any $\boldsymbol{\beta} \in \mathcal{B}^0$,

$$-\sum_{k=2}^{K}(\widehat{\boldsymbol{\nu}}_k - \widehat{\boldsymbol{\nu}}_1)^{\intercal}\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\nu}}_k - \widehat{\boldsymbol{\nu}}_1) + \lambda\sum_{j=1}^{p}\|\boldsymbol{\beta}_{\cdot j}\| \leq L(\boldsymbol{\beta}) \leq L(\boldsymbol{\beta}^0) \tag{E.7}$$

which implies that

$$\sum_{j=1}^{p}\|\boldsymbol{\beta}_{\cdot j}\| \leq \frac{1}{\lambda}\sum_{k=2}^{K}(\widehat{\boldsymbol{\nu}}_k - \widehat{\boldsymbol{\nu}}_1)^{\intercal}\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\nu}}_k - \widehat{\boldsymbol{\nu}}_1) + L(\boldsymbol{\beta}^0) \text{ for any } \boldsymbol{\beta} \in \mathcal{B}^0. \tag{E.8}$$

Therefore, $\mathcal{B}^0$ is bounded and hence compact. □

**Proof of Theorem 1.** Let

$$f_0(\boldsymbol{\beta}) = \sum_{k=2}^{K}\{\boldsymbol{\beta}_k^{\intercal}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k - 2(\boldsymbol{\nu}_k - \boldsymbol{\nu}_1)^{\intercal}\boldsymbol{\beta}_k\}, f_j(\boldsymbol{\beta}_{\cdot j}) = \lambda\|\boldsymbol{\beta}_{\cdot j}\|, j = 1, \ldots, p \tag{E.9}$$

Apparently $L(\boldsymbol{\beta}) = f_0(\boldsymbol{\beta}) + \sum_{j=1}^{p} f_j(\boldsymbol{\beta}_{\cdot j})$. Because $f_0(\boldsymbol{\beta})$ is differentiable, by Proposition E.1, we have that $L(\boldsymbol{\beta})$ is regular at each $\boldsymbol{\beta}$. By Lemma 3, if $(n-K)p_{-m} > p_m$ for any $m$, with a probability of 1, $\widehat{\boldsymbol{\Sigma}}$ is positive definite. Combine this fact with Lemma E.1 and we have $\mathcal{B}^0 = \{\boldsymbol{\beta} : L(\boldsymbol{\beta}) \leq L(\boldsymbol{\beta}^0)\}$ is compact. Further note that $L(\boldsymbol{\beta})$ is continuous and convex. By Proposition E.2, the blockwise coordinate descent algorithm converges to a stationary point of $L(\boldsymbol{\beta})$. Because $L(\boldsymbol{\beta})$ is strictly convex, the stationary point is the global minimizer. □

# F Technical Lemmas for the proof of Theorems 2 and 3

Our proofs rely on the following result that is well-known in the literature; see Bickel and Levina (2004) for example.

**Lemma F.1.** *For bivariate normal random variables $(Z_1^i, Z_2^i)_{i=1}^n$, i.i.d. with mean zero and bounded variances, there exists $\epsilon_0 > 0$ such that for any $\epsilon$, $\epsilon$, $0 < \epsilon < \epsilon_0$, with a probability greater than $1 - 2\exp(-Cn\epsilon^2)$, we have*

$$|\frac{1}{n}\sum_{i=1}^{n} Z_1^i - \mathrm{E}Z_1| \leq \epsilon \tag{F.1}$$

$$|\frac{1}{n}\sum_{i=1}^{n} Z_1^i Z_2^i - \mathrm{E}(Z_1^i Z_2^i)| \leq \epsilon \tag{F.2}$$

In what follows, we prove some large deviation inequalities for matrix/tensor normal distributions.

**Lemma F.2.** *For $\mathbf{Z}^i \in \mathbb{R}^{p_1 \times p_2}$, $\mathbf{Z}^i \sim TN(0, \mathbf{I}_{p_1}, \mathbf{I}_{p_2})$, $i = 1, \ldots, n$, i.i.d., and $\mathbf{\Lambda} \in \mathbb{R}^{p_1 \times p_1}$ a diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_{p_1} \geq 0$, there exists $\epsilon_0 > 0$ such that for any $0 < \epsilon < \epsilon_0$, we have*

$$\Pr(\|\frac{1}{np_1}\sum_{i=1}^{n}(\mathbf{Z}^i)^\intercal \mathbf{\Lambda}\mathbf{Z}^i - \frac{1}{p_1}\mathrm{E}((\mathbf{Z}^i)^\intercal \mathbf{\Lambda}\mathbf{Z}^i)\|_{\max} > \epsilon) \leq 2p_2^2 \exp(-\frac{Cnp_1^2\epsilon^2}{\sum_{l=1}^{p_1}\lambda_l^2}) \tag{F.3}$$

**Proof of Lemma F.2.** Note that

$$\|\frac{1}{np_1}\sum_{i=1}^{n}(\mathbf{Z}^i)^\intercal \mathbf{\Lambda}\mathbf{Z}^i - \frac{1}{p_1}\mathrm{E}((\mathbf{Z}^i)^\intercal \mathbf{\Lambda}\mathbf{Z}^i)\|_{\max} \tag{F.4}$$

$$= \max_{s,k}|\frac{1}{np_1}\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l Z_{ls}^i Z_{lk}^i - \mathrm{E}(\frac{1}{p_1}\sum_{l=1}^{p_1}\lambda_l Z_{ls}^i Z_{lk}^i)| \equiv \max_{s,k} L(s,k) \tag{F.5}$$

We analyze $L(s,k)$ under two scenarios: $s = k$ and $s \neq k$. We first consider the case $s = k$.

$$L(s,s) = |\frac{1}{np_1}\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l (Z_{ls}^i)^2 - E\left(\frac{1}{p_1}\sum_{l=1}^{p_1}\lambda_l (Z_{ls}^i)^2\right)| \tag{F.6}$$

Because $(Z_{ls}^i)^2 \sim \chi^2(1)$, we have, for any $t > 0$, $Ee^{t\lambda_l (Z_{ls}^i)^2} = (1 - 2\lambda_l t)^{-1/2}$. Thanks to the independence among $Z_{ls}^i$ for $(i,l)$, we have

$$\mathrm{E}\{\exp(t\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l (Z_{ls}^i)^2)\} = \prod_{l=1}^{p_1}(1 - 2\lambda_l t)^{-n/2} \tag{F.7}$$

It follows from Chernoff bound that

$$\Pr\left(\frac{1}{np_1}\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l (Z_{ls}^i)^2 - E\left(\frac{1}{p_1}\sum_{l=1}^{p_1}\lambda_l (Z_{ls}^i)^2\right) \geq \epsilon\right) \tag{F.8}$$

$$\leq \left(\prod_{l=1}^{p_1}(1 - 2\lambda_l t)^{-n/2}\right)\exp(-tn\sum_{l=1}^{p_1}\lambda_l - tnp_1\epsilon) \tag{F.9}$$

It can be verified that, for $0 < t < \dfrac{1}{4\lambda_l}$, we have

$$(1 - 2\lambda_l t)^{-1/2} \leq \exp(\lambda_l t + 4\lambda_l^2 t^2) \tag{F.10}$$

It follows that

$$\Pr\left(\frac{1}{np_1}\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l(Z_{ls}^i)^2 - E\left(\frac{1}{p_1}\sum_{l=1}^{p_1}\lambda_l(Z_{ls}^i)^2\right) \geq \epsilon\right) \tag{F.11}$$

$$\leq \exp(\sum_{l=1}^{p_1} 4\lambda_l^2 nt^2 - tnp_1\epsilon) \tag{F.12}$$

$$= \exp(-\frac{np_1^2\epsilon^2}{16\sum_{l=1}^{p_1}\lambda_l^2}) \tag{F.13}$$

where the last equality is obtained by letting $t = \frac{p_1\epsilon}{8\sum_{l=1}^{p_1}\lambda_l^2}$.

Similarly, we can show that

$$\Pr\left(\frac{1}{p_1}\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l(Z_{ls}^i)^2 - E\left(\sum_{l=1}^{p_1}\lambda_l(Z_{ls}^i)^2\right) \leq -\epsilon\right) \leq \exp(-\frac{Cnp_1^2\epsilon^2}{\sum_{l=1}^{p_1}\lambda_l^2}).$$

Hence, for any $s$, $\Pr(L(s,s) \geq \epsilon) \leq 2\exp(-\frac{Cnp_1^2\epsilon^2}{\sum_{l=1}^{p_1}\lambda_l^2})$.

Now we consider the case $s \neq k$, where $L(s,k) = |\frac{1}{np_1}\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l Z_{ls}^i Z_{lk}^i|$.

Note that

$$E[\exp(t\lambda_l Z_{ls}^i Z_{lk}^i)] = E\left\{E[\exp(t\lambda_l Z_{ls}^i Z_{lk}^i) \mid Z_{lk}^i]\right\} \tag{F.14}$$

$$= E\left\{\exp\left(\frac{t^2\lambda_l^2(Z_{lk}^i)^2}{2}\right)\right\} = \frac{1}{\sqrt{1-\lambda_l^2 t^2}} \tag{F.15}$$

By the Chernoff bound, we have that

$$\Pr(\frac{1}{np_1}\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l Z_{ls}^i Z_{lk}^i \geq \epsilon) \leq \{\prod_{l=1}^{p_1}(1-\lambda_l^2 t^2)^{-n/2}\}\exp(-tnp_1\epsilon) \tag{F.16}$$

It can be verified that, if $t^2 \leq \frac{1}{2\lambda_l^2}$, then we have

$$(1-\lambda_l^2 t^2)^{-1/2} \leq \exp(2\lambda_l^2 t^2) \tag{F.17}$$

It follows that

$$\Pr(\frac{1}{np_1}\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l Z_{ls}^i Z_{lk}^i \geq \epsilon) \leq \exp(2n\sum_{l=1}^{p_1}\lambda_l^2 t^2 - tnp_1\epsilon) = \exp(-\frac{np_1^2\epsilon^2}{8\sum_{l=1}^{p_1}\lambda_l^2}) \tag{F.18}$$

if we take $t = \frac{p_1\epsilon}{4\sum_{l=1}^{p_1}\lambda_l^2}$.

Similarly, we can show that

$$\Pr(\frac{1}{np_1}\sum_{i=1}^{n}\sum_{l=1}^{p_1}\lambda_l Z_{ls}^i Z_{lk}^i \le -\epsilon) \le \exp(-\frac{Cnp_1^2\epsilon^2}{\sum_{l=1}^{p_1}\lambda_l^2}) \tag{F.19}$$

Consequently, $\Pr(L(s,k) > \epsilon) \le 2\exp(-\frac{Cnp_1^2\epsilon^2}{\sum_{l=1}^{p_1}\lambda_l^2})$ for any $s \neq k$.

Therefore, apply the union bound and we have

$$\Pr(\|\frac{1}{np_1}\sum_{i=1}^{n}(\mathbf{Z}^i)^\intercal\mathbf{\Lambda}\mathbf{Z}^i - \frac{1}{p_1}E((\mathbf{Z}^i)^\intercal\mathbf{\Lambda}\mathbf{Z}^i)\|_{\max} > \epsilon) \le \Pr(\max_{s,k} L(s,k) > \epsilon) \tag{F.20}$$

$$\le \quad 2p_2^2\exp(-\frac{Cnp_1^2\epsilon^2}{\sum_{l=1}^{p_1}\lambda_l^2}) \tag{F.21}$$

□

**Lemma F.3.** *Consider* $\mathbf{W}^1, \ldots, \mathbf{W}^n \in \mathbb{R}^{p_1\times\cdots\times p_M}$ *i.i.d. with the distribution* $\mathbf{TN}(0, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \ldots, \mathbf{\Sigma}_M)$, *where diagonal elements of* $\mathbf{\Sigma}_m$, $m = 1, \ldots, M$, *are all ones. Define*

$$\widehat{\mathbf{\Sigma}}_j = \frac{1}{np_{-j}}\sum_{i=1}^{n}(\mathbf{W}_{(j)}^i)(\mathbf{W}_{(j)}^i)^\intercal \tag{F.22}$$

$$\widehat{\boldsymbol{\mu}}_{(j)} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{W}_{(j)}^i \tag{F.23}$$

*where* $p_{-j} = \prod_{m\neq j} p_m$ *Then there exists* $\epsilon_0 > 0$ *such that for any* $0 < \epsilon < \epsilon_0$, *with probability greater than* $1 - 2p_j^2\exp(-\frac{Cnp_{-j}^2\epsilon^2}{\sum_{l=1}^{p_{-j}}\lambda_l^2})$, *we have*

$$\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|_{\max} \le \|\mathbf{\Sigma}_j^{1/2}\|_1^2\epsilon \tag{F.24}$$

*In the meantime, for* $2/n < \epsilon < \epsilon_0$, *with a proabability greater than* $1-2p^2\exp(-Cn^2p_{-j}\epsilon^2)$, *we have*

$$\|\frac{1}{p_{-j}}\widehat{\boldsymbol{\mu}}_{(j)}\widehat{\boldsymbol{\mu}}_{(j)}^\intercal\|_{\max} \le \lambda_{\max}^{M-1}\|\mathbf{\Sigma}_j^{1/2}\|_1^2\epsilon \tag{F.25}$$

*where* $\lambda_{\max}$ *is the largest eigenvalue of* $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_M$.

**Proof of Lemma F.3.** By definition, we have

$$\mathbf{W}^i = [\![\mathbf{Z}^i; \mathbf{\Sigma}_1^{1/2}, \ldots, \mathbf{\Sigma}_M^{1/2}]\!] \tag{F.26}$$

where $\mathbf{Z}^i \sim TN(0, \mathbf{I}_{p_1}, \ldots, \mathbf{I}_{p_M})$. Further decompose $\otimes_{m\neq j}\mathbf{\Sigma}_m = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^\intercal$, where $\mathbf{\Lambda}$ is a diagonal matrix with diagonal elements $\lambda_l, l = 1, \ldots, p_{-j}$ and $\mathbf{\Gamma}$ is an orthonormal matrix. It is easy to see that $\lambda_l \leq \lambda_{\max}^{M-1}$.

We first show (F.24). Note that

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_j &= \mathbf{\Sigma}_j^{1/2}\left(\frac{1}{np_{-j}}\sum_{i=1}^n(\mathbf{Z}_{(j)}^i)\left(\otimes_{m\neq j}\mathbf{\Sigma}_m\right)(\mathbf{Z}_{(j)}^i)^\intercal\right)\mathbf{\Sigma}_j^{1/2} &\text{(F.27)}\\
&= \mathbf{\Sigma}_j^{1/2}\left(\frac{1}{np_{-j}}\sum_{i=1}^n(\widetilde{\mathbf{Z}}_{(j)}^i)\mathbf{\Lambda}(\widetilde{\mathbf{Z}}_{(j)}^i)^\intercal\right)\mathbf{\Sigma}_j^{1/2} &\text{(F.28)}
\end{aligned}$$

where $\widetilde{\mathbf{Z}}_{(j)}^i = \mathbf{Z}_{(j)}^i\mathbf{\Gamma}$. By properties of matrix normal distribution, we have $\widetilde{\mathbf{Z}}_{(j)}^i \sim \mathbf{TN}(0, \mathbf{I}_{p_j}, \mathbf{I}_{p_{-j}})$, and hence is still a matrix with i.i.d standard normal random variables. Consequently,

$$\|\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j\|_{\max} \leq \|\mathbf{\Sigma}_j^{1/2}\|_1^2 \cdot \|\frac{1}{np_{-j}}\sum_{i=1}^n(\widetilde{\mathbf{Z}}_{(j)}^i)\mathbf{\Lambda}(\widetilde{\mathbf{Z}}_{(j)}^i)^\intercal - E(\widetilde{\mathbf{Z}}_{(j)}^i)\mathbf{\Lambda}(\widetilde{\mathbf{Z}}_{(j)}^i)^\intercal\|_{\max} \tag{F.29}$$

By Lemma F.2, we have

$$\Pr(\|\frac{1}{np_{-j}}\sum_{i=1}^n(\widetilde{\mathbf{Z}}_{(j)}^i)\mathbf{\Lambda}(\widetilde{\mathbf{Z}}_{(j)}^i)^\intercal - \frac{1}{p_{-j}}E(\widetilde{\mathbf{Z}}_{(j)}^i)\mathbf{\Lambda}(\widetilde{\mathbf{Z}}_{(j)}^i)^\intercal\|_{\max} > \epsilon) \leq 2p_j^2\exp(-\frac{Cnp_{-j}^2\epsilon^2}{\sum_{l=1}^{p_j}\lambda_l^2})$$

And the conclusion follows.

Now we show (F.25). Note that

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_{(j)}\widehat{\boldsymbol{\mu}}_{(j)}^\intercal &= \mathbf{\Sigma}_j^{1/2}(\frac{1}{n}\sum_{i=1}^n\widetilde{\mathbf{Z}}_{(j)}^i)(\otimes_{m\neq j}\mathbf{\Sigma}_m)(\frac{1}{n}\sum_{i=1}^n\widetilde{\mathbf{Z}}_{(j)}^i)^\intercal\mathbf{\Sigma}_j^{1/2} &\text{(F.30)}\\
&= \mathbf{\Sigma}_j^{1/2}(\frac{1}{n}\sum_{i=1}^n\widetilde{\mathbf{Z}}_{(j)}^i)\mathbf{\Lambda}(\frac{1}{n}\sum_{i=1}^n\widetilde{\mathbf{Z}}^i)^\intercal\mathbf{\Sigma}_j^{1/2} &\text{(F.31)}\\
&= \mathbf{\Sigma}_j^{1/2}\widetilde{\boldsymbol{\mu}}_{(j)}\mathbf{\Lambda}\widetilde{\boldsymbol{\mu}}_{(j)}^\intercal\mathbf{\Sigma}_j^{1/2} &\text{(F.32)}
\end{aligned}$$

where $\widetilde{\boldsymbol{\mu}}_{(j)} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{Z}}^i_{(j)}$. Therefore,

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}_{(j)}\widehat{\boldsymbol{\mu}}_{(j)}^{\intercal}\|_{\max} &\leq \|\boldsymbol{\Sigma}_j^{1/2}\|_1^2\|\widetilde{\boldsymbol{\mu}}_{(j)}\boldsymbol{\Lambda}\widetilde{\boldsymbol{\mu}}_{(j)}^{\intercal}\|_{\max} = \|\boldsymbol{\Sigma}_1^{1/2}\|_1^2\max_{s,k}|\sum_{l=1}^{p_{-j}}\lambda_l\widetilde{\mu}_{(j),sl}\widetilde{\mu}_{(j),kl}| && \text{(F.33)}\\
&\leq \|\boldsymbol{\Sigma}_j^{1/2}\|_1^2\max_{s,k}\frac{1}{2}(\sum_{l=1}^{p_{-j}}\lambda_l\widetilde{\mu}^2_{(j),sl}+\sum_{l=1}^{p_{-j}}\lambda_l\widetilde{\mu}^2_{(j),kl}) && \text{(F.34)}\\
&\leq \|\boldsymbol{\Sigma}_j^{1/2}\|_1^2\max_{s}\sum_{l=1}^{p_{-j}}\lambda_l\widetilde{\mu}^2_{(j),sl} \leq \lambda_{\max}^{M-1}\|\boldsymbol{\Sigma}_j^{1/2}\|_1^2\max_{s}\sum_{l=1}^{p_{-j}}\widetilde{\mu}^2_{(j),sl} && \text{(F.35)}
\end{aligned}$$

For the term $\sum_{l=1}^{p_{-j}}\widetilde{\mu}^2_{(j),sl}$, note that $n\sum_{l=1}^{p_{-j}}\widetilde{\mu}^2_{(j),sl}\sim\chi^2(p_{-j})$. Then by the Chernoff bound, there exists $\epsilon_0>0$ such that for $2/n<\epsilon<\epsilon_0$, we have

$$\begin{aligned}
&\Pr(\frac{1}{p_{-j}}\sum_{l=1}^{p_{-j}}\widetilde{\mu}^2_{(j),sl}>\epsilon) = \Pr(n\sum_{l=1}^{p_{-j}}\widetilde{\mu}^2_{(j),sl}-p_{-j}>np_{-j}\epsilon-p_{-j}) && \text{(F.36)}\\
\leq\ &\Pr(n\sum_{l=1}^{p_{-j}}\widetilde{\mu}^2_{(j),sl}-p_{-j}>np_{-j}\epsilon/2)\leq\exp(-Cn^2p_{-j}\epsilon^2) && \text{(F.37)}
\end{aligned}$$

And the conclusion follows. □

**Lemma F.4.** *For matrices $\boldsymbol{\Sigma}_m,\widehat{\boldsymbol{\Sigma}}_m\in\mathbb{R}^{p_m}$, define $\sigma_{\max}=\max_{m,j,k}|\sigma_{m,j,k}|$. If $\|\widehat{\boldsymbol{\Sigma}}_m-\boldsymbol{\Sigma}_m\|_{\max}\leq\epsilon\leq\max_{m,j,k}|\sigma_{m,jk}|$ for $m=1,\ldots,M$, then we have*

$$\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_{\max}\leq M2^M\sigma_{\max}^{M-1}\epsilon \tag{F.38}$$

*where $\widehat{\boldsymbol{\Sigma}}=\otimes_{m=M}^1\widehat{\boldsymbol{\Sigma}}_m$ and $\boldsymbol{\Sigma}=\otimes_{m=M}^1\boldsymbol{\Sigma}_m$.*

**Proof of Lemma F.4.**

$$\begin{aligned}
&\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_{\max} = \max_{j_m,k_m,m=1,\ldots,M}|\prod_{m=1}^{M}\widehat{\sigma}_{m,j_m,k_m}-\prod_{m=1}^{M}\sigma_{m,j_m,k_m}| && \text{(F.39)}\\
=\ &\max_{j_m,k_m,m=1,\ldots,M}|\sum_{m=1}^{M}(\prod_{l=1}^{m-1}\sigma_{l,j_lk_l})(\widehat{\sigma}_{m,j_mk_m}-\sigma_{m,j_mk_m})(\prod_{l=m+1}^{M}\widehat{\sigma}_{l,j_lk_l})| && \text{(F.40)}\\
\leq\ &\max_{j_m,k_m,m=1,\ldots,M}\sum_{m=1}^{M}|\prod_{l\neq m}\max\{|\widehat{\sigma}_{l,j_l,k_l}|,|\sigma_{l,j_l,k_l}|\}(\widehat{\sigma}_{m,j_m,k_m}-\sigma_{m,j_m,k_m})| && \text{(F.41)}\\
\leq\ &M2^M\sigma_{\max}^{M-1}\epsilon && \text{(F.42)}
\end{aligned}$$

□

**Lemma F.5.** *Consider random variables $(W, \mathbf{V})$ where $W \in \mathbb{R}$, $\mathbf{V} \in \mathbb{R}^q$ and $\mathbf{V} \sim N(0, \mathbf{\Omega})$, $W \mid \mathbf{V} = \mathbf{v} \sim N(\boldsymbol{\alpha}^\intercal \mathbf{v}, \sigma^2)$. Given $n$ independent samples $\{W^i, \mathbf{V}^i\}_{i=1}^n$, define*

$$\widehat{\boldsymbol{\alpha}} = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{V}^i (\mathbf{V}^i)^\intercal \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{V}^i W^i \right) \tag{F.43}$$

*Then*

$$\Pr\left( \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{\max} > 2\|\mathbf{\Omega}^{-1}\|_1 (1 + \|\mathbf{\Omega}^{-1}\|_1 \cdot \|\boldsymbol{\delta}\|_\infty)\epsilon \right) < 2q^2 \exp(-\frac{Cn\epsilon^2}{q^2}) + 2q \exp(-2Cn\epsilon^2) \tag{F.44}$$

*where $\boldsymbol{\delta} = E(\mathbf{V}W)$.*

**Proof Lemma F.5.** Write $\widehat{\mathbf{\Omega}} = \frac{1}{n} \sum_{i=1}^n \mathbf{V}^i (\mathbf{V}^i)^\intercal$ and $\widehat{\boldsymbol{\delta}} = \frac{1}{n} \sum_{i=1}^n \mathbf{V}^i W^i$. By Lemma F.1, for any $(j, k)$, we have

$$\Pr(|\widehat{\omega}_{jk} - \omega_{jk}| > \epsilon) \leq 2 \exp(-Cn\epsilon^2) \tag{F.45}$$

$$\Pr(|\widehat{\delta}_k - \delta_k| > \epsilon) \leq 2 \exp(-Cn\epsilon^2) \tag{F.46}$$

Consequently,

$$\Pr(\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_1 > \epsilon) \leq 2q^2 \exp(-\frac{Cn\epsilon^2}{q^2}) \tag{F.47}$$

$$\Pr(\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_\infty > \epsilon) \leq 2q \exp(-Cn\epsilon^2) \tag{F.48}$$

It follows that

$$\begin{aligned} \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_\infty &= \|\widehat{\mathbf{\Omega}}^{-1}\widehat{\boldsymbol{\delta}} - \mathbf{\Omega}^{-1}\boldsymbol{\delta}\|_\infty = \|\widehat{\mathbf{\Omega}}^{-1}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + (\widehat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1})\boldsymbol{\delta}\|_\infty && \text{(F.49)} \\ &\leq \|\widehat{\mathbf{\Omega}}^{-1}\|_1 \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_\infty + \|\widehat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}\|_1 \cdot \|\boldsymbol{\delta}\|_\infty && \text{(F.50)} \\ &\leq (\|\widehat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}\|_1 + \|\mathbf{\Omega}^{-1}\|_1)\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_\infty + \|\widehat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}\|_1 \|\boldsymbol{\delta}\|_\infty && \text{(F.51)} \end{aligned}$$

Now note that

$$\begin{aligned} \|\widehat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}\|_1 &= \|\widehat{\mathbf{\Omega}}^{-1}(\mathbf{\Omega} - \widehat{\mathbf{\Omega}})\mathbf{\Omega}^{-1}\|_1 \leq \|\widehat{\mathbf{\Omega}}^{-1}\|_1 \cdot \|\mathbf{\Omega} - \widehat{\mathbf{\Omega}}\|_1 \cdot \|\mathbf{\Omega}^{-1}\|_1 && \text{(F.52)} \\ &\leq (\|\widehat{\mathbf{\Omega}}^{-1} - \mathbf{\Omega}^{-1}\|_1 + \|\mathbf{\Omega}^{-1}\|_1)\|\mathbf{\Omega} - \widehat{\mathbf{\Omega}}\|_1 \cdot \|\mathbf{\Omega}^{-1}\|_1 && \text{(F.53)} \end{aligned}$$

It follows that if $\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 < \epsilon$ for $\epsilon < \dfrac{1}{2\|\boldsymbol{\Omega}^{-1}\|_1}$, we must have

$$\|\widehat{\boldsymbol{\Omega}}^{-1} - \boldsymbol{\Omega}^{-1}\|_1 \leq 2\|\boldsymbol{\Omega}^{-1}\|_1^2 \epsilon \tag{F.54}$$

Hence, under the event $\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 < \epsilon, \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_\infty < \epsilon$, we have

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{\max} \leq 2\|\boldsymbol{\Omega}^{-1}\|_1(1 + \|\boldsymbol{\Omega}^{-1}\|_1 \cdot \|\boldsymbol{\delta}\|_\infty)\epsilon. \tag{F.55}$$

□

**Lemma F.6.** *Consider* $(\mathbf{V}, \mathbf{W})$ *where* $\mathbf{V} \in \mathbb{R}^q, \mathbf{W} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$. *Assume that* $\mathbf{V} \sim N(0, \boldsymbol{\Omega})$, $\mathbf{W} \mid \mathbf{V} = \mathbf{v} \sim TN(\boldsymbol{\alpha} \bar{\times}_{(M+1)} \mathbf{v}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$. *Given* $n$ *independent samples* $\{\mathbf{V}^i, \mathbf{W}^i\}_{i=1}^n$, *define*

$$\widehat{\boldsymbol{\alpha}}_{j_1 \cdots j_M} = (\frac{1}{n}\sum_{i=1}^n \mathbf{V}^i(\mathbf{V}^i)^\intercal)^{-1}(\frac{1}{n}\sum_{i=1}^n \mathbf{V}^i W^i_{j_1 \cdots j_M}) \tag{F.56}$$

$$\widehat{\boldsymbol{\Sigma}}_j(\widehat{\boldsymbol{\alpha}}) = \frac{1}{np_{-j}}\sum_{i=1}^n (\mathbf{W}^i - \widehat{\boldsymbol{\alpha}} \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)} (\mathbf{W}^i - \widehat{\boldsymbol{\alpha}} \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)}^\intercal \tag{F.57}$$

$$\widehat{\boldsymbol{\Sigma}}_j(\boldsymbol{\alpha}) = \frac{1}{np_{-j}}\sum_{i=1}^n (\mathbf{W}^i - \boldsymbol{\alpha} \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)} (\mathbf{W}^i - \boldsymbol{\alpha} \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)}^\intercal \tag{F.58}$$

*Then we have*

$$\Pr(\|\widehat{\boldsymbol{\Sigma}}_j(\widehat{\boldsymbol{\alpha}}) - \widehat{\boldsymbol{\Sigma}}_j(\boldsymbol{\alpha})\|_{\max} > q^2\epsilon) \leq Cpq^2 \exp(-Cn\frac{\epsilon}{q^2}) \tag{F.59}$$

**Proof of Lemma F.6.** Define events:

$$\mathcal{E}_1 = \{\|\widehat{\boldsymbol{\alpha}}_{j_1 \cdots j_M} - \boldsymbol{\alpha}_{j_1 \cdots j_M}\|_\infty \leq \epsilon \text{ for all } \{j_1, \ldots, j_M\}\} \tag{F.60}$$

$$\mathcal{E}_2 = \{\max_{s,l} \frac{1}{n}\sum_{i=1}^n (W^i_{(j),sl})^2 \leq 1 + \epsilon\} \tag{F.61}$$

$$\mathcal{E}_3 = \{\max_{l} \frac{1}{n}\sum_{i=1}^n (V^i_l)^2 \leq 1 + \epsilon\} \tag{F.62}$$

By Lemma F.5, $\Pr(\mathcal{E}_1) \geq 1 - 2p(q^2 \exp(-Cn\frac{\epsilon^2}{q^2}) + q\exp(-2Cn\epsilon^2))$. By Lemma F.1, we have $\Pr(\mathcal{E}_2) \geq 1 - p\exp(-Cn\epsilon^2), \Pr(\mathcal{E}_3) \geq 1 - q\exp(-Cn\epsilon^2)$. Hence,

$$\Pr(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - 2Cpq^2 \exp(-Cn\frac{\epsilon^2}{q^2}) \tag{F.63}$$

In what follows we assume that $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ has happened.

Straightforward calculation shows that

$$
\begin{aligned}
& \widehat{\boldsymbol{\Sigma}}_j(\widehat{\boldsymbol{\alpha}}) - \widehat{\boldsymbol{\Sigma}}_j(\boldsymbol{\alpha}) \\
= \; & \frac{1}{np_{-j}} \sum_{i=1}^{n} \mathbf{W}^i_{(j)} ((\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{V}^i)^{\intercal}_{(j)} \\
& + \frac{1}{np_{-j}} \sum_{i=1}^{n} ((\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)} (\mathbf{W}^i_{(j)})^{\intercal} \\
& + \left( \frac{1}{np_{-j}} \sum_{i=1}^{n} (\widehat{\boldsymbol{\alpha}} \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)} (\widehat{\boldsymbol{\alpha}} \bar{\times}_{(M+1)} \mathbf{V}^i)^{\intercal}_{(j)} - \frac{1}{np_{-j}} \sum_{i=1}^{n} (\boldsymbol{\alpha} \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)} (\boldsymbol{\alpha} \bar{\times}_{(M+1)} \mathbf{V}^i)^{\intercal}_{(j)} \right) \\
\equiv \; & L_1 + L_2 + L_3
\end{aligned}
$$

We analyze these three terms one by one.

$$
\begin{aligned}
\|L_1\|_{\max} &= \max_{s,k} |\frac{1}{np_{-j}} \sum_{i=1}^{n} \sum_{l=1}^{p_{-j}} W^i_{(j),sl} [(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{V}^i]_{(j),kl}| \\
&\leq \max_{k_1,\ldots,k_M} \|\boldsymbol{\alpha}_{k_1 \cdots k_M} - \widehat{\boldsymbol{\alpha}}_{k_1 \cdots k_M}\|_\infty \max_{s,l} \frac{1}{np_{-j}} \sum_{i=1}^{n} \sum_{l=1}^{p_{-j}} |W^i_{(j),sl}| \cdot \|\mathbf{V}^i\|_1 \\
&\leq \max_{k_1,\ldots,k_M} \|\boldsymbol{\alpha}_{k_1 \cdots k_M} - \widehat{\boldsymbol{\alpha}}_{k_1 \cdots k_M}\|_\infty \max_{s,l} \frac{1}{np_{-j}} \sum_{i=1}^{n} \sum_{l=1}^{p_{-j}} |W^i_{(j),sl}| \cdot \sqrt{q} \|\mathbf{V}^i\|_2 \\
&\leq \max_{k_1,\ldots,k_M} \|\boldsymbol{\alpha}_{k_1 \cdots k_M} - \widehat{\boldsymbol{\alpha}}_{k_1 \cdots k_M}\|_\infty \max_{s,l} \frac{1}{np_{-j}} \sum_{i=1}^{n} \sum_{l=1}^{p_{-j}} \{(W^i_{(j),sl})^2/2 + q\|\mathbf{V}^i\|_2^2\} \\
&\leq \epsilon \left( \max_{s,l} \frac{1}{n} \sum_{i=1}^{n} (W^i_{(j),sl})^2 + \max_{l} \frac{q^2}{n} \sum_{i=1}^{n} (V^i_l)^2 \right) \\
&\leq Cq^2\epsilon
\end{aligned}
$$

Hence, $\|L_1\|_{\max} \leq Cq^2\epsilon$. Meanwhile, $L_2 = L_1^{\intercal}$. Hence, $\|L_2\|_{\max} = \|L_1\|_{\max} \leq Cq^2\epsilon$.

For $L_3$, note that

$$
\begin{aligned}
L_3 &= \frac{1}{np_{-j}} \sum_{i=1}^{n} ((\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)} (\widehat{\boldsymbol{\alpha}} \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)}^{\intercal} && \text{(F.64)} \\
&\quad + \frac{1}{np_{-j}} \sum_{i=1}^{n} (\boldsymbol{\alpha} \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)} ((\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j)}^{\intercal} && \text{(F.65)} \\
&\equiv L_4 + L_5 && \text{(F.66)}
\end{aligned}
$$

$$
\begin{aligned}
\|L_4\|_{\max} &= \max_{k,s} \frac{1}{np_{-j}} | \sum_{i=1}^{n} \sum_{l=1}^{p_{-j}} ((\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \bar{\times}_{(M+1)} \mathbf{V}^i)_{(j),kl} \cdot \widehat{\boldsymbol{\alpha}}_{(j),sl}^{\intercal} \mathbf{V}^i | && \text{(F.67)} \\
&\leq \max_{k_1,\ldots,k_M} \|\widehat{\boldsymbol{\alpha}}_{k_1,\ldots,k_M} - \boldsymbol{\alpha}_{k_1,\ldots,k_M}\|_\infty \cdot (\max_{k_1,\ldots,k_M} \|\boldsymbol{\alpha}_{k_1,\ldots,k_M}\|_\infty + \epsilon) && \text{(F.68)} \\
&\quad \cdot \frac{1}{np_{-j}} \sum_{i=1}^{n} \sum_{l=1}^{p_{-j}} \|\mathbf{V}^i\|_1^2 && \text{(F.69)} \\
&\leq \epsilon (\max_{k_1,\ldots,k_M} \|\boldsymbol{\alpha}_{k_1,\ldots,k_M}\|_\infty + \epsilon) \max_l \frac{q^2}{n} \sum_{i=1}^{n} (V_l^i)^2 && \text{(F.70)} \\
&\leq 2q^2\epsilon (\max_{k_1,\ldots,k_M} \|\boldsymbol{\alpha}_{k_1,\ldots,k_M}\|_\infty + \epsilon) && \text{(F.71)} \\
&\leq Cq^2\epsilon && \text{(F.72)}
\end{aligned}
$$

Similarly we can provide a bound for $\|L_5\|_{\max}$. Putting all the pieces together, we have the desired conclusion. □

# G Proofs for Theorems 1 and 2

Now we consider the CATCH model in (2.1, 2.2). Throughout this section, we assume that $n_k > \frac{\pi_k n}{2}$, where $n_k = \sum_{i=1}^{n} 1(Y^i = k)$. Note that this event happens with a probability greater than $1 - 2K\exp(-Cn\pi_{\min}^2)$, as

$$
\begin{aligned}
&\Pr(n_k < \frac{\pi_k n}{2}, k = 1, \ldots, K) \leq \sum_{k=1}^{K} \Pr(|\frac{n_k}{n} - \pi_k| > \frac{\pi_k}{2}) && \text{(G.1)} \\
\leq\ & 2\sum_{k=1}^{K} \exp(-Cn\pi_k^2) \leq 2K\exp(-Cn\pi_{\min}^2) && \text{(G.2)}
\end{aligned}
$$

where we apply Hoeffding's inequality.

Define $\widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{Y^i=k}(\mathbf{X}^i - \widehat{\boldsymbol{\alpha}}\bar{\times}_{(M+1)}\mathbf{U}^i)$, $\widehat{\boldsymbol{\Sigma}}_j^{(k)} = \frac{1}{n_k p_{-j}}\sum_{Y^i=k}(\mathbf{X}^i - \widehat{\boldsymbol{\alpha}}\bar{\times}_{(M+1)}\mathbf{U}^i - \widehat{\boldsymbol{\mu}}_k)_{(j)}(\mathbf{X}^i - \widehat{\boldsymbol{\alpha}}\bar{\times}_{(M+1)}\mathbf{U}^i - \widehat{\boldsymbol{\mu}}_k)_{(j)}^\intercal$. We have the following lemma.

**Lemma G.1.** *For* $0 < \epsilon < \min\{\epsilon_0, \frac{\pi_{\min}}{2}\}$*, we have*

$$\Pr(\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{\max} > \epsilon) \leq 2q^2 p \exp(-\frac{Cn\epsilon^2}{q^2}) + 2K\exp(-Cn\pi_{\min}^2) \tag{G.3}$$

**Proof of Lemma G.1.** The proof is similar to that of Lemma F.5 and is hence omitted here. □

**Lemma G.2.** *With a probability greater than* $1 - Cq^2 p\exp(-C\frac{n\epsilon^2}{q^2}) - 2K\exp(-Cn\pi_{\min}^2)$*, we have*

$$\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_{\max} \leq (2 + \|\boldsymbol{\phi}_k\|_\infty)q\epsilon \quad \textit{for } k = 1, \ldots, K \tag{G.4}$$

**Proof of Lemma G.2.** Note that

$$\begin{aligned}
& \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_{\max} && \text{(G.5)}\\
\leq\ & \|\frac{1}{n_k}\sum_{Y^i=k}(\mathbf{X}^i - \widehat{\boldsymbol{\alpha}}\bar{\times}_{(M+1)}\mathbf{U}^i) - \frac{1}{n_k}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\alpha}\bar{\times}_{(M+1)}\mathbf{U}^i)\|_{\max} && \text{(G.6)}\\
& + \|\frac{1}{n_k}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\alpha}\bar{\times}_{(M+1)}\mathbf{U}^i) - \boldsymbol{\mu}_k\|_{\max} && \text{(G.7)}\\
\leq\ & \|(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\bar{\times}_{(M+1)}\boldsymbol{\phi}_k\|_{\max} + \|(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\bar{\times}_{(M+1)}(\boldsymbol{\phi}_k - \widehat{\boldsymbol{\phi}}_k)\|_{\max} && \text{(G.8)}\\
& + \|\frac{1}{n_k}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\alpha}\bar{\times}_{(M+1)}\mathbf{U}^i) - \boldsymbol{\mu}_k\|_{\max} && \text{(G.9)}\\
\leq\ & (2 + \|\boldsymbol{\phi}_k\|_\infty)q\epsilon && \text{(G.10)}
\end{aligned}$$

under the event $n_k > \frac{\pi_k n}{2}$, $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{\max} \leq \epsilon$, $\|\widehat{\boldsymbol{\phi}}_k - \boldsymbol{\phi}_k\|_{\max} \leq \epsilon$, $\|\frac{1}{n_k}\sum_{Y^i=k}(\mathbf{X}^i - \boldsymbol{\alpha}\bar{\times}_{(M+1)}\mathbf{U}^i) - \boldsymbol{\mu}_k\|_{\max} \leq \epsilon$. Combine Lemma G.1 with Lemma F.1 and we have the desired conclusion.

□

**Lemma G.3.** *There exists* $\epsilon_0 > 0$ *such that for* $\frac{4}{\pi_k n} < \epsilon < \epsilon_0$*, we have*

$$\Pr(\|\widehat{\boldsymbol{\Sigma}}_j^{(k)} - \boldsymbol{\Sigma}_j\|_{\max} > Cq^2\epsilon) \leq 2p_j^2\exp(-Cnp_{-j}\epsilon^2) + Cq^2 p\exp(-\frac{Cn\epsilon^2}{q^2}) + 2K\exp(-Cn\pi_{\min}^2) \tag{G.11}$$

**Proof of Lemma G.3.** For simplicity, write

$$\widetilde{\mathbf{E}}^i = \mathbf{X}^i - \boldsymbol{\alpha} \bar{\times}_{(M+1)} \mathbf{U}^i - \boldsymbol{\mu}_k \tag{G.12}$$

Throughout this proof, we assume that the following events have happened:

$$\begin{aligned}
\mathcal{E}_4 &= \{\|\frac{1}{np_{-j}} \sum_{Y^i=k} \widetilde{\mathbf{E}}^i_{(j)} (\widetilde{\mathbf{E}}^i_{(j)})^\intercal - \boldsymbol{\Sigma}_j\|_{\max} < \epsilon\} && \text{(G.13)} \\
\mathcal{E}_5 &= \{\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_{\max} < \epsilon\} && \text{(G.14)} \\
\mathcal{E}_6 &= \{\frac{1}{n_k} \sum_{Y^i=k} (\widetilde{\mathbf{E}}^i_{(j),lt})^2 < 1+\epsilon, \text{ for any } l,t\} && \text{(G.15)} \\
\mathcal{E}_7 &= \{\frac{1}{n_k} \sum_{Y^i=k} \|\mathbf{U}^i\|_2^2 \le (1+\epsilon)q\} && \text{(G.16)} \\
\mathcal{E}_8 &= \{\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_{\max} < \epsilon, \text{ for any } k\} && \text{(G.17)} \\
\mathcal{E}_9 &= \{\|\widehat{\boldsymbol{\phi}}_k - \boldsymbol{\phi}_k\|_{\max} < \epsilon, \text{ for any } k\} && \text{(G.18)}
\end{aligned}$$

When $n_k > \frac{\pi_k n}{2}$, by Lemma F.3, $\Pr(\mathcal{E}_4^C) \le 2p_j^2 \exp(-Cnp_{-j}\epsilon^2)$. By Lemma G.1, $\Pr(\mathcal{E}_5^C) \le 2q^2 p \exp(-\frac{Cn\epsilon^2}{q^2})$. By standard Chernoff bound, $\Pr(\mathcal{E}_6^C) \le 2p\exp(-Cn\epsilon^2)$, $\Pr(\mathcal{E}_7^C) \le 2q\exp(-Cn\epsilon^2)$, $\Pr(\mathcal{E}_8^C) \le 2p\exp(-Cn\epsilon^2)$, $\Pr(\mathcal{E}_9^C) \le 2q\exp(-Cn\epsilon^2)$. Hence, all these events happen simultaneously with a probability greater than

$$1 - 2p_j^2 \exp(-Cnp_{-j}\epsilon^2) - 2q^2 p \exp(-\frac{Cn\epsilon^2}{q^2}) - Cp\exp(-Cn\epsilon^2) - 2K\exp(-Cn\pi_{\min}^2)$$

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_j^{(k)} &= \frac{1}{n_k p_{-j}} \sum_{Y^i=k} (\mathbf{X}^i - \widehat{\boldsymbol{\alpha}} \bar{\times}_{(M+1)} \mathbf{U}^i - \widehat{\boldsymbol{\mu}}_k)_{(j)} (\mathbf{X}^i - \widehat{\boldsymbol{\alpha}} \bar{\times}_{(M+1)} \mathbf{U}^i - \widehat{\boldsymbol{\mu}}_k)_{(j)}^\intercal \\
&= \frac{1}{n_k p_{-j}} \sum_{Y^i=k} \widetilde{\mathbf{E}}^i_{(j)} (\widetilde{\mathbf{E}}^i_{(j)})^\intercal + \frac{2}{n_k p_{-j}} \sum_{Y^i=k} \widetilde{\mathbf{E}}^i_{(j)} \left((\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i + (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)\right)_{(j)}^\intercal \\
&\quad + \frac{1}{n_k p_{-j}} \sum_{Y^i=k} \left((\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i + (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)\right)_{(j)} \left((\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i + (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)\right)_{(j)}^\intercal \\
&\equiv L_5 + L_6 + L_7
\end{aligned}$$

Under events $\mathcal{E}_4$–$\mathcal{E}_9$, we have, for $L_5$,

$$\|L_5 - \mathbf{\Sigma}_j\|_{\max} \leq \epsilon. \tag{G.19}$$

For $L_6$,

$$\begin{aligned} \|L_6\|_{\max} &\leq \|\frac{1}{n_k p_{-j}} \sum_{Y^i=k} \widetilde{\mathbf{E}}^i_{(j)} [(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i]_{(j)} \|_{\max} && \text{(G.20)} \\ &\quad + \|\frac{1}{n_k p_{-j}} \sum_{Y^i=k} \widetilde{\mathbf{E}}^i_{(j)} (\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k)^{\intercal}_{(j)}] \|_{\max} && \text{(G.21)} \end{aligned}$$

We analyze these two terms separately. First,

$$\begin{aligned} & \|\frac{1}{n_k p_{-j}} \sum_{Y^i=k} \widetilde{\mathbf{E}}^i_{(j)} [(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i]_{(j)} \|_{\max} && \text{(G.22)} \\ = \; & \max_{l,s} |\frac{1}{n_k p_{-j}} \sum_{Y^i=k} \sum_{t=1} \widetilde{E}^i_{(j),lt} [(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i]_{(j),st}| && \text{(G.23)} \\ \leq \; & \|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_{\max} \max_{l,s} \frac{1}{n_k p_{-j}} \sum_{Y^i=k} \sum_{t=1} |\widetilde{E}^i_{(j),lt}| \cdot \|\mathbf{U}^i]_{(j),st}\|_1 && \text{(G.24)} \\ \leq \; & \|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_{\max} \cdot \max_{l,s} \left\{ \frac{1}{n_k p_{-j}} \sum_{Y^i=k} \sum_{t=1}^{p_{-j}} (\widetilde{E}^i_{(j),lt})^2 + \frac{q}{n_k p_{-j}} \sum_{Y^i=k} \sum_{t=1}^{p_{-j}} \|\mathbf{U}^i\|_2^2 \right\} && \text{(G.25)} \\ \leq \; & \|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_{\max} \left\{ \max_{l,t} \frac{1}{n_k} \sum_{Y^i=k} (\widetilde{E}^i_{(j),lt})^2 + \frac{q}{n_k} \sum_{Y^i=k} \|\mathbf{U}^i\|_2^2 \right\} && \text{(G.26)} \\ \leq \; & 2(1+q^2)\epsilon && \text{(G.27)} \end{aligned}$$

Second,

$$\|\frac{1}{n_k p_{-j}} \sum_{Y^i=k} \widetilde{\mathbf{E}}^i_{(j)} (\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k)^\intercal_{(j)}\|_{\max} \tag{G.28}$$

$$= \max_{l,s} |\frac{1}{n_k p_{-j}} \sum_{Y^i=k} \sum_{t=1}^{p_{-j}} \widetilde{E}^i_{(j),lt} (\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k)_{(j),st}| \tag{G.29}$$

$$\leq \|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k\|_{\max} \cdot \max_{l} \frac{1}{n_k p_{-j}} \sum_{Y^i=k} \sum_{t=1}^{p_{-j}} |\widetilde{E}^i_{(j),lt}| \tag{G.30}$$

$$\leq \|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k\|_{\max} \cdot \max_{l,t} \frac{1}{n_k - 1} \sum_{Y^i=k} |\widetilde{E}^i_{(j),lt}| \tag{G.31}$$

$$\leq \|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k\|_{\max} \cdot \max_{l,t} \frac{1}{n_k - 1} \sqrt{\sum_{Y^i=k} (\widetilde{E}^i_{(j),lt})^2 \sum_{Y^i=k} 1^2} \tag{G.32}$$

$$\leq \|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k\|_{\max} \max_{l,t} \sqrt{\frac{2}{n_k} \sum_{Y^i=k} (\widetilde{E}^i_{(j),lt})^2} \leq 2\epsilon \tag{G.33}$$

Therefore,

$$\|L_6\|_{\max} \leq 2(2 + q^2)\epsilon \tag{G.34}$$

For $L_7$, we further decompose it to be

$$L_7 = \frac{1}{n_k p_{-j}} \sum_{Y^i=k} \left((\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i\right)_{(j)} \left((\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i\right)^\intercal_{(j)} \tag{G.35}$$

$$+ \frac{2}{n_k p_{-j}} \sum_{Y^i=k} \left((\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i\right)_{(j)} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)^\intercal_{(j)} \tag{G.36}$$

$$+ \frac{1}{p_{-j}} \sum_{Y^i=k} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)_{(j)} (\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)^\intercal_{(j)} \tag{G.37}$$

$$\equiv L_8 + L_9 + L_{10} \tag{G.38}$$

$$\|L_8\|_{\max} = \max_{l,s} \frac{1}{n_k p_{-j}} \sum_{Y^i=k} \sum_{t=1}^{p_{-j}} [(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i]_{(j),lt} [(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \mathbf{U}^i]_{(j),st}$$

$$\leq \frac{\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|^2_{\max}}{n_k p_{-j}} \sum_{Y^i=k} \sum_{t=1}^{p_{-j}} \|\mathbf{U}^i\|_1^2 \leq 2q^2\epsilon^2$$

$$
\begin{aligned}
\|L_9\|_{\max} &= \max_{l,s} \frac{4}{p_{-j}} |\sum_{t=1}^{p_{-j}} [(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \bar{\times}_{(M+1)} \widehat{\boldsymbol{\phi}}_k]_{(j),lt} (\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k)_{(j),st}| && \text{(G.39)} \\
&\leq \frac{4\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_{\max}}{p_{-j}} \sum_{t=1}^{p_{-j}} \|\widehat{\boldsymbol{\phi}}_k\|_1 \cdot |(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)_{(j),st}| && \text{(G.40)} \\
&\leq 4\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}\|_{\max} \cdot (\|\boldsymbol{\phi}_k\|_1 + q\epsilon) \cdot \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_{\max} && \text{(G.41)} \\
&\leq 4\epsilon^2 (\|\boldsymbol{\phi}_k\|_1 + q\epsilon) && \text{(G.42)}
\end{aligned}
$$

By Lemma F.3, we have that, with a probability greater than $1 - 2p^2 \exp(-Cn^2 p_{-j} \epsilon^2)$, $\|L_{10}\|_{\max} \leq \epsilon$.

Putting all the results together, we have the desired conclusion.

□

**Lemma G.4.** *There exists $\epsilon_0 > 0$ such that for $\frac{4}{\pi_k n} < \epsilon < \epsilon_0$, we have*

$$
\Pr(\|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|_{\max} > Cq^2\epsilon) \leq 2Kp_j^2 \exp(-Cnp_{-j}\epsilon^2) + CKq^2 p \exp(-\frac{Cn\epsilon^2}{q^2}) + 2K \exp(-Cn\pi_{\min}^2) \tag{G.43}
$$

**Proof of Lemma G.4.** Because $\widehat{\boldsymbol{\Sigma}}_j = \sum_{k=1}^K \frac{n_k - 1}{n - K} \boldsymbol{\Sigma}_j^{(k)}$, we have

$$
\|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|_{\max} = \|\sum_{k=1}^K \frac{n_k}{n} (\widehat{\boldsymbol{\Sigma}}_j^{(k)} - \boldsymbol{\Sigma}_j)\|_{\max} \leq \left( \sum_{k=1}^K \frac{n_k}{n} \right) \max_k \|\widehat{\boldsymbol{\Sigma}}_j^{(k)} - \boldsymbol{\Sigma}_j\|_{\max} \leq q^2 \epsilon
$$

provided that $\max_k \|\widehat{\boldsymbol{\Sigma}}_j^{(k)} - \boldsymbol{\Sigma}_j\|_{\max} \leq q^2\epsilon$. By Lemma G.3 we have the desired conclusion. □

**Lemma G.5.** *There exists $\epsilon_0 > 0$ such that for $\frac{4}{\pi_k n} < \epsilon < \epsilon_0$, we have*

$$
\Pr(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} > \epsilon) \leq 2K \sum_{m=1}^M p_m^2 \exp(-Cnp_{-m}\epsilon^2) + CMKq^2 p \exp(-\frac{Cn\epsilon^2}{q^2}) + 2KM \exp(-Cn\pi_{\min}^2) \tag{G.44}
$$

**Proof of Lemma G.5.** Combine Lemma G.4 with Lemma F.4 and we have the desired conclusion.

□

To prove Theorem 2, we define the oracle estimator

$$(\hat{\boldsymbol{\beta}}_{2,\mathcal{D}}^{(\text{oracle})}, \cdots, \hat{\boldsymbol{\beta}}_{K,\mathcal{D}}^{(\text{oracle})}) = \arg \min_{\boldsymbol{\beta}_{2,\mathcal{D}}, \cdots, \boldsymbol{\beta}_{K,\mathcal{D}}} \sum_{k=2}^{K} \{\boldsymbol{\beta}_{k,\mathcal{D}}^{\intercal} \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_{k,\mathcal{D}} - 2(\widehat{\boldsymbol{\nu}}_{k,\mathcal{D}} - \hat{\boldsymbol{\nu}}_{1,\mathcal{D}})^{\intercal} \boldsymbol{\beta}_{k,\mathcal{D}} + \lambda \sum_{j \in \mathcal{D}} \|\boldsymbol{\beta}_{\cdot j}\|\} \tag{G.45}$$

We also collect two technical propositions proved in Mai et al. (2017).

**Proposition G.1.** *For a set of real numbers* $\{a_1, \ldots, a_N\}$*, if* $\sum_{i=1}^{N} a_i^2 \leq \kappa^2 < 1$*, then* $\sum_{i=1}^{N}(a_i + b)^2$ *as long as* $b < \dfrac{1-\kappa}{\sqrt{N}}$.

**Proposition G.2.** *If* $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} < \dfrac{\epsilon}{d}$, $\|(\hat{\boldsymbol{\nu}}_k - \hat{\boldsymbol{\nu}}_1) - (\boldsymbol{\nu}_k - \boldsymbol{\nu}_1)\|_\infty < \epsilon/d$ *for* $\epsilon < \min\{\dfrac{1}{2\varphi}, \dfrac{\lambda}{1+\varphi\Delta}\}$, *we have that*

$$\|\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}} \hat{\boldsymbol{\Sigma}}_{\mathcal{D}\mathcal{D}}^{-1} - \boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}} \boldsymbol{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1}\|_\infty < \frac{\varphi\epsilon}{1-\varphi\epsilon} \tag{G.46}$$

$$\|\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^0 - \boldsymbol{\beta}_{k,\mathcal{D}}^0\|_1 < \frac{\varphi\epsilon(1+\varphi\Delta)}{1-\varphi\epsilon} \tag{G.47}$$

$$\|\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{\textit{oracle}} - \boldsymbol{\beta}_{k,\mathcal{D}}\|_\infty \leq 4\lambda\varphi \tag{G.48}$$

The following lemmas will be helpful in proving Theorem 2

**Lemma G.6.** *For* $(\hat{\boldsymbol{\beta}}_{2,\mathcal{D}}^{(\textit{oracle})}, \cdots, \hat{\boldsymbol{\beta}}_{K,\mathcal{D}}^{(\textit{oracle})})$ *defined in* (G.45)*, if*

$$\max_{j \in \mathcal{D}^c} [\sum_{k=2}^{K} \{(\widehat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}} \hat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{(\textit{oracle})})_j - (\widehat{\nu}_{kj} - \widehat{\nu}_{1j})\}^2]^{1/2} < \lambda/2 \tag{G.49}$$

*then we have that* $\hat{\mathcal{D}} = \mathcal{D}$ *and* $\operatorname{vec}(\widehat{\mathbf{B}}_k)_{\mathcal{D}} = \widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{(\textit{oracle})}$.

*Proof of Lemma G.6.* When (G.49) is true, $\widehat{\boldsymbol{\beta}}_k = (\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{(\text{oracle})}, 0), k = 2, \ldots, K$ satisfies the Karush-Kuhn-Tucker condition. Hence we have the desired conclusion. □

**Proof of Theorem 2.** By Lemmas F.4, G.2, & G.5, for any $0 < \epsilon < \min\{\frac{1}{2\varphi}, \frac{\lambda}{1+\varphi\Delta}\}$, with a probability greater than

$$1 - 2K\sum_{m=1} p_m^2 \exp(-C\frac{np_{-m}\epsilon^2}{d^2}) - CKq^2p\exp\{-\frac{Cn\epsilon^2}{d^2q^2}\} - 2Kp\exp(-Cn\epsilon^2)$$

we have that,

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \leq \frac{\epsilon}{d}, \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_{\max} \leq \frac{\epsilon}{d}. \tag{G.50}$$

By Lemma G.6 and (G.48), it suffices to show that, under the event in (G.50) with a properly chosen $\epsilon$, (G.49) must be true. To this end, note that

$$\begin{aligned} 0 &= \boldsymbol{\beta}_{\mathcal{D}^C} = (\boldsymbol{\Sigma}^{-1}(\boldsymbol{\nu}_k - \boldsymbol{\nu}_1))_{\mathcal{D}^C} \\ &= -(\boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}^C} - \boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}^C})^{-1}(\boldsymbol{\Sigma}_{\mathcal{D}^C}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}(\boldsymbol{\nu}_{k,\mathcal{D}} - \boldsymbol{\nu}_{1,\mathcal{D}}) - (\boldsymbol{\nu}_{k,\mathcal{D}^C} - \boldsymbol{\nu}_{1,\mathcal{D}^C})) \end{aligned}$$

which implies that

$$\boldsymbol{\Sigma}_{\mathcal{D}^C}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}(\boldsymbol{\nu}_{k,\mathcal{D}} - \boldsymbol{\nu}_{1,\mathcal{D}}) = \boldsymbol{\nu}_{k,\mathcal{D}^C} - \boldsymbol{\nu}_{1,\mathcal{D}^C} \tag{G.51}$$

Now we assume that (G.50) holds. For any $j \in \mathcal{D}^C$, we have that

$$|(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{(\text{oracle})})_j - (\widehat{\nu}_k - \widehat{\nu}_1)| \leq |\widehat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{\text{oracle}} - \boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\boldsymbol{\beta}_{k,\mathcal{D}}| + |(\widehat{\nu}_{kj} - \widehat{\nu}_{1j}) - (\nu_{kj} - \nu_{1j})|$$

where we make use of (G.51). Under (G.50), $|(\widehat{\nu}_{kj} - \widehat{\nu}_{1j}) - (\nu_{kj} - \nu_{1j})| \leq \epsilon$. Meanwhile, it is easy to show that

$$\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{\text{oracle}} = \widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^0 - \frac{\lambda}{2}\widehat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}\widehat{\mathbf{t}}_{k,\mathcal{D}} \tag{G.52}$$

where $\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^0 = \widehat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}^{-1}(\widehat{\boldsymbol{\nu}}_{k,\mathcal{D}} - \widehat{\boldsymbol{\nu}}_{1,\mathcal{D}})$ and $\widehat{\mathbf{t}}_{k,\mathcal{D}}$ is the sub-gradient of the group lasso penalty at $\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{\text{oracle}}$. Hence,

$$|\widehat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{\text{oracle}} - \boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\boldsymbol{\beta}_{k,\mathcal{D}}| \leq |(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^0)_j - (\boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\boldsymbol{\beta}_{k,\mathcal{D}})_j| + \frac{\lambda}{2}|(\widehat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}^{-1}\widehat{\mathbf{t}}_{k,\mathcal{D}})_j|$$

For the first term,

$$\begin{aligned} & |(\widehat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^0)_j - (\boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\boldsymbol{\beta}_{k,\mathcal{D}})| \\ \leq\ & \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\|_\infty\|\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^0 - \boldsymbol{\beta}_{k,\mathcal{D}}\|_\infty + \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{D}^C,\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\|_\infty\|\boldsymbol{\beta}_{k,\mathcal{D}}\|_\infty + \|\boldsymbol{\Sigma}_{\mathcal{D}^C,\mathcal{D}}(\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^0 - \boldsymbol{\beta}_{k,\mathcal{D}})\|_\infty \\ \leq\ & C\epsilon \end{aligned}$$

For the second term,

$$|(\widehat{\mathbf{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}}^{-1}\widehat{\mathbf{t}}_{k,\mathcal{D}})_j| \leq |(\mathbf{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\mathbf{t}_{k,\mathcal{D}})_j| + |(\widehat{\mathbf{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}}^{-1}\widehat{\mathbf{t}}_{k,\mathcal{D}})_j - (\mathbf{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\mathbf{t}_{k,\mathcal{D}})_j|$$

Now note that

$$\begin{aligned} & |(\widehat{\mathbf{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}}^{-1}\widehat{\mathbf{t}}_{k,\mathcal{D}})_j - (\mathbf{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\mathbf{t}_{k,\mathcal{D}})_j| \\ \leq\ & \|\hat{\mathbf{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\hat{\mathbf{\Sigma}}_{\mathcal{D}\mathcal{D}}^{-1} - \mathbf{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1}\|_\infty\|\widehat{\mathbf{t}}_{k,\mathcal{D}} - \mathbf{t}_{k,\mathcal{D}}\|_\infty + \|\mathbf{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1}\|_\infty\|\widehat{\mathbf{t}}_{k,\mathcal{D}} - \mathbf{t}_{k,\mathcal{D}}\| \\ & +\|\hat{\mathbf{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\hat{\mathbf{\Sigma}}_{\mathcal{D}\mathcal{D}}^{-1} - \mathbf{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1}\|_\infty|(t_{k,\mathcal{D}})_j| \end{aligned}$$

Because, by (G.48),

$$|\widehat{t}_{kj} - t_{kj}| = |\frac{\widehat{\beta}_{kj}^{\text{oracle}}}{\|\widehat{\boldsymbol{\beta}}_{\cdot j}^{\text{oracle}}\|} - \frac{\beta_{kj}}{\|\boldsymbol{\beta}_{\cdot j}\|}| \leq \frac{2\|\widehat{\boldsymbol{\beta}}_{\cdot j}^{\text{oracle}} - \boldsymbol{\beta}_{\cdot j}\|}{\|\widehat{\boldsymbol{\beta}}_{\cdot j}^{\text{oracle}}\|} \leq C\lambda\varphi$$

we have

$$|(\widehat{\mathbf{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}}^{-1}\widehat{\mathbf{t}}_{k,\mathcal{D}})_j| \leq C\varphi\lambda \tag{G.53}$$

Combine the bounds for the two terms and we have

$$|(\hat{\mathbf{\Sigma}}_{\mathcal{D}^C,\mathcal{D}}\widehat{\boldsymbol{\beta}}_{k,\mathcal{D}}^{(\text{oracle})})_j - (\widehat{\nu}_k - \widehat{\nu}_1)| \leq \frac{\lambda}{2}|(\mathbf{\Sigma}_{\mathcal{D}^C,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D}\mathcal{D}}^{-1}\mathbf{t}_{k,\mathcal{D}})_j| + C\lambda^2 \tag{G.54}$$

Let $\epsilon = \dfrac{1+\varphi\Delta}{4\varphi}\lambda$. Condition (C1) and Proposition G.1 together imply that there exists a generic constant $M_0 > 0$, such that when $\lambda < M_0(1-\kappa)$, (G.49) holds. Let $\epsilon = \dfrac{1+\varphi\Delta}{4\varphi}\lambda$ and we have the first conclusion.

By the first conclusion, if $\{\dfrac{d^2\log(pd)}{n}\}^{1/2} \ll \lambda \ll b_{\min}$, $\lambda \to 0$, we have that $\widehat{\mathcal{D}} = \mathcal{D}$ and $\|\text{vec}(\widehat{\mathbf{B}}_k) - \text{vec}(\mathbf{B}_k)\|_\infty \to 0$. □

**Proof of Theorem 3.** By Lemmas F.4, G.2, & G.5, for any $0 < \epsilon < \min\{\dfrac{1}{2\varphi}, \dfrac{\lambda}{1+\varphi\Delta}\}$, with a probability greater than

$$1 - 2K\sum_{m=1} p_m^2\exp(-C\frac{np_{-m}\epsilon^2}{d^2}) - CKq^2p\exp\{-\frac{Cn\epsilon^2}{d^2q^2}\} - 2Kp\exp(-Cn\epsilon^2)$$

we have that the events in (G.50) have happened. Meanwhile, by the Bernstein inequality, with a probability greater than $1 - 2K\exp(-Cn\epsilon^2)$, we have that

$$|\widehat{\pi}_k - \pi_k| \leq \epsilon \tag{G.55}$$

By the proofs of Theorem 2 in Mai et al. (2017), we have that (G.50) & (G.55) imply $|R_n - R| \leq C\lambda^{1/3}$. Let $\lambda = \dfrac{1+\varphi\Delta}{4\varphi}\lambda$ and we have the first conclusion. The second conclusion is a direct consequence of the first conclusion. □

# References


Bertsekas, D. P. (2016). *Nonlinear Programming*. Athena Scientific, 3 edition.

Bickel, P. and Levina, E. (2004). Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2009). E1071: Misc functions of the department of statistics (e1071), tu wien.

Goeman, J., Meijer, R., and Chaturvedi, N. (2012). penalized: L1 (lasso and fused lasso) and l2 (ridge) penalized estimation in glms and in the cox model. *CRAN R package*.

Lai, Z., Xu, Y., Yang, J., Tang, J., and Zhang, D. (2013). Sparse tensor discriminant analysis. *IEEE Transactions on Image Processing*, 22(10):3904–3915.

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

Li, Q. and Schonfeld, D. (2014). Multilinear discriminant analysis for higher-order tensor data classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2524–2537.

Liaw, A. and Wiener, M. (2014). Package 'randomforest': Breiman and cutler's random forests for classification and regression. 4:6–10.

Mai, Q., Yang, Y., and Zou, H. (2017). Multiclass sparse discriminant analysis. *Statistica Sinica*, In press.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475–494.

Wimalawarne, K., Tomioka, R., and Sugiyama, M. (2016). Theoretical and experimental analyses of tensor-based regression and classification. *Neural Computation*, 28(4):686–715.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

Zhong, W. and Suslick, K. S. (2015). Matrix discriminant analysis with application to colorimetric sensor array data. *Technometrics*, 57(4):524–534.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552.