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 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.

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

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 2.9, we have following steps.

1). Assume we know the true value of $\Sigma$, and $\mu_k$, then 2.9 can be written as

$$\left( \hat{\beta}_2, \hat{\delta}_2, \ldots, \hat{\beta}_K, \hat{\delta}_K \right) = \arg\min_{\beta_k \in \mathbb{R}^p, \delta_k \in \mathbb{R}^p} \sum_{k=2}^K \left( \frac{1}{2} (\beta_k + \delta_k)^T \Sigma (\beta_k + \delta_k) - (\mu_k - \mu_1)^T (\beta_k + \delta_k) + \lambda_2 ||\beta_k||^2_2 \right) + \lambda_1 \sum_{j=1}^p ||\delta_j||^2_2$$

(S2.1)

To simplify the notation, denote the $\ell_2$ norms as $||\beta_k||^2_2 = \sum_k \beta_{kj}^2$, and
$||\delta_j||_2 = \sqrt{\sum_{k=2}^{K} \delta_{kj}}$, then $\lambda_2||\beta_k||_2^2$ and $\lambda_1 \sum_{j=1}^{p} ||\delta_j||_2$ in (S2.1) refer to the ridge and group lasso penalties respectively. Let $\delta_k$ be fixed for all $k = 2, \ldots, K$, then (S2.1) is a differentiable function of $\beta_k$ for any given $\delta_k$ and any certain $k$. Thus, the optimizer for $\beta_k$ is

$$\hat{\beta}_k(\delta_k) = (2\lambda_2 I_p + \Sigma)^{-1}(\mu_k - \mu_1 - \Sigma \delta_k). \quad (S2.2)$$

by taking the derivative of $\beta_k$ in (S2.1). It shows that $\beta_k$ is only determined by $\delta_k$ for any certain $k$.

2). Denote $Q = 2\lambda_2 I_p + \Sigma$ and $\mu_{dk} = \mu_k - \mu_1$, then $\hat{\beta}_k(\delta) + \delta_k = Q^{-1}\mu_{dk} + Q^{-1}(Q - \Sigma)\delta_k$. Bring $\hat{\beta}_k(\delta_k)$ into (S2.1), then for any $k$, we have

$$\lambda_2||\hat{\beta}_k(\delta_k)||_2^2 = \lambda_2(\mu_{dk}^T Q^{-1} Q^{-1} \mu_{dk} - 2\mu_{dk}^T Q^{-1} \Sigma \delta_k + \delta_k^T \Sigma Q^{-1} Q^{-1} \Sigma \delta_k), \quad (S2.3)$$

$$\quad (\mu_k - \mu_1)^T(\hat{\beta}_k(\delta_k) + \delta_k) = \mu_{dk}^T Q^{-1} \mu_{dk} + \mu_{dk}^T Q^{-1} (Q - \Sigma) \delta_k, \quad (S2.4)$$

and

$$\frac{1}{2}(\hat{\beta}_k(\delta_k) + \delta_k)^T \Sigma (\hat{\beta}_k(\delta_k) + \delta_k) = \frac{1}{2}[\mu_{dk}^T Q^{-1} \Sigma Q^{-1} \mu_{dk} + 2\mu_{dk}^T Q^{-1} \Sigma Q^{-1} (Q - \Sigma) \delta_k + \delta_k^T (Q - \Sigma) Q^{-1} \Sigma Q^{-1} (Q - \Sigma) \delta_k]. \quad (S2.5)$$
Thus, (S2.1) becomes
\[
\arg\min_{\delta_2, \ldots, \delta_K} \sum_{k=2}^{K} \{ \lambda_2 \delta_k^T \Sigma Q^{-1} \delta_k - 2 \lambda_2 \mu_{dk}^T Q^{-1} \delta_k - \frac{1}{2} \mu_{dk}^T Q^{-1} \mu_{dk} \} + \lambda_1 \sum_{j=1}^{p} ||\delta_j||_2. 
\]

(S2.6)

3). Remove the constant \( \frac{1}{2} \mu_{dk}^T Q^{-1} \mu_{dk} \). Let \( \hat{\Sigma}(\lambda_2) = 2 \lambda_2 \Sigma Q^{-1} \), and \( \hat{\mu}_{dk}(\lambda_2) = \mu_{dk}^T (2 \lambda_2 Q^{-1}) \), then the optimizer for \( (\hat{\delta}_2, \ldots, \hat{\delta}_K) \) is
\[
(\hat{\delta}_2, \ldots, \hat{\delta}_K) = \arg\min_{\delta_2, \ldots, \delta_K} \sum_{k=2}^{K} \{ \frac{1}{2} [\delta_k^T \hat{\Sigma} \delta_k] - \hat{\mu}_{dk} \delta_k \} + \lambda_1 \sum_{j=1}^{p} ||\delta_j||_2. 
\]

(S2.7)

Denote \( \hat{Q} = 2 \lambda_2 I_p + \hat{\Sigma} \), \( \hat{\Sigma} = 2 \lambda_2 \hat{\Sigma} Q^{-1} \), \( \hat{\mu}_{dk} = \hat{\mu}_k - \hat{\mu}_1 \) and \( \hat{\mu}_{dk} = \hat{\mu}_{dk}^T (2 \lambda_2 \hat{Q}^{-1}) \), then the empirical version for S2.2 and S2.7 are
\[
\hat{\beta}_k(\hat{\delta}_k) = (2 \lambda_2 I_p + \hat{\Sigma})^{-1} (\hat{\mu}_k - \hat{\mu}_1 - \hat{\Sigma} \hat{\delta}_k), 
\]

(S2.8)

and
\[
(\hat{\delta}_2, \ldots, \hat{\delta}_K) = \arg\min_{\delta_2, \ldots, \delta_K} \sum_{k=2}^{K} \{ \frac{1}{2} [\delta_k^T \Sigma \delta_k] - \mu_{dk} \delta_k \} + \lambda_1 \sum_{j=1}^{p} ||\delta_j||_2, 
\]

(S2.9)

which is the conclusion given in Lemma 1.

Thus, (S2.1) becomes
\[
\arg\min_{\delta_2, \ldots, \delta_K} \sum_{k=2}^{K} \{ \lambda_2 \delta_k^T \Sigma Q^{-1} \delta_k - 2 \lambda_2 \mu_{dk}^T Q^{-1} \delta_k - \frac{1}{2} \mu_{dk}^T Q^{-1} \mu_{dk} \} + \lambda_1 \sum_{j=1}^{p} ||\delta_j||_2.
\]

(S2.6)

3). Remove the constant \( \frac{1}{2} \mu_{dk}^T Q^{-1} \mu_{dk} \). Let \( \hat{\Sigma}(\lambda_2) = 2 \lambda_2 \Sigma Q^{-1} \), and \( \hat{\mu}_{dk}(\lambda_2) = \mu_{dk}^T (2 \lambda_2 Q^{-1}) \), then the optimizer for \( (\hat{\delta}_2, \ldots, \hat{\delta}_K) \) is
\[
(\hat{\delta}_2, \ldots, \hat{\delta}_K) = \arg\min_{\delta_2, \ldots, \delta_K} \sum_{k=2}^{K} \{ \frac{1}{2} [\delta_k^T \hat{\Sigma} \delta_k] - \hat{\mu}_{dk} \delta_k \} + \lambda_1 \sum_{j=1}^{p} ||\delta_j||_2.
\]

(S2.7)

Denote \( \hat{Q} = 2 \lambda_2 I_p + \hat{\Sigma} \), \( \hat{\Sigma} = 2 \lambda_2 \hat{\Sigma} Q^{-1} \), \( \hat{\mu}_{dk} = \hat{\mu}_k - \hat{\mu}_1 \) and \( \hat{\mu}_{dk} = \hat{\mu}_{dk}^T (2 \lambda_2 \hat{Q}^{-1}) \), then the empirical version for S2.2 and S2.7 are
\[
\hat{\beta}_k(\hat{\delta}_k) = (2 \lambda_2 I_p + \hat{\Sigma})^{-1} (\hat{\mu}_k - \hat{\mu}_1 - \hat{\Sigma} \hat{\delta}_k),
\]

(S2.8)

and
\[
(\hat{\delta}_2, \ldots, \hat{\delta}_K) = \arg\min_{\delta_2, \ldots, \delta_K} \sum_{k=2}^{K} \{ \frac{1}{2} [\delta_k^T \Sigma \delta_k] - \mu_{dk} \delta_k \} + \lambda_1 \sum_{j=1}^{p} ||\delta_j||_2,
\]

(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 \( \approx, \lesssim \) and \( \gtrsim \), where \( A \approx \xi \) 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 \( \delta^* \) as

\[
D = \{ j \mid \delta_{kj}^* \neq 0 \text{ for some } k \},
\]

then the size of \( D \) is \(|D| = s\).

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

\[
\theta_k^* = \beta_k^* + \delta_k^* = \Sigma^{-1}(\mu_k - \mu_1).
\]

Meanwhile, \( \hat{\theta}_k = \hat{\beta}_k(\hat{\delta}_k) + \hat{\delta}_k \), where \( \hat{\beta}_k(\hat{\delta}_k) \) is given in (2.14) and \( \hat{\delta}_k \) is estimated by (2.15). Then we have the following result for \( ||\hat{\theta}_k - \theta_k^*||_2 \).

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

\[
||\hat{\theta}_k - \theta_k^*||_2 \leq ||\hat{\delta}_k - \delta_k^*||_2 + ||\hat{\beta}_k(\delta_k^*) - \beta_k^*||_2,
\]

where \( \hat{\delta}_k \) is given in (2.15) and \( \hat{\beta}_k \) is given in (2.14).

**Proof.** Recall that \( \hat{Q} = 2\lambda_2 I_p + \hat{\Sigma} \). By (2.14), we have

\[
\hat{\theta}_k - \theta_k^* = \hat{\beta}_k(\hat{\delta}_k) + \hat{\delta}_k - \beta_k^* - \delta_k^*
\]

\[
= \hat{Q}^{-1}(\hat{\mu}_k - \hat{\mu}_1) + \hat{Q}^{-1}(\hat{Q} - \hat{\Sigma})\delta_k - \delta_k^* - \beta_k^*
\]

\[
= \hat{Q}^{-1}(\hat{Q} - \hat{\Sigma})(\hat{\delta}_k - \delta_k^*) + \hat{Q}^{-1}(\hat{\mu}_k - \hat{\mu}_1) - \hat{Q}^{-1}\hat{\Sigma}\delta_k^* - \beta_k^*
\]

\[
= \hat{Q}^{-1}(\hat{Q} - \hat{\Sigma})(\hat{\delta}_k - \delta_k^*) + \beta_k(\delta_k^*) - \beta_k^*.
\]
Therefore,
\[ \|\hat{\theta}_k - \theta_k^*\|_2 \leq \|\hat{Q}^{-1}(\hat{Q} - \hat{\Sigma})(\hat{\delta}_k - \delta_k^*)\|_2 + \|\hat{\beta}_k(\delta_k^*) - \beta_k^*\|_2 \]
\[ = \|\hat{Q}^{-1}(\hat{Q} - \hat{\Sigma})\|_2\|\hat{\delta}_k - \delta_k^*\|_2 + \|\hat{\beta}_k(\delta_k^*) - \beta_k^*\|_2. \]  
\[(S3.14)\]

Since \( \|\hat{Q}^{-1}(\hat{Q} - \hat{\Sigma})\|_2 = \|((2\lambda_2 I_p + \hat{\Sigma})^{-1}2\lambda_2 I_p)\|_2 \), and the eigenvalues of the positive definite matrix \((2\lambda_2 I_p + \hat{\Sigma})^{-1}2\lambda_2 I_p\) are all smaller than or equal to 1, we have
\[ \|((2\lambda_2 I_p + \hat{\Sigma})^{-1}2\lambda_2 I_p)\|_2 \leq \sqrt{\max \text{eig}\{(2\lambda_2 I_p + \hat{\Sigma})^{-1}2\lambda_2 I_p\}} \] 
\[ \leq 1. \]  
\[(S3.15)\]

Thus,
\[ \|\hat{\theta}_k - \theta_k^*\|_2 \leq \|\hat{\delta}_k - \delta_k^*\|_2 + \|\hat{\beta}_k(\delta_k^*) - \beta_k^*\|_2. \]  
\[(S3.16)\]

We will then focus on the event \( \{\|\hat{\theta}_k - \theta_k^*\|_2 \lesssim \epsilon\} \). 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 \( \pi_k \)
and \( \hat{\pi}_k = \frac{n_k}{n} \) be the estimator for \( \pi_k \). 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 \( \bm{\mu}_k \) and \( \hat{\bm{\mu}}_k \) be defined as in Section 2, then when Assumption (A3) holds, we have

\[
\Pr(||\bm{\mu}_k - \hat{\bm{\mu}}_k||_{\max} \gtrsim \epsilon) \leq C_p \exp\{-C_n \epsilon^2\}, \tag{S3.19}
\]

and

\[
\Pr(||\bm{\mu}_k - \hat{\bm{\mu}}_k||_2 \lesssim \epsilon) \geq 1 - C_p \exp\{-C_n \epsilon^2/p\} \tag{S3.20}
\]

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

**Proof.** Because \( \bm{X} \mid Y = k \sim N(\bm{\mu}_k, \Sigma) \tag{S3.21} \)

we have \( \hat{\mu}_{jk} - \mu_k \sim N(0, \frac{1}{n_k} \sigma_{jj}^2) \) for any \( j \), with \( \sigma_{jj}^2 \) being the \( (j, j) \)th element of \( \Sigma \). 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 \( \mathcal{A} = \{n_k: n_k \geq n\pi_k/2\} \).
Then we have

\[ \Pr(\{|\hat{\mu}_{jk} - \mu_k| \geq \epsilon\}) = \mathbb{E}[\Pr(\{|\hat{\mu}_{jk} - \mu_k| \geq \epsilon \mid n_k\})] \]

\[ \leq \mathbb{E}[C \exp\{-\frac{n_k \epsilon^2}{\sigma_{jj}^2}\}I_{A^c}] + \mathbb{E}[C \exp\{-\frac{n_k \epsilon^2}{\sigma_{jj}^2}\}I_{A^c}] \]

\[ \leq C \exp\{-\frac{n \pi_k \epsilon^2}{\sigma_{jj}^2}\} + C \mathbb{E}[I_{A^c}] \]

(S3.22)

By Proposition 1, we can show that \( \mathbb{E}[I_{A^c}] \leq \Pr(n_k \leq n \pi_k/2) \leq \exp\{-\frac{n \pi_k^2}{4}\}. \)

By Assumption (A3), we can bound \( \pi_k \) away from 0 and 1. Then

\[ \exp\{-\frac{n \pi_k^2}{4}\} \leq C \exp\{-Cn \epsilon^2/\sigma_{jj}^2\} \]

(S3.23)

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

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

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

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

\[ \Pr(\{|\mu_k - \hat{\mu}_k|_\max \geq \epsilon\}) \leq \sum_{j=1}^p \Pr(\{|\hat{\mu}_{jk} - \mu_k| \geq \epsilon\}) \leq Cp \exp\{-Cn \epsilon^2\}. \]

(S3.25)

Furthermore, as \( \{|\mu_k - \hat{\mu}_k|_2 \geq \epsilon\} \subset \bigcup_{j=1}^p \{|\hat{\mu}_{jk} - \mu_k| \geq \epsilon/\sqrt{p}\}, \) we also
have
\[
\Pr(\{||\mu_k - \hat{\mu}_k||_2 \geq \epsilon\}) \leq \sum_{j=1}^{p} \Pr(\{|\hat{\mu}_{jk} - \mu_k| \geq \epsilon/\sqrt{p}\}) \leq Cp \exp\{-Cn\epsilon^2/p\}.
\]  
(S3.26)

As for the estimator for the covariance matrix, we utilize the following proposition to show its convergence result with respect to the $\ell_2$ norm.

**Proposition 2.** (Proposition 2.1 in Vershynin (2012)) Assume that $X = \{X_1, \ldots, X_p\}$ follows the multivariate normal distribution with covariance $\Lambda$, and $\hat{\Lambda}$ is its sample estimator with sample size $m$, then we have
\[
\Pr(||\Lambda - \hat{\Lambda}||_2 > \epsilon) \leq 2 \exp\{2p - Cm\epsilon^2\}
\]  
(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 $\Sigma$ and $\hat{\Sigma}$ be defined in Section 2 and Assumptions (A1) $\&$ (A3) hold. Then we have
\[
||\Sigma - \hat{\Sigma}||_2 \lesssim \sqrt{\frac{p \log p}{n}}
\]  
(S3.28)

with probability at least $1 - O(p^{-1})$. 
Proof. As $\hat{\Sigma} = \sum_{k=1}^{K} \hat{\pi}_k \sum_{i \in C_k} \frac{(X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T}{n_k - 1} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\Sigma}_k$, we have $||\Sigma - \hat{\Sigma}||_2 \leq \sum_{k=1}^{K} \hat{\pi}_k ||\hat{\Sigma}_k - \Sigma||_2$, which indicates that $\{||\Sigma - \hat{\Sigma}||_2 > \epsilon\} \subset \bigcup_{k=1}^{K} \{||\Sigma - \Sigma_k||_2 > \epsilon\}$. Then following the same strategy as we use in (S3.22), with $A = \{n_k: n_k \geq n\pi_k/2\}$, we can show that

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

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

$$||\Sigma - \hat{\Sigma}||_2 \lesssim \sqrt{\frac{p \log p}{n}}$$

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

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

$$||\hat{\Sigma}^{-1}||_2 \leq C$$

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

**Proof.** Denote that the eigenvalues of $\Sigma$ are $\{D_{jj}\}$ and the eigenvalues of $\hat{\Sigma}$ are $\{\hat{D}_{jj}\}$ for $1 \leq j \leq p$. Without loss of generality, we assume that
\[ D_{11} \geq \ldots \geq D_{jj} \geq D_{pp} \quad \text{and} \quad \hat{D}_{11} \geq \ldots \geq \hat{D}_{jj} \geq \hat{D}_{pp}. \] We could show that \[ \frac{1}{D_{pp}} = \|\Sigma^{-1}\|_2 \leq U \] 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 \( \|\hat{\Sigma}^{-1}\|_2 = \frac{1}{\hat{D}_{pp}} \), it is sufficient to show that \( \hat{D}_{pp} \geq C > 0 \) for some constant \( C \). By equation (11) in Fulton (2000), we know
\[
D_{pp} - \hat{D}_{pp} \leq \max \text{eig}\{\Sigma - \hat{\Sigma}\} = \|\Sigma - \hat{\Sigma}\|_2. \tag{S3.32}
\]
Then, if \( \|\Sigma - \hat{\Sigma}\|_2 \lesssim \epsilon \), we have
\[
\hat{D}_{pp} \geq D_{pp} - \|\Sigma - \hat{\Sigma}\|_2 \geq u - C_1 \epsilon \geq C \tag{S3.33}
\]
for some constant \( C_1 \) and \( C \) with probability at least \( 1 - O(p^{-1}) \).

\[ \square \]

With these preparations, we can now stick to \( \|\hat{\theta}_k - \theta_k^*\|_2 \) 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 \( \epsilon \) that satisfies the condition in Lemma 3, we have
\[
\Pr(\|\hat{\theta}_k - \theta_k^*\|_2 \gtrsim \epsilon) \leq \Pr(\|\hat{\delta}_k - \delta_k^*\|_2 \gtrsim \epsilon) + \Pr(\|\hat{\mu}_k - \mu_k\|_2 \gtrsim \epsilon) \tag{S3.34}
\]
\[
+ \Pr(\|\hat{\Sigma} - \Sigma\|_2 \gtrsim \epsilon) + \Pr(\|\hat{\Sigma} - \Sigma\|_2 \gtrsim \epsilon).
\]
Proof. By Lemma 1, we could show that

$$\Pr(\{||\hat{\theta}_k - \theta^*_k||_2 \gtrsim \epsilon\})$$

$$\leq \Pr(\{||\hat{\delta}_k - \delta^*_k||_2 + ||\hat{\beta}_k(\delta^*_k) - \beta^*_k||_2 \gtrsim \epsilon\})$$  \hspace{1cm} (S.3.35)

$$\leq \Pr(\{||\hat{\delta}_k - \delta^*_k||_2 \gtrsim \epsilon\} \cup \{||\hat{\beta}_k(\delta^*_k) - \beta^*_k||_2 \gtrsim \epsilon\}).$$

We first consider the term $$||\hat{\beta}_k(\delta^*_k) - \beta^*_k||_2$$. Recall that the true $$\beta^*_k$$ satisfies the expression $$\beta^*_k = \Sigma^{-1}(\mu_k - \mu_1 - \Sigma\delta^*_k)$$, and we have

$$||\hat{\beta}_k - \beta^*_k||_2 \leq ||((2\lambda_2I_p + \hat{\Sigma})^{-1} - \Sigma^{-1})(\mu_k - \mu_1 - \hat{\Sigma}\delta^*_k)||_2$$

$$+ ||\Sigma^{-1}[(\hat{\mu}_k - \mu_k) - (\hat{\mu}_1 - \mu_1) + (\hat{\Sigma} - \Sigma)\delta^*_k]||_2$$

$$\leq ||(2\lambda_2I_p + \hat{\Sigma})^{-1} - \Sigma^{-1}||_2(||\hat{\mu}_k - \hat{\mu}_1 - \hat{\Sigma}\delta^*_k||_2$$

$$+ ||\Sigma^{-1}||_2(||\hat{\mu}_k - \mu_k||_2 + ||\hat{\mu}_1 - \mu_1||_2 + ||\hat{\Sigma} - \Sigma||_2||\delta^*_k||_2).$$

(S.3.36)

For the term $$||(2\lambda_2I_p + \hat{\Sigma})^{-1} - \Sigma^{-1}||_2$$, as $$||\Sigma^{-1}||_2$$ is bounded if Assumption (A1) holds, and $$||\hat{\Sigma}^{-1}||_2 \leq C$$ if $$\{||\hat{\Sigma} - \Sigma||_2 \lesssim \epsilon\}$$ according to Lemma 3, we have

$$||(2\lambda_2I_p + \hat{\Sigma})^{-1} - \Sigma^{-1}||_2$$

$$= ||(2\lambda_2I_p + \hat{\Sigma})^{-1}(2\lambda_2I_p + \hat{\Sigma} - \Sigma)\Sigma^{-1}||_2$$

$$\leq ||(2\lambda_2I_p + \hat{\Sigma})^{-1}||_2(2\lambda_2 + ||\hat{\Sigma} - \Sigma||_2)||\Sigma^{-1}||_2$$  \hspace{1cm} (S.3.37)

$$\leq ||\hat{\Sigma}^{-1}||_2(2\lambda_2 + ||\hat{\Sigma} - \Sigma||_2)||\Sigma^{-1}||_2$$

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

Furthermore, \|\delta_k^*\|_2 is finite when Assumptions (A1) & (A3) hold, and
\|\hat{\mu}_k - \hat{\mu}_1 - \hat{\Sigma}\delta_k^*\|_2 will be finite when all estimators are sufficiently close to
the truth. Intuitively, the scale of \|\hat{\beta}_k(\delta_k^*) - \beta_k^*\|_2 can be controlled when
\hat{\mu}_k, \hat{\mu}_k, and \hat{\Sigma} are sufficiently close to the truth and \lambda_2 is sufficiently small.
To be specific, denote the event

\[ B_k = \{||\hat{\mu}_k - \mu_k||_2 \lesssim \epsilon\} \cap \{||\hat{\mu}_1 - \mu_1||_2 \lesssim \epsilon\} \cap \{||\hat{\Sigma} - \Sigma||_2 \lesssim \epsilon\}. \quad (S3.38) \]

Then \( B_k \) is a sufficient condition for \{||\hat{\beta}_k(\delta_k^*) - \beta_k^*||_2 \lesssim \epsilon\} when Assumptions (A1) & (A3) hold and \lambda_2 \leq \epsilon. Inversely, we have

\[ \Pr(||\hat{\theta}_k - \theta_k^*||_2 \gtrsim \epsilon) \]
\[ \leq \Pr(||\hat{\delta}_k - \delta_k^*||_2 \gtrsim \epsilon) + \Pr(||\hat{\beta}_k(\delta_k^*) - \beta_k^*||_2 \gtrsim \epsilon) \]
\[ \leq \Pr(||\hat{\delta}_k - \delta_k^*||_2 \gtrsim \epsilon) + \Pr(||\hat{\mu}_k - \mu_k||_2 \gtrsim \epsilon) \]
\[ + \Pr(||\hat{\mu}_1 - \mu_1||_2 \gtrsim \epsilon) + \Pr(||\hat{\Sigma} - \Sigma||_2 \gtrsim \epsilon) \]

when Assumptions (A1) & (A3) hold and \lambda_2 \leq \epsilon for any \epsilon that satisfies the
condition in Lemma 3. \[ \square \]

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

**Proposition 3.** (Lemma A.1 in Min and Mai (2022) when $s = p$) Let $\tilde{\mu}_k$, $\bar{\mu}_k$, $\tilde{\Sigma}$, and $\bar{\Sigma}$ denoted as in (S2.7) and (2.15). Assume that $\sqrt{\frac{p \log p}{n}} \leq C$, and choose $\lambda_1 = O(\sqrt{\log p})$. If $\tilde{\Sigma}$ and $\tilde{\mu}_{dk}$ satisfy

1. $||\tilde{\mu}_{dk} - \bar{\mu}_{dk}||_{max} \lesssim \sqrt{\frac{\log p}{n}}$;

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

3. $\text{tr}((\hat{\delta}_k - \delta^*_k)^T \tilde{\Sigma}(\hat{\delta}_k - \delta^*_k)) \gtrsim ||\hat{\delta}_k - \delta^*_k||_2^2$

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

$$||\hat{\delta}_k - \delta^*_k||_2 \lesssim \sqrt{\frac{p \log p}{n}}$$  \hspace{1cm} (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 $\{||\hat{\delta}_k - \delta^*_k||_2 \gtrsim \epsilon\}$. 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. $||\tilde{\mu}_{dk} - \bar{\mu}_{dk}||_{max} \lesssim \sqrt{\frac{\log p}{n}}$;

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


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

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

\[
||\hat{\mu}_{dk} - \hat{\mu}_{dk}||_{max} \leq 2\lambda_2 \left||\hat{\mu}_{dk}^T ((2\lambda_2 I_p + \hat{\Sigma})^{-1} - (2\lambda_2 I_p + \Sigma)^{-1})\right||_{max} \\
+ 2\lambda_2 \left||(\mu_{dk} - \hat{\mu}_{dk})^T (2\lambda_2 I_p + \Sigma)^{-1}||_{max}\right|
\]

\[
\leq 2\lambda_2 \left||(\mu_{dk}^T ((2\lambda_2 I_p + \hat{\Sigma})^{-1} (\hat{\Sigma} - \Sigma) (2\lambda_2 I_p + \Sigma)^{-1}) ||_2 \\
+ 2\lambda_2 \left||(\mu_{dk} - \hat{\mu}_{dk})^T (2\lambda_2 I_p + \Sigma)^{-1}||_2\right|
\]

\[
\leq 2\lambda_2 \left||(\hat{\mu}_{dk} - \mu_{dk}) - (\hat{\mu}_{dk} - \hat{\mu}_{dk})^T (2\lambda_2 I_p + \Sigma)^{-1} ||_2\right|
\]

\[
+ 2\lambda_2 \left||(\mu_{dk} - \hat{\mu}_{dk}) - (\mu_{dk} - \hat{\mu}_{dk})^T (2\lambda_2 I_p + \Sigma)^{-1} ||_2\right|
\]

\[
\leq 2\lambda_2 ||(\hat{\Sigma} - \Sigma)||_2 ||(2\lambda_2 I_p + \hat{\Sigma})^{-1} ||_2 ||\hat{\Sigma} - \Sigma||_2 ||(2\lambda_2 I_p + \Sigma)^{-1} ||_2 \\
+ 2\lambda_2 ||(\mu_{dk} - \hat{\mu}_{dk})||_2 ||(2\lambda_2 I_p + \Sigma)^{-1} ||_2.
\]

(S3.41)

We now prove that these terms of \( \ell_2 \) norm are finite with high probability. Firstly, by Corollary 1, \( ||\hat{\Sigma} - \Sigma||_2 \lesssim \frac{p \log p}{n} \leq C \) with probability at least \( 1 - O(p^{-1}) \). We have that \( ||(2\lambda_2 I_p + \Sigma)^{-1} ||_2 \leq ||\Sigma^{-1}||_2 \leq C \) when Assumption (A1) holds. Similarly, by Lemma 3, we could show that \( ||(2\lambda_2 I_p + \hat{\Sigma})^{-1} ||_2 \leq ||\hat{\Sigma}^{-1}||_2 \leq C \) with probability at least \( 1 - O(p^{-1}) \) when \( ||\hat{\Sigma} - \Sigma||_2 \lesssim \frac{p \log p}{n} \). Furthermore, we have

\[
||\hat{\mu}_{dk}||_2 \leq ||\mu_{dk} - \hat{\mu}_{dk}||_2 + ||\mu_{dk}||_2 \leq ||\mu_k - \hat{\mu}_k||_2 + ||\mu_k - \hat{\mu}_k||_2 + ||\mu_{dk}||_2,
\]

(S3.42)

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

We then obtain that

$$||\hat{\mu}_{dk} - \bar{\mu}_{dk}||_{max} \lesssim \lambda_2$$  \hspace{1cm} (S3.43)

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

$$||\hat{\mu}_{dk} - \bar{\mu}_{dk}||_{max} \lesssim \sqrt{\log p/n}$$  \hspace{1cm} (S3.44)

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

Now we consider the second condition, the bound of $|| (\hat{\Sigma} - \bar{\Sigma}) \delta_k^* ||_{max}$.

Let $e_i$ to be the orthonormal basis with $i$th element being 1, and we have

$$||(\hat{\Sigma} - \bar{\Sigma}) \delta_k^*||_{max}
= \max_j \{ e_j (\hat{\Sigma} - \bar{\Sigma}) \delta_k^* \}
\leq \max_j \sqrt{e_j^T (\hat{\Sigma} - \bar{\Sigma}) e_j \delta_k^* \delta_k^T (\hat{\Sigma} - \bar{\Sigma}) \delta_k^*}
\leq \max |\text{eig}\{\hat{\Sigma} - \bar{\Sigma}\}|
= 2\lambda_2 \max |\text{eig}\{\Sigma (2\lambda_2 I_p + \Sigma)^{-1} - \hat{\Sigma} (2\lambda_2 I_p + \hat{\Sigma})^{-1}\}|.$$

According to Corollary 1 and Lemma 3, $\text{eig}\{\hat{\Sigma} (2\lambda_2 I_p + \hat{\Sigma})^{-1}\}$ is bounded with probability at least $1 - O(p^{-1})$ when $||\Sigma - \hat{\Sigma}||_2 \lesssim \sqrt{p \log p/n}$ by setting
\[ \epsilon = O(\sqrt{\frac{p \log p}{n}}). \] We thus have

\[
2\lambda_2 \max |\text{eig}\{\Sigma(2\lambda_2 I_p + \Sigma)^{-1} - \hat{\Sigma}(2\lambda_2 I_p + \hat{\Sigma})^{-1}\}| \leq 2\lambda_2 (|\max\text{eig}\{\Sigma(2\lambda_2 I_p + \Sigma)^{-1}\}| + |\max\text{eig}\{\hat{\Sigma}(2\lambda_2 I_p + \hat{\Sigma})^{-1}\}|) \lesssim \lambda_2 \]

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

Then we obtain

\[
|| (\hat{\Sigma} - \Sigma) \delta_k^* ||_{\text{max}} \lesssim \sqrt{\frac{\log p}{n}} \quad (S3.47)
\]

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

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 and Mai (2022)) Continue to use the notations and settings in Proposition 3. If

1. \(|| \hat{\mu}_{dk} - \mu_{dk} ||_{\text{max}} \lesssim \sqrt{\frac{\log p}{n}}\),

2. \(|| (\hat{\Sigma} - \Sigma) \delta_k^* ||_{\text{max}} \lesssim \sqrt{\frac{\log p}{n}}\),

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

\[
\sum_{j \in \mathcal{D}^c} ||\hat{\delta}_j||_2 \leq \sum_{j \in \mathcal{D}} ||\hat{\delta}_j||_2. \quad (S3.48)
\]
We now check the third condition in Proposition 3. This can be done by directly generalizing the result of Lemma A.6 in Min and Mai (2022).

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

$$1/C - c\epsilon \leq \min \text{eig}\{\Sigma\} \leq \max \text{eig}\{\Sigma\} \leq C + c\epsilon \quad (S3.49)$$

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

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

$$\Sigma = 2\lambda_2 \hat{P}^T \hat{D}(2\lambda_2 I_p + \hat{D})^{-1} \hat{P}. \quad (S3.50)$$

Then,

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

and

$$\max \text{eig}\{\Sigma\} \leq \max_j \{\hat{D}_{jj}\}/(2\lambda_2 + \min_j \{\hat{D}_{jj}\}). \quad (S3.52)$$

We start with bounding all eigenvalues of $\hat{\Sigma}$ away from 0 and infinity.
When Assumption (A1) holds, we have \(0 < 1/U \leq \|\Sigma\|_2 \leq u\), and 
\(0 < 1/u \leq \|\Sigma^{-1}\|_2 \leq U\). Hence, for \(\hat{\Sigma}\), we have

\[
\max \text{eig}\{\hat{\Sigma}\} = \max \{\hat{D}_{jj}\} = \|\hat{\Sigma}\|_2 \leq \|\Sigma - \hat{\Sigma}\|_2 + \|\Sigma\|_2. \tag{S3.53}
\]

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

\[
\max \text{eig}\{\hat{\Sigma}\} \leq C_1 + c_1 \sqrt{\frac{p \log p}{n}}. \tag{S3.54}
\]

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

\[
\min \text{eig}\{\hat{\Sigma}\} \geq C_2 - c_2 \sqrt{\frac{p \log p}{n}} > 0. \tag{S3.55}
\]

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

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

and

\[
\max \text{eig}\{\hat{\Sigma}\} \leq C_M \max_j \{\hat{D}_{jj}\} \leq C_M (C_1 + c_1 \sqrt{\frac{p \log p}{n}}). \tag{S3.57}
\]

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

\[
1/C - c\epsilon \leq \min \text{eig}\{\hat{\Sigma}\} \leq \max \text{eig}\{\hat{\Sigma}\} \leq C + c\epsilon \tag{S3.58}
\]
for some constants $C$ and $c$ and some $\epsilon = O\left(\sqrt{\frac{\log p}{n}}\right)$ 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 and Mai (2022)) Continue to use the notations and settings in Proposition 3. If

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

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

then we have

$$\text{tr}((\hat{\delta}_k - \delta_k^*)^T \Sigma (\hat{\delta}_k - \delta_k^*)) \gtrsim ||\hat{\delta}_k - \delta_k^*||_2^2. \quad (S3.59)$$

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

**Proof of Theorem 1.** If Assumptions (A1), (A2) and (A3) hold and let $\lambda_2 =$
S3. PROOF OF THEOREM 1

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

\[
\Pr(||\hat{\theta}_k - \theta^*_k||_2 \gtrsim \epsilon) \\
\leq \Pr(||\hat{\delta}_k - \delta^*_k||_2 \gtrsim \epsilon) + \Pr(||\hat{\mu}_k - \mu_k||_2 \gtrsim \epsilon) \\
+ \Pr(||\hat{\mu}_1 - \mu_1||_2 \gtrsim \epsilon) + \Pr(||\hat{\Sigma} - \Sigma||_2 \gtrsim \epsilon) \\
\lesssim \Pr(||\hat{\delta}_k - \delta^*_k||_2 \gtrsim \epsilon) \land \Pr(||\hat{\mu}_k - \mu_k||_2 \gtrsim \epsilon) \\
\land \Pr(||\hat{\mu}_1 - \mu_1||_2 \gtrsim \epsilon) \land \Pr(||\hat{\Sigma} - \Sigma||_2 \gtrsim \epsilon)
\]

(S3.60)

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

\[
\Pr(||\hat{\delta}_k - \delta^*_k||_2 \gtrsim \sqrt{\frac{p \log p}{n}}) \land \Pr(||\hat{\mu}_k - \mu_k||_2 \gtrsim \sqrt{\frac{p \log p}{n}}) \\
\land \Pr(||\hat{\mu}_1 - \mu_1||_2 \gtrsim \sqrt{\frac{p \log p}{n}}) \land \Pr(||\hat{\Sigma} - \Sigma||_2 \gtrsim \sqrt{\frac{p \log p}{n}})
\]

(S3.61)

\( \lesssim 1/p. \)

We eventually have

\[
||\hat{\theta}_k - \theta^*_k||_2 \lesssim \sqrt{\frac{p \log p}{n}}
\]

(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

