LINEAR DISCRIMINANT ANALYSIS WITH SPARSE AND DENSE SIGNALS

Ning Wang, Shaokang Ren, and Qing Mai

Beijing Normal University, Microsoft, and Florida State University

Supplementary Material

Section S1 contains some additional simulation results, Section S2 contains the detailed derivation of lemmas, and Section S3 contains the technical proof of the main theorem.

S1 Additional Simulation Results

In this section, we modify Models (D1)-(D3) to test the performance of the proposed methods when the sample sizes in each class are extremely unbalanced. We adopt all the parameters in Models (D1)-(D3) except $n_1 = 10$ and $n_2 = 90$. The resulting models are referred to as Models (D1')-(D3'). Table S.1 presents the miss-classification error rates. From the table, we can see that most of the competitors give error rates very close to 0.1, which indicates that they classify all the samples into one class. As a comparison, SD-LDA is closer to the Bayes error than all the competitors.

Table S.1: The prediction accuracy result. The means and standard errors (in the parentheses) of the prediction error of 100 replicates are reported in percentage.

Models	q	BE	SD-LDA	MSDA	Lasso	elastic-net	SVM	SOS
D1'	0.1	8.24	9.38 (0.09)	10.68 (0.21)	9.96 (0.03)	9.92 (0.03)	10.07 (0.05)	11.83 (0.26)
	0.2	0.71	4.25 (0.11)	10.72 (0.22)	9.96 (0.05)	9.84 (0.04)	10.04 (0.04)	11.62 (0.25)
D2'	0.1	6.35	8.73 (0.11)	10.67 (0.24)	9.85 (0.07)	9.70 (0.07)	10.00 (0.00)	11.16 (0.21)
	0.2	1.66	6.08 (0.97)	9.71 (0.21)	9.51 (0.11)	9.02 (0.12)	10.00(0.00)	9.05 (0.24)
D3'	0.1	6.90	9.04 (0.11)	10.57 (0.26)	9.93 (0.04)	9.83 (0.04)	10.00 (0.00)	11.59 (0.24)
	0.2	2.85	7.45 (0.10)	10.23 (0.20)	9.79 (0.08)	9.54 (0.08)	10.00 (0.00)	9.77 (0.18)

S2 Algorithm to Solve SD-LDA

In this section, we provide the proof of Lemma 1. The result in Example 1 is a direct result of Lemma 1 by replacing parameters with the certain ones in the example. Therefore, we will not give additional proof to it.

Proof. To derive the optimizer to 2.7, we have following steps.

1). Assume we know the true value of $\boldsymbol{\Sigma}$, and $\boldsymbol{\mu}_k$, then 2.7 can be written as

$$
\begin{aligned}
&(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\delta}}_2, \ldots, \hat{\boldsymbol{\beta}}_K, \hat{\boldsymbol{\delta}}_K) = \underset{\boldsymbol{\beta}_k \in \mathbb{R}^p, \boldsymbol{\delta}_k \in \mathbb{R}^p}{\arg\min} \\
&\sum_{k=2}^{K} \{ \frac{1}{2} (\boldsymbol{\beta}_k + \boldsymbol{\delta}_k)^T \boldsymbol{\Sigma} (\boldsymbol{\beta}_k + \boldsymbol{\delta}_k) - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^T (\boldsymbol{\beta}_k + \boldsymbol{\delta}_k) + \lambda_2 ||\boldsymbol{\beta}_k||_2^2 \} \quad \text{(S2.1)} \\
&+\lambda_1 \sum_{j=1}^{p} ||\boldsymbol{\delta}_{\cdot j}||_2.
\end{aligned}
$$

To simplify the notation, denote the ℓ_2 norms as $||\boldsymbol{\beta}_k||_2^2 = \sum_k \beta_{kj}^2$, and

$||\boldsymbol{\delta}_{\cdot j}||_2 = \sqrt{\sum_{k=2}^{K}\delta_{kj}^2}$, then $\lambda_2||\boldsymbol{\beta}_k||_2^2$ and $\lambda_1\sum_{j=1}^{p}||\boldsymbol{\delta}_{\cdot j}||_2$ in (S2.1) refer to the ridge and group lasso penalties respectively. Let $\boldsymbol{\delta}_k$ be fixed for all $k = 2, \ldots, K$, then (S2.1) is and differentiable function of $\boldsymbol{\beta}_k$ for any given $\boldsymbol{\delta}_k$ and any certain k. Thus, the optimizer for $\boldsymbol{\beta}_k$ is

$$\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k) = (2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma})^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}\boldsymbol{\delta}_k). \tag{S2.2}$$

by taking the derivative of $\boldsymbol{\beta}_k$ in (S2.1). It shows that $\boldsymbol{\beta}_k$ is only determined by $\boldsymbol{\delta}_k$ for any certain k.

2). Denote $\boldsymbol{Q} = 2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_{dk} = \boldsymbol{\mu}_k - \boldsymbol{\mu}_1$, then $\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}) + \boldsymbol{\delta}_k = \boldsymbol{Q}^{-1}\boldsymbol{\mu}_{dk} + \boldsymbol{Q}^{-1}(\boldsymbol{Q} - \Sigma)\boldsymbol{\delta}_k$. Bring $\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k)$ into (S2.1), then for any k, we have

$$\lambda_2||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k)||_2^2 = \lambda_2(\boldsymbol{\mu}_{dk}^T\boldsymbol{Q}^{-1}\boldsymbol{Q}^{-1}\boldsymbol{\mu}_{dk} - 2\boldsymbol{\mu}_{dk}\boldsymbol{Q}^{-1}\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{\delta}_k + \boldsymbol{\delta}_k^T\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{\delta}_k), \tag{S2.3}$$

$$(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^T(\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k) + \boldsymbol{\delta}_k) = \boldsymbol{\mu}_{dk}^T\boldsymbol{Q}^{-1}\boldsymbol{\mu}_{dk} + \boldsymbol{\mu}_{dk}^T\boldsymbol{Q}^{-1}(\boldsymbol{Q} - \boldsymbol{\Sigma})\boldsymbol{\delta}_k, \tag{S2.4}$$

and

$$\begin{aligned}&\frac{1}{2}(\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k) + \boldsymbol{\delta}_k)^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k) + \boldsymbol{\delta}_k)\\ =&\frac{1}{2}[\boldsymbol{\mu}_{dk}^T\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}\boldsymbol{\mu}_{dk} + 2\boldsymbol{\mu}_{dk}^T\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}(\boldsymbol{Q} - \boldsymbol{\Sigma})\boldsymbol{\delta}_k\\ &+\boldsymbol{\delta}_k^T(\boldsymbol{Q} - \boldsymbol{\Sigma})\boldsymbol{Q}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}(\boldsymbol{Q} - \boldsymbol{\Sigma})\boldsymbol{\delta}_k].\end{aligned} \tag{S2.5}$$

Thus, (S2.1) becomes

$$\operatorname*{arg\,min}_{\boldsymbol{\delta}_2,\ldots,\boldsymbol{\delta}_K}\sum_{k=2}^{K}\{\lambda_2\boldsymbol{\delta}_k^T\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}\boldsymbol{\delta}_k-2\lambda_2\boldsymbol{\mu}_{dk}^T\boldsymbol{Q}^{-1}\boldsymbol{\delta}_k-\frac{1}{2}\boldsymbol{\mu}_{dk}^T\boldsymbol{Q}^{-1}\boldsymbol{\mu}_{dk}\}+\lambda_1\sum_{j=1}^{p}||\boldsymbol{\delta}_{\cdot j}||_2. \tag{S2.6}$$

3). Remove the constant $\frac{1}{2}\boldsymbol{\mu}_{dk}^T\boldsymbol{Q}^{-1}\boldsymbol{\mu}_{dk}$. Let $\tilde{\boldsymbol{\Sigma}}(\lambda_2)=2\lambda_2\boldsymbol{\Sigma}\boldsymbol{Q}^{-1}$, and $\tilde{\boldsymbol{\mu}}_{dk}(\lambda_2)=\boldsymbol{\mu}_{dk}^T(2\lambda_2\boldsymbol{Q}^{-1})$, then the optimizer for $(\hat{\boldsymbol{\delta}}_2,\ldots,\hat{\boldsymbol{\delta}}_K)$ is

$$(\hat{\boldsymbol{\delta}}_2,\ldots,\hat{\boldsymbol{\delta}}_K)=\operatorname*{arg\,min}_{\boldsymbol{\delta}_2,\ldots,\boldsymbol{\delta}_K}\sum_{k=2}^{K}\{\frac{1}{2}[\boldsymbol{\delta}_k^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\delta}_k]-\tilde{\boldsymbol{\mu}}_{dk}\boldsymbol{\delta}_k\}+\lambda_1\sum_{j=1}^{p}||\boldsymbol{\delta}_{\cdot j}||_2. \tag{S2.7}$$

Denote $\hat{\boldsymbol{Q}}=2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}}$, $\bar{\boldsymbol{\Sigma}}=2\lambda_2\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{Q}}^{-1}$, $\hat{\boldsymbol{\mu}}_{dk}=\hat{\boldsymbol{\mu}}_k-\hat{\boldsymbol{\mu}}_1$ and $\bar{\boldsymbol{\mu}}_{dk}=\hat{\boldsymbol{\mu}}_{dk}^T(2\lambda_2\hat{\boldsymbol{Q}}^{-1})$, then the empirical version for S2.2 and S2.7 are

$$\hat{\boldsymbol{\beta}}_k(\hat{\boldsymbol{\delta}}_k)=(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}(\hat{\boldsymbol{\mu}}_k-\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\delta}}_k), \tag{S2.8}$$

and

$$(\hat{\boldsymbol{\delta}}_2,\ldots,\hat{\boldsymbol{\delta}}_K)=\operatorname*{arg\,min}_{\boldsymbol{\delta}_2,\ldots,\boldsymbol{\delta}_K}\sum_{k=2}^{K}\{\frac{1}{2}[\boldsymbol{\delta}_k^T\bar{\boldsymbol{\Sigma}}\boldsymbol{\delta}_k]-\bar{\boldsymbol{\mu}}_{dk}\boldsymbol{\delta}_k\}+\lambda_1\sum_{j=1}^{p}||\boldsymbol{\delta}_{\cdot j}||_2, \tag{S2.9}$$

which is the conclusion given in Lemma 1 □

S3 Proof of Theorem 1

In this section, we give the detailed proof of Theorem 1. To simplify the notation, we let C and c denote constants that can vary from place to place. Further define the notations $\eqsim$, $\lesssim$ and $\gtrsim$, where $A\eqsim\xi$ indicates that $A=C\xi$ for some constant $C>0$, $A\lesssim\xi$ indicates that $A\leq C\xi$, and $A \gtrsim \xi$ indicates that $A \geq C\xi$ likewise.Furthermore, define the set of non-zero elements in $\boldsymbol{\delta}^*_\cdot$ as

$$\mathcal{D} = \{j \mid \delta^*_{kj} \neq 0 \text{ for some } k\}, \tag{S3.10}$$

then the size of $\mathcal{D}$ is $|\mathcal{D}| = s$.

For any $k \geq 2$, we have the true discriminant directions as

$$\boldsymbol{\theta}^*_k = \boldsymbol{\beta}^*_k + \boldsymbol{\delta}^*_k = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1). \tag{S3.11}$$

Meanwhile, $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\beta}}_k(\hat{\boldsymbol{\delta}}_k) + \hat{\boldsymbol{\delta}}_k$, where $\hat{\boldsymbol{\beta}}_k(\hat{\boldsymbol{\delta}}_k)$ is given in (2.12) and $\hat{\boldsymbol{\delta}}_k$ is estimated by (2.13). Then we have the following result for $||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*_k||_2$.

Lemma 1. *Denote $\boldsymbol{\theta}^* = \boldsymbol{\beta}^* + \boldsymbol{\delta}^*$ and $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\beta}}_k(\hat{\boldsymbol{\delta}}_k) + \hat{\boldsymbol{\delta}}_k$. For any $k \geq 2$, we have*

$$||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*_k||_2 \leq ||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}^*_k||_2 + ||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}^*_k) - \boldsymbol{\beta}^*_k||_2, \tag{S3.12}$$

where $\hat{\boldsymbol{\delta}}_k$ is given in (2.13) and $\hat{\boldsymbol{\beta}}_k$ is given in (2.12).

Proof. Recall that $\hat{\boldsymbol{Q}} = 2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}}$. By (2.12), we have

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*_k =& \hat{\boldsymbol{\beta}}_k(\hat{\boldsymbol{\delta}}_k) + \hat{\boldsymbol{\delta}}_k - \boldsymbol{\beta}^*_k - \boldsymbol{\delta}^*_k \\
=& \hat{\boldsymbol{Q}}^{-1}(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1) + \hat{\boldsymbol{Q}}^{-1}(\hat{\boldsymbol{Q}} - \hat{\boldsymbol{\Sigma}})\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}^*_k - \boldsymbol{\beta}^*_k \\
=& \hat{\boldsymbol{Q}}^{-1}(\hat{\boldsymbol{Q}} - \hat{\boldsymbol{\Sigma}})(\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}^*_k) + \hat{\boldsymbol{Q}}^{-1}(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1) - \hat{\boldsymbol{Q}}^{-1}\hat{\boldsymbol{\Sigma}}\boldsymbol{\delta}^*_k - \boldsymbol{\beta}^*_k \\
=& \hat{\boldsymbol{Q}}^{-1}(\hat{\boldsymbol{Q}} - \hat{\boldsymbol{\Sigma}})(\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}^*_k) + \hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}^*_k) - \boldsymbol{\beta}^*_k.
\end{aligned} \tag{S3.13}$$

Therefore,

$$\begin{aligned}||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2 \leq & ||\hat{\boldsymbol{Q}}^{-1}(\hat{\boldsymbol{Q}} - \hat{\Sigma})(\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*)||_2 + ||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k^*) - \boldsymbol{\beta}_k^*||_2 \\ = & ||\hat{\boldsymbol{Q}}^{-1}(\hat{\boldsymbol{Q}} - \hat{\Sigma})||_2 ||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 + ||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k^*) - \boldsymbol{\beta}_k^*||_2.\end{aligned} \tag{S3.14}$$

Since $||\hat{\boldsymbol{Q}}^{-1}(\hat{\boldsymbol{Q}} - \hat{\Sigma})||_2 = ||(2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1} 2\lambda_2 \boldsymbol{I}_p||_2$, and the eigenvalues of the positive definite matrix $(2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1} 2\lambda_2 \boldsymbol{I}_p$ are all smaller than or equal to 1, we have

$$\begin{aligned}& ||(2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1} 2\lambda_2 \boldsymbol{I}_p||_2 \\ \leq & \sqrt{\max \operatorname{eig}\{[(2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1} 2\lambda_2 \boldsymbol{I}_p]^T (2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1} 2\lambda_2 \boldsymbol{I}_p\}} \\ \leq & 1.\end{aligned} \tag{S3.15}$$

Thus,

$$||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2 \leq ||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 + ||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k^*) - \boldsymbol{\beta}_k^*||_2. \tag{S3.16}$$

□

We will then focus on the event $\{||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2 \lesssim \epsilon\}$. Before specifying its theoretical properties, we first introduce some general results of the estimators for the class means and covariance matrix, as well as some necessary propositions.

Proposition 1. *(Hoeffding's inequality) Let n_k be the sum of n independent and identically distributed Bernoulli random variables with probability π_k*

and $\hat{\pi}_k = \frac{n_k}{n}$ be the estimator for π_k. Then we have

$$\Pr(n_k \leq n\pi_k/2) \leq \exp\{-n\pi_k^2/4\}, \tag{S3.17}$$

and

$$\Pr(|\hat{\pi}_k - \pi_k| \geq \pi_k\epsilon) \leq 2\exp\{-\frac{n\pi_k\epsilon^2}{3}\}, \tag{S3.18}$$

for any k and $\epsilon > 0$.

Lemma 2. *Let $\boldsymbol{\mu}_k$ and $\hat{\boldsymbol{\mu}}_k$ be defined as in Section 2, then when Assumption (A3) holds, we have*

$$\Pr(\{||\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k||_{max} \gtrsim \epsilon\}) \leq Cp\exp\{-Cn\epsilon^2\}, \tag{S3.19}$$

and

$$\Pr(\{||\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k||_2 \lesssim \epsilon\}) \geq 1 - Cp\exp\{-Cn\epsilon^2/p\} \tag{S3.20}$$

for any k and any $\epsilon = o(1)$.

Proof. Because

$$\boldsymbol{X} \mid Y = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \tag{S3.21}$$

we have $\hat{\mu}_{jk} - \mu_k \sim N(0, \frac{1}{n_k}\sigma_{jj}^2)$ for any j, with σ_{jj}^2 being the (j,j)th element of $\boldsymbol{\Sigma}$. Therefore, $\Pr(\{|\hat{\mu}_{jk} - \mu_k| \geq \epsilon \mid n_k\}) \leq C\exp\{-n_k\epsilon^2/\sigma_{jj}^2\}$ according to the properties of sub-Gaussian distribution.

To handle the random variable n_k, we define $\mathcal{A} = \{n_k \colon n_k \geq n\pi_k/2\}$.

Then we have

$$\begin{aligned}\Pr(\{|\hat{\mu}_{jk}-\mu_k|\geq\epsilon\}) =&\mathrm{E}[\Pr(\{|\hat{\mu}_{jk}-\mu_k|\geq\epsilon\mid n_k\})]\\ \leq&\mathrm{E}[C\exp\{-\frac{n_k\epsilon^2}{\sigma_{jj}^2}\}I_{\mathcal{A}}]+\mathrm{E}[C\exp\{-\frac{n_k\epsilon^2}{\sigma_{jj}^2}\}I_{\mathcal{A}^c}]\\ \leq&C\exp\{-\frac{n\pi_k\epsilon^2}{\sigma_{jj}^2}\}+C\mathrm{E}[I_{\mathcal{A}^c}]\end{aligned} \tag{S3.22}$$

By Proposition 1, we can show that $\mathrm{E}[I_{\mathcal{A}^c}]\leq\Pr(n_k\leq n\pi_k/2)\leq\exp\{-\frac{n\pi_k^2}{4}\}$. By Assumption (A3), we can bound π_k away from 0 and 1. Then

$$\exp\{-\frac{n\pi_k^2}{4}\}<C\exp\{-Cn\epsilon^2/\sigma_{jj}^2\} \tag{S3.23}$$

when $\epsilon=o(1)$. Then we have

$$\Pr(\{|\hat{\mu}_{jk}-\mu_k|\geq\epsilon\})\leq C\exp\{-Cn\epsilon^2/\sigma_{jj}^2\} \tag{S3.24}$$

for any $\epsilon=o(1)$.

Since σ_{jj} are bounded for any j by Assumption (A1), we can directly show that

$$\Pr(\{||\boldsymbol{\mu}_k-\hat{\boldsymbol{\mu}}_k||_{max}\gtrsim\epsilon\})\leq\sum_{j=1}^{p}\Pr(\{|\hat{\mu}_{jk}-\mu_k|\geq\epsilon\})\leq Cp\exp\{-Cn\epsilon^2\}. \tag{S3.25}$$

Furthermore, as $\{||\boldsymbol{\mu}_k-\hat{\boldsymbol{\mu}}_k||_2\geq\epsilon\}\subset\bigcup_{j=1}^{p}\{|\hat{\mu}_{jk}-\mu_k|\geq\epsilon/\sqrt{p}\}$, we also have

$$\Pr(\{||\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k||_2 \gtrsim \epsilon\}) \leq \sum_{j=1}^{p} \Pr(\{|\hat{\mu}_{jk} - \mu_k| \geq \epsilon/\sqrt{p}\}) \leq Cp \exp\{-Cn\epsilon^2/p\}. \tag{S3.26}$$

□

As for the estimator for the covariance matrix, we utilize the following proposition to show its convergence result with respect to the ℓ_2 norm.

Proposition 2. *(Proposition 2.1 in Vershynin (2012)) Assume that $\boldsymbol{X} = \{X_1, \ldots, X_p\}$ follows the multivariate normal distribution with covariance $\boldsymbol{\Lambda}$, and $\hat{\boldsymbol{\Lambda}}$ is its sample estimator with sample size m, then we have*

$$\Pr(||\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}||_2 > \epsilon) \leq 2\exp\{2p - Cm\epsilon^2\} \tag{S3.27}$$

for some constant C.

With this conclusion, we could show the asymptotic result of the covariance estimator in our proposal.

Corollary 1. *Let $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ be defined in Section 2 and Assumptions (A1) & (A3) hold. Then we have*

$$||\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}||_2 \lesssim \sqrt{\frac{p \log p}{n}} \tag{S3.28}$$

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

Proof. As $\hat{\mathbf{\Sigma}} = \sum_{k=1}^{K} \hat{\pi}_k \sum_{i \in \mathcal{C}_k} \frac{(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_k)^T}{n_k - 1} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mathbf{\Sigma}}_k$, we have $||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_2 \leq \sum_{k=1}^{K} \hat{\pi}_k ||\hat{\mathbf{\Sigma}}_k - \mathbf{\Sigma}||_2$, which indicates that $\{||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_2 > \epsilon\} \subset \bigcup_{k=1}^{K} \{||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_k||_2 > \epsilon\}$. Then following the same strategy as we use in (S3.22), with $\mathcal{A} = \{n_k\colon n_k \geq n\pi_k/2\}$, we can show that

$$\begin{aligned}\Pr(||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_2 > \epsilon) =& \mathbb{E}[\Pr(||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_2 > \epsilon \mid n_1, \ldots, n_K)] \\ \leq& \sum_{k=1}^{K} \mathbb{E}[\Pr(||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_k||_2 > \epsilon \mid n_k)] \\ \leq& C \exp\{2p - Cn\pi_k \epsilon^2\} + C \exp\{-\frac{n\pi_k^2}{4}\} \\ \leq& C \exp\{2p - Cn\epsilon^2\}\end{aligned} \tag{S3.29}$$

when $\epsilon = o(1)$, according to Propositions 1 & 2. As pointed out by Vershynin (2012), let $\epsilon^2 = (4/C)\log(2p)p/n = O(\frac{p \log p}{n})$, and we can eventually obtain

$$||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_2 \lesssim \sqrt{\frac{p \log p}{n}} \tag{S3.30}$$

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

Lemma 3. *Let $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ be defined in Section 2 and $p < n$, $||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_2 \lesssim \epsilon$ indicates that*

$$||\hat{\mathbf{\Sigma}}^{-1}||_2 \leq C \tag{S3.31}$$

for some constant C and any $0 < \epsilon < u/C$ if Assumption (A1) holds.

Proof. Denote that the eigenvalues of $\mathbf{\Sigma}$ are $\{D_{jj}\}$ and the eigenvalues of $\hat{\mathbf{\Sigma}}$ are $\{\hat{D}_{jj}\}$ for $1 \leq j \leq p$. Without loss of generality, we assume that

$D_{11} \geq \dots D_{jj} \dots \geq D_{pp}$ and $\hat{D}_{11} \geq \dots \hat{D}_{jj} \dots \geq \hat{D}_{pp}$. We could show that $\frac{1}{D_{pp}} = ||\boldsymbol{\Sigma}^{-1}||_2 \leq U$ if Assumption (A1) holds. Therefore, we have $0 < u \leq D_{pp} \leq D_{11} \leq U$ for some constants u and U.

Similarly, as $||\hat{\boldsymbol{\Sigma}}^{-1}||_2 = \frac{1}{\hat{D}_{pp}}$, it is sufficient to show that $\hat{D}_{pp} \geq C > 0$ for some constant C. By equation (11) in Fulton (2000), we know

$$D_{pp} - \hat{D}_{pp} \leq \max \operatorname{eig}\{\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\} = ||\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}||_2. \tag{S3.32}$$

Then, if $||\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}||_2 \lesssim \epsilon$, we have

$$\hat{D}_{pp} \geq D_{pp} - ||\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}||_2 \geq u - C_1\epsilon \geq C \tag{S3.33}$$

for some constant C_1 and C with probability at least $1 - O(p^{-1})$.

□

With these preparations, we can now stick to $||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2$ and its convergence. We first show the following decomposition so that a sufficient condition could be later introduced to bound its convergence rate.

Lemma 4. *When Assumptions (A1) and (A2) hold, $p < n$, and $\lambda_2 \leq \epsilon$ for any ϵ that satisfies the condition in Lemma 3, we have*

$$\begin{aligned} &\Pr(||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2 \gtrsim \epsilon) \\ \leq &\Pr(\{||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 \gtrsim \epsilon\}) + \Pr(\{||\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k||_2 \gtrsim \epsilon\}) \\ + &\Pr(\{||\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1||_2 \gtrsim \epsilon\}) + \Pr(\{||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \gtrsim \epsilon\}). \end{aligned} \tag{S3.34}$$

Proof. By Lemma 1, we could show that

$$\begin{aligned}&\Pr(\{||\hat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k^*||_2\gtrsim\epsilon\})\\ \le&\Pr(\{||\hat{\boldsymbol{\delta}}_k-\boldsymbol{\delta}_k^*||_2+||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k^*)-\boldsymbol{\beta}_k^*||_2\gtrsim\epsilon\})\\ \le&\Pr(\{||\hat{\boldsymbol{\delta}}_k-\boldsymbol{\delta}_k^*||_2\gtrsim\epsilon\}\cup\{||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k^*)-\boldsymbol{\beta}_k^*||_2\gtrsim\epsilon\}).\end{aligned}\tag{S3.35}$$

We first consider the term $||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k^*)-\boldsymbol{\beta}_k^*||_2$. Recall that the true $\boldsymbol{\beta}_k^*$ satisfies the expression $\boldsymbol{\beta}_k^*=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}\boldsymbol{\delta}_k^*)$, and we have

$$\begin{aligned}||\hat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^*||_2\le&||((2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}-\boldsymbol{\Sigma}^{-1})(\hat{\boldsymbol{\mu}}_k-\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\Sigma}}\boldsymbol{\delta}_k^*)||_2\\ &+||\boldsymbol{\Sigma}^{-1}[(\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)-(\hat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)+(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\delta}_k^*)]||_2\\ \le&||(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}-\boldsymbol{\Sigma}^{-1}||_2||\hat{\boldsymbol{\mu}}_k-\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\Sigma}}\boldsymbol{\delta}_k^*||_2\\ &+||\boldsymbol{\Sigma}^{-1}||_2(||\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k||_2+||\hat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1||_2+||\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}||_2||\boldsymbol{\delta}_k^*||_2).\end{aligned}\tag{S3.36}$$

For the term $||(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}-\boldsymbol{\Sigma}^{-1}||_2$, as $||\boldsymbol{\Sigma}^{-1}||_2$ is bounded if Assumption (A1) holds, and $||\hat{\boldsymbol{\Sigma}}^{-1}||_2\le C$ if $\{||\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}||_2\lesssim\epsilon\}$ according to Lemma 3, we have

$$\begin{aligned}&||(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}-\boldsymbol{\Sigma}^{-1}||_2\\ =&||(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}||_2\\ \le&||(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}||_2(2\lambda_2+||\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}||_2)||\boldsymbol{\Sigma}^{-1}||_2\\ \le&||\hat{\boldsymbol{\Sigma}}^{-1}||_2(2\lambda_2+||\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}||_2)||\boldsymbol{\Sigma}^{-1}||_2\\ \lesssim&||\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}||_2\end{aligned}\tag{S3.37}$$

when $\{||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \lesssim \epsilon\}$ holds and $\lambda_2 \lesssim \epsilon$.

Furthermore, $||\boldsymbol{\delta}_k^*||_2$ is finite when Assumptions (A1) & (A3) hold, and $||\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\Sigma}}\boldsymbol{\delta}_k^*||_2$ will be finite when all estimators are sufficiently close to the truth. Intuitively, the scale of $||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k^*) - \boldsymbol{\beta}_k^*||_2$ can be controlled when $\hat{\boldsymbol{\mu}}_k$, $\hat{\boldsymbol{\mu}}_k$, and $\hat{\boldsymbol{\Sigma}}$ are sufficiently close to the truth and λ_2 is sufficiently small. To be specific, denote the event

$$\boldsymbol{B}_k = \{||\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k||_2 \lesssim \epsilon\} \cap \{||\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1||_2 \lesssim \epsilon\} \cap \{||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \lesssim \epsilon\}. \tag{S3.38}$$

Then $\boldsymbol{B}_k$ is a sufficient condition for $\{||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k^*) - \boldsymbol{\beta}_k^*||_2 \lesssim \epsilon\}$ when Assumptions (A1) & (A3) hold and $\lambda_2 \le \epsilon$. Inversely, we have

$$\begin{aligned}
&\Pr(||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2 \gtrsim \epsilon)\\
\le&\Pr(\{||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 \gtrsim \epsilon\}) + \Pr(\{||\hat{\boldsymbol{\beta}}_k(\boldsymbol{\delta}_k^*) - \boldsymbol{\beta}_k^*||_2 \gtrsim \epsilon\})\\
\le&\Pr(\{||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 \gtrsim \epsilon\}) + \Pr(\{||\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k||_2 \gtrsim \epsilon\})\\
+&\Pr(\{||\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1||_2 \gtrsim \epsilon\}) + \Pr(\{||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \gtrsim \epsilon\})
\end{aligned} \tag{S3.39}$$

when Assumptions (A1) & (A3) hold and $\lambda_2 \le \epsilon$ for any ϵ that satisfies the condition in Lemma 3. □

The bound for $\Pr(\{||\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k||_2 \gtrsim \epsilon\})$ for any k is given in Lemma 2 and the bound for $\Pr(\{||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \gtrsim \epsilon\})$ is given in Corollary 1. Hence, we only need to consider the event $\{||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 \gtrsim \epsilon\}$ so as to derive the bound of $\Pr(||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2 \gtrsim \epsilon)$.

We start with the following proposition.

Proposition 3. *(Lemma A.1 in Min et al. (2023) when $s = p$) Let $\tilde{\boldsymbol{\mu}}_k$, $\bar{\boldsymbol{\mu}}_k$, $\tilde{\boldsymbol{\Sigma}}$, and $\bar{\boldsymbol{\Sigma}}$ denoted as in (S2.7) and (2.13). Assume that $\sqrt{\frac{p\log p}{n}} \leq C$, and choose $\lambda_1 = O(\sqrt{\frac{\log p}{n}})$. If $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\mu}}_{dk}$ satisfy*

1. $||\tilde{\boldsymbol{\mu}}_{dk} - \bar{\boldsymbol{\mu}}_{dk}||_{max} \lesssim \sqrt{\frac{\log p}{n}}$;
2. $||(\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}})\boldsymbol{\delta}_k^*||_{max} \lesssim \sqrt{\frac{\log p}{n}}$;
3. $\mathrm{tr}((\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*)^T \tilde{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*)) \gtrsim ||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2^2$

with probability at least $1 - O(p^{-1})$, then we also have

$$||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 \lesssim \sqrt{\frac{p\log p}{n}} \tag{S3.40}$$

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

Proposition 3 shows that we simply need to check the three conditions to derive the bound for $\{||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 \gtrsim \epsilon\}$. We introduce the following lemma to verify first two conditions.

Lemma 5. *Continue to use the notations and settings in Proposition 3. With probability at least $1 - O(p^{-1})$, we have*

1. $||\tilde{\boldsymbol{\mu}}_{dk} - \bar{\boldsymbol{\mu}}_{dk}||_{max} \lesssim \sqrt{\frac{\log p}{n}}$,
2. $||(\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}})\boldsymbol{\delta}_k^*||_{max} \lesssim \sqrt{\frac{\log p}{n}}$,

when Assumptions (A1)–(A3) hold and $\lambda_2 = O(\sqrt{\frac{\log p}{n}})$.

Proof. We start with the first condition. By the definition given in Section 2.3, we have

$$\begin{aligned}||\tilde{\boldsymbol{\mu}}_{dk} - \bar{\boldsymbol{\mu}}_{dk}||_{max} \leq& 2\lambda_2||\hat{\boldsymbol{\mu}}_{dk}^T((2\lambda_2\boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1} - (2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma})^{-1})||_{max}\\ &+2\lambda_2||(\boldsymbol{\mu}_{dk} - \hat{\boldsymbol{\mu}}_{dk})^T(2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma})^{-1}||_{max}\\ \leq& 2\lambda_2||\hat{\boldsymbol{\mu}}_{dk}(2\lambda_2\boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma})^{-1}||_2\\ &+2\lambda_2||(\boldsymbol{\mu}_{dk} - \hat{\boldsymbol{\mu}}_{dk})^T(2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma})^{-1}||_2\\ \leq& 2\lambda_2||\hat{\boldsymbol{\mu}}_{dk}||_2||(2\lambda_2\boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1}||_2||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2||(2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma})^{-1}||_2\\ &+2\lambda_2||\boldsymbol{\mu}_{dk} - \hat{\boldsymbol{\mu}}_{dk}||_2||(2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma})^{-1}||_2.\end{aligned} \tag{S3.41}$$

We now prove that these terms of ℓ_2 norm are finite with high probability. Firstly, by Corollary 1, $||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \lesssim \sqrt{\frac{p\log p}{n}} \leq C$ with probability at least $1 - O(p^{-1})$. We have that $||(2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma})^{-1}||_2 \leq ||\boldsymbol{\Sigma}^{-1}||_2 \leq C$ when Assumption (A1) holds. Similarly, by Lemma 3, we could show that $||(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}||_2 \leq ||\hat{\boldsymbol{\Sigma}}^{-1}||_2 \leq C$ with probability at least $1-O(p^{-1})$ when $||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \lesssim \sqrt{\frac{p\log p}{n}}$. Furthermore, we have

$$||\hat{\boldsymbol{\mu}}_{dk}||_2 \leq ||\boldsymbol{\mu}_{dk} - \hat{\boldsymbol{\mu}}_{dk}||_2 + ||\boldsymbol{\mu}_{dk}||_2 \leq ||\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_k||_2 + ||\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1||_2 + ||\boldsymbol{\mu}_{dk}||_2, \tag{S3.42}$$

which is finite with probability at least $1 - O(p^{-1})$ according to Lemma 2 when Assumptions (A1) & (A3) hold by letting $\epsilon = O(\sqrt{\frac{p\log p}{n}})$.

We then obtain that

$$||\tilde{\boldsymbol{\mu}}_{dk} - \bar{\boldsymbol{\mu}}_{dk}||_{max} \lesssim \lambda_2 \tag{S3.43}$$

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

$$||\tilde{\boldsymbol{\mu}}_{dk} - \bar{\boldsymbol{\mu}}_{dk}||_{max} \lesssim \sqrt{\frac{\log p}{n}} \tag{S3.44}$$

with probability at least $1 - O(p^{-1})$ for $\lambda_2 = O(\sqrt{\frac{\log p}{n}})$.

Now we consider the second condition, the bound of $||(\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}})\boldsymbol{\delta}_k^*||_{max}$. Let $\boldsymbol{e}_i$ to be the orthonormal basis with ith element being 1, and we have

$$\begin{aligned}
&||(\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}})\boldsymbol{\delta}_k^*||_{max} \\
=&\max_j \{\boldsymbol{e}_j(\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}})\boldsymbol{\delta}_k^*\} \\
\le&\max_j \sqrt{\boldsymbol{e}_j^T(\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}})\boldsymbol{e}_j \boldsymbol{\delta}_k^{*T}(\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}})\boldsymbol{\delta}_k^*} \\
\le&\max|\operatorname{eig}\{\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}}\}| \\
=&2\lambda_2 \max|\operatorname{eig}\{\boldsymbol{\Sigma}(2\lambda_2\boldsymbol{I}_p + \boldsymbol{\Sigma})^{-1} - \hat{\boldsymbol{\Sigma}}(2\lambda_2\boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1}\}|.
\end{aligned} \tag{S3.45}$$

According to Corollary 1 and Lemma 3, $\operatorname{eig}\{\hat{\boldsymbol{\Sigma}}(2\lambda_2\boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1}\}$ is bounded with probability at least $1 - O(p^{-1})$ when $||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \lesssim \sqrt{\frac{p\log p}{n}}$ by setting

$\epsilon = O(\sqrt{\frac{p\log p}{n}})$. We thus have

$$\begin{aligned}
&2\lambda_2 \max|\operatorname{eig}\{\boldsymbol{\Sigma}(2\lambda_2\boldsymbol{I}_p+\boldsymbol{\Sigma})^{-1} - \hat{\boldsymbol{\Sigma}}(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}\}| \\
\leq& 2\lambda_2(|\max\operatorname{eig}\{\boldsymbol{\Sigma}(2\lambda_2\boldsymbol{I}_p+\boldsymbol{\Sigma})^{-1}\}| + |\max\operatorname{eig}\{\hat{\boldsymbol{\Sigma}}(2\lambda_2\boldsymbol{I}_p+\hat{\boldsymbol{\Sigma}})^{-1}\}|) \\
\lesssim& \lambda_2
\end{aligned} \tag{S3.46}$$

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

Then we obtain

$$||(\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}})\boldsymbol{\delta}_k^*||_{max} \lesssim \sqrt{\frac{\log p}{n}} \tag{S3.47}$$

with probability at least $1-O(p^{-1})$ for $\lambda_2 = O(\sqrt{\frac{\log p}{n}})$. □

Recall that $\mathcal{D}$ is defined as the set of the sparse signal, as given in (S3.10). Then the two conditions in Lemma 5 imply following result.

Proposition 4. *(Lemma A.4. in Min et al. (2023)) Continue to use the notations and settings in Proposition 3. If*

1. $||\tilde{\boldsymbol{\mu}}_{dk} - \bar{\boldsymbol{\mu}}_{dk}||_{max} \lesssim \sqrt{\frac{\log p}{n}}$,
2. $||(\tilde{\boldsymbol{\Sigma}} - \bar{\boldsymbol{\Sigma}})\boldsymbol{\delta}_k^*||_{max} \lesssim \sqrt{\frac{\log p}{n}}$,

with some $\lambda_1 = O(\sqrt{\frac{\log p}{n}})$, we have that

$$\sum_{j\in\mathcal{D}^{\mathcal{C}}} ||\hat{\delta}_{\cdot j}||_2 \leq \sum_{j\in\mathcal{D}} ||\hat{\delta}_{\cdot j}||_2. \tag{S3.48}$$

We now check the third condition in Proposition 3. This can be done by directly generalizing the result of Lemma A.6 in Min et al. (2023).

Lemma 6. *For $\bar{\boldsymbol{\Sigma}}$, under the condition given in Proposition 3, with probability at least $1 - O(p^{-1})$, we have*

$$1/C - c\epsilon \leq \min \operatorname{eig}\{\bar{\boldsymbol{\Sigma}}\} \leq \max \operatorname{eig}\{\bar{\boldsymbol{\Sigma}}\} \leq C + c\epsilon \tag{S3.49}$$

for some $\epsilon = O(\sqrt{\frac{p \log p}{n}})$ and any $\lambda_2 = O(\sqrt{\frac{\log p}{n}})$ when Assumptions (A1) & (A3) hold.

Proof. Recall that $\bar{\boldsymbol{\Sigma}} = 2\lambda_2 \hat{\boldsymbol{\Sigma}}(2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{\Sigma}})^{-1}$. When the condition $\sqrt{\frac{p \log p}{n}} = o(1)$ in Proposition 3 holds, we have $n > p$ and thus, $\hat{\boldsymbol{\Sigma}}$ is positive definite. Let $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ be some orthogonal matrices such that $\boldsymbol{\Sigma} = \boldsymbol{P}^T \boldsymbol{D} \boldsymbol{P}$ and $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{P}}^T \hat{\boldsymbol{D}} \hat{\boldsymbol{P}}$, where $\boldsymbol{D} = \{D_{jj}\}$ and $\hat{\boldsymbol{D}} = \{\hat{D}_{jj}\}$ are diagonal matrices. Following this decomposition, we have

$$\bar{\boldsymbol{\Sigma}} = 2\lambda_2 \hat{\boldsymbol{P}}^T \hat{\boldsymbol{D}} (2\lambda_2 \boldsymbol{I}_p + \hat{\boldsymbol{D}})^{-1} \hat{\boldsymbol{P}}. \tag{S3.50}$$

Then,

$$\min \operatorname{eig}\{\bar{\boldsymbol{\Sigma}}\} \geq \min_j \{\hat{D}_{jj}\} / (2\lambda_2 + \max_j \{\hat{D}_{jj}\}), \tag{S3.51}$$

and

$$\max \operatorname{eig}\{\bar{\boldsymbol{\Sigma}}\} \leq \max_j \{\hat{D}_{jj}\} / (2\lambda_2 + \min_j \{\hat{D}_{jj}\}). \tag{S3.52}$$

We start with bounding all eigenvalues of $\hat{\boldsymbol{\Sigma}}$ away from 0 and infinity.

When Assumption (A1) holds, we have $0 < 1/U \leq ||\mathbf{\Sigma}||_2 \leq u$, and $0 < 1/u \leq ||\mathbf{\Sigma}^{-1}||_2 \leq U$. Hence, for $\hat{\mathbf{\Sigma}}$, we have

$$\max \operatorname{eig}\{\hat{\mathbf{\Sigma}}\} = \max_j \{\hat{D}_{jj}\} = ||\hat{\mathbf{\Sigma}}||_2 \leq ||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_2 + ||\mathbf{\Sigma}||_2. \tag{S3.53}$$

According to Corollary 1, we have $||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_2 \lesssim \sqrt{\frac{p \log p}{n}}$ with probability at least $1 - O(p^{-1})$ if we let $\epsilon = O(\sqrt{\frac{p \log p}{n}})$. Therefore, with probability at least $1 - O(p^{-1})$, we have

$$\max \operatorname{eig}\{\hat{\mathbf{\Sigma}}\} \leq C_1 + c_1 \sqrt{\frac{p \log p}{n}}. \tag{S3.54}$$

On the other hand, (S3.33) in the proof of Lemma 3 show that

$$\min \operatorname{eig}\{\hat{\mathbf{\Sigma}}\} \geq C_2 - c_2 \sqrt{\frac{p \log p}{n}} > 0. \tag{S3.55}$$

for some constant C_2 and c_2.

Then both $\min_j\{\hat{D}_{jj}\}$ and $\max_j\{\hat{D}_{jj}\}$ are bounded away from 0 and infinite, and we have

$$\min \operatorname{eig}\{\bar{\mathbf{\Sigma}}\} \geq C_m \min_j \{\hat{D}_{jj}\} \geq C_m (C_2 - c_2 \sqrt{\frac{p \log p}{n}}), \tag{S3.56}$$

and

$$\max \operatorname{eig}\{\bar{\mathbf{\Sigma}}\} \leq C_M \max_j \{\hat{D}_{jj}\} \leq C_M (C_1 + c_1 \sqrt{\frac{p \log p}{n}}). \tag{S3.57}$$

Thus, by taking $C = \max\{C_m C_1, C_M C_2\}$, and $c = \max\{C_m c_1, C_M c_2\}$, we eventually get

$$1/C - c\epsilon \leq \min \operatorname{eig}\{\bar{\mathbf{\Sigma}}\} \leq \max \operatorname{eig}\{\bar{\mathbf{\Sigma}}\} \leq C + c\epsilon \tag{S3.58}$$

for some constants C and c and some $\epsilon = O(\sqrt{\frac{p\log p}{n}})$ with probability at least $1 - O(p^{-1})$. □

Combining the results in Proposition 4 and Lemma 6, we can evaluate the third condition in Proposition 3 by the following property.

Proposition 5. *(Result of Lemma A.6 in Min et al. (2023)) Continue to use the notations and settings in Proposition 3. If*

1. $\sum_{j\in\mathcal{D}^c} ||\hat{\delta}_{\cdot j}||_2 \leq \sum_{j\in\mathcal{D}} ||\hat{\delta}_{\cdot j}||_2$,

2. $1/C - c\epsilon \leq \min \mathrm{eig}\{\bar{\mathbf{\Sigma}}\} \leq \max \mathrm{eig}\{\bar{\mathbf{\Sigma}}\} \leq C + c\epsilon$ *for some* $\epsilon = o(1)$,

then we have

$$\mathrm{tr}((\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*)^T \tilde{\mathbf{\Sigma}}(\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*)) \gtrsim ||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2^2. \tag{S3.59}$$

Combining the results in Propositions 3 & 5, Lemmas 2 & 5, and Corollary 1, we could eventually obtain the consistency result of $\hat{\boldsymbol{\theta}}_k$.

Proof of Theorem 1. If Assumptions (A1), (A2) and (A3) hold and let $\lambda_2 =$

$O(\sqrt{\frac{\log p}{n}})$ and $\lambda_1 = O(\sqrt{\frac{\log p}{n}})$, we have

$$\begin{aligned}
&\Pr(||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2 \gtrsim \epsilon) \\
\leq &\Pr(\{||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 \gtrsim \epsilon\}) + \Pr(\{||\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k||_2 \gtrsim \epsilon\}) \\
+ &\Pr(\{||\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1||_2 \gtrsim \epsilon\}) + \Pr(\{||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \gtrsim \epsilon\}) \qquad \text{(S3.60)} \\
\lesssim &\Pr(\{||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 \gtrsim \epsilon\} \wedge \Pr(\{||\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k||_2 \gtrsim \epsilon\}) \\
\wedge &\Pr(\{||\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1||_2 \gtrsim \epsilon\}) \wedge \Pr(\{||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \gtrsim \epsilon\})
\end{aligned}$$

By Lemma 4.

By Propositions 2, 3, and 5, Lemmas 2&5, and Corollary 1, we set $\epsilon = O(\sqrt{\frac{p \log p}{n}})$, and let $\lambda_2 = O(\sqrt{\frac{\log p}{n}})$ and $\lambda_1 = O(\sqrt{\frac{\log p}{n}})$, then

$$\begin{aligned}
&\Pr(\{||\hat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k^*||_2 \gtrsim \sqrt{\frac{p\log p}{n}}\}) \wedge \Pr(\{||\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k||_2 \gtrsim \sqrt{\frac{p\log p}{n}}\}) \\
\wedge &\Pr(\{||\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1||_2 \gtrsim \sqrt{\frac{p\log p}{n}}\}) \wedge \Pr(\{||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2 \gtrsim \sqrt{\frac{p\log p}{n}}\}) \qquad \text{(S3.61)} \\
\lesssim &1/p.
\end{aligned}$$

We eventually have

$$||\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*||_2 \lesssim \sqrt{\frac{p\log p}{n}} \qquad \text{(S3.62)}$$

with probability at least $1 - O(p^{-1})$ for any $\lambda_2 = O(\sqrt{\frac{\log p}{n}})$ and $\lambda_1 = O(\sqrt{\frac{\log p}{n}})$. □

Bibliography

Fulton, W. (2000). Eigenvalues, invariant factors, highest weights, and schubert calculus. *Bulletin of the American Mathematical Society 37*(3), 209–249.

Min, K., Q. Mai, and L. Junge (2023). Optimality in high-dimensional tensor discriminant analysis. *Pattern Recognition 143*(1), 109803.

Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability 25*(3), 655–686.