Appendices

A.1 Connections with Fisher's discriminant analysis

For simplicity, in this subsection we denote $\boldsymbol{\eta}$ as the discriminant directions defined by Fisher's discriminant analysis in (4), and $\boldsymbol{\theta}$ as the discriminant directions defined by Bayes rule. Our method gives a sparse estimate of $\boldsymbol{\theta}$. In this section, we discuss the connection between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, and hence the connection between our method and Fisher's discriminant analysis. We first comment on the advantage of directly estimating $\boldsymbol{\theta}$ rather than estimating $\boldsymbol{\eta}$. Then we discuss how to estimate $\boldsymbol{\eta}$ once $\hat{\boldsymbol{\theta}}$ is available.

There are two advantages of estimating $\boldsymbol{\theta}$ rather than $\boldsymbol{\eta}$. Firstly, estimating $\boldsymbol{\theta}$ allows for simultaneous estimation of all the discriminant directions. Note that (4) requires that $\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\eta}_l = 0$ for any $l < k$. This requirement almost necessarily leads to a sequential optimization problem, which is indeed the case for sparse optimal scoring and ℓ_1 penalized Fisher's discriminant analysis. In our proposal, the discriminant direction $\boldsymbol{\theta}_k$ is determined by the covariance matrix and the mean vectors $\boldsymbol{\mu}_k$ within Class k, but is not related to $\boldsymbol{\theta}_l$ for any $l \neq k$. Hence, our proposal can simultaneously estimate all the directions by solving a convex problem. Secondly, it is easy to study the theoretical properties if we focus on $\boldsymbol{\theta}$. On the population level, $\boldsymbol{\theta}$ can be written out in explicit forms and hence it is easy to calculate the difference between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ in the theoretical studies. Since $\boldsymbol{\eta}$ do not have closed-form solutions even when we know all the parameters, it is relatively harder to study its theoretical properties.

Moreover, if one is specifically interested in the discriminant directions $\boldsymbol{\eta}$, it is very easy to obtain a sparse estimate of them once we have a sparse estimate of $\boldsymbol{\theta}$. For convenience, for any positive integer m, denote 0_m as an m-dimensional vector with all entries being 0, 1_m as an m-dimensional vector with all entries being 1, and $\mathbf{I}_m$ as the $m \times m$ identity matrix. The following lemma provides an approach to estimating $\boldsymbol{\eta}$ once $\hat{\boldsymbol{\theta}}$ is available. The proof is relegated to Section A.2.

Lemma 3. *The discriminant directions $\boldsymbol{\eta}$ contain all the right eigenvectors of $\boldsymbol{\theta}_0\boldsymbol{\Pi}\boldsymbol{\delta}_0^{\mathrm{T}}$ corresponding to positive eigenvalues, where $\boldsymbol{\theta}_0 = (0_p, \boldsymbol{\theta})$, $\boldsymbol{\Pi} = \mathbf{I}_K - \frac{1}{K}1_K1_K^{\mathrm{T}}$, and $\boldsymbol{\delta}_0 = (\boldsymbol{\mu}_1 - \bar{\boldsymbol{\mu}}, \ldots, \boldsymbol{\mu}_K - \bar{\boldsymbol{\mu}})$ with $\bar{\boldsymbol{\mu}} = \sum_{k=1}^K \pi_k \boldsymbol{\mu}_k$.*

Therefore, once we have obtained a sparse estimate of $\boldsymbol{\theta}$, we can estimate $\boldsymbol{\eta}$ as follows. Without loss of generality write $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_{\hat{\mathcal{D}}}^{\mathrm{T}}, 0)^{\mathrm{T}}$, where $\hat{\mathcal{D}} = \{j : \hat{\boldsymbol{\theta}}_{\cdot j} \neq 0\}$. Then $\hat{\boldsymbol{\theta}}_0 = (0, \hat{\boldsymbol{\theta}})$. On the other hand, set $\hat{\boldsymbol{\delta}}_0 = (\hat{\boldsymbol{\mu}}_1 - \hat{\bar{\boldsymbol{\mu}}}, \ldots, \hat{\boldsymbol{\mu}}_K - \hat{\bar{\boldsymbol{\mu}}})$ where $\hat{\boldsymbol{\mu}}_k$ are sample estimates and $\hat{\bar{\boldsymbol{\mu}}} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\boldsymbol{\mu}}_k$. It follows that $\hat{\boldsymbol{\theta}}_0 \boldsymbol{\Pi} \hat{\boldsymbol{\delta}}_0 = ((\hat{\boldsymbol{\theta}}_{0,\hat{\mathcal{D}}} \boldsymbol{\Pi} \hat{\boldsymbol{\delta}}_{0,\hat{\mathcal{D}}}^{\mathrm{T}})^{\mathrm{T}}, 0)^{\mathrm{T}}$. Consequently, we can perform eigen-decomposition on $\hat{\boldsymbol{\theta}}_{0,\hat{\mathcal{D}}} \boldsymbol{\Pi} \hat{\boldsymbol{\delta}}_{0,\hat{\mathcal{D}}}^{\mathrm{T}}$ to obtain $\hat{\boldsymbol{\eta}}_{\hat{\mathcal{D}}}$. Because $\hat{\mathcal{D}}$ is a small subset of the original dataset, this decomposition will be computationally efficient. Then $\hat{\boldsymbol{\eta}}$ would be $(\hat{\boldsymbol{\eta}}_{\hat{\mathcal{D}}}^{\mathrm{T}}, 0)^{\mathrm{T}}$.

A.2 Technical Proofs

Proof of Proposition 1. We first show (15).

For a vector $\boldsymbol{\theta} \in \mathbb{R}^p$, define

$$L^{\mathrm{MSDA}}(\boldsymbol{\theta}, \lambda) = \frac{1}{2} \boldsymbol{\theta}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}} \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_1, \tag{22}$$

$$L^{\mathrm{ROAD}}(\boldsymbol{\theta}, \lambda) = \boldsymbol{\theta}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_1 \tag{23}$$

Set $\tilde{\boldsymbol{\theta}} = c_0(\lambda)^{-1} \hat{\boldsymbol{\theta}}^{\mathrm{MSDA}}(\lambda)$. Since $\tilde{\boldsymbol{\theta}}^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_1) = 1$, it suffices to check that, for any $\tilde{\boldsymbol{\theta}}'$ such that $(\tilde{\boldsymbol{\theta}}')^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_1) = 1$, we have $L^{\mathrm{ROAD}}(\tilde{\boldsymbol{\theta}}, \frac{2\lambda}{|c_0(\lambda)|}) \leq L^{\mathrm{ROAD}}(\tilde{\boldsymbol{\theta}}', \frac{2\lambda}{|c_0(\lambda)|})$. Now for any such $\tilde{\boldsymbol{\theta}}'$,

$$L^{\mathrm{MSDA}}(c_0(\lambda)\tilde{\boldsymbol{\theta}}', \lambda) = c_0(\lambda)^2 L^{\mathrm{ROAD}}(\tilde{\boldsymbol{\theta}}', \frac{2\lambda}{|c_0(\lambda)|}) - c_0(\lambda) \tag{24}$$

Similarly,

$$L^{\mathrm{MSDA}}(c_0(\lambda)\tilde{\boldsymbol{\theta}}, \lambda) = c_0(\lambda)^2 L^{\mathrm{ROAD}}(\tilde{\boldsymbol{\theta}}, \frac{2\lambda}{|c_0(\lambda)|}) - c_0(\lambda). \tag{25}$$

Since $L^{\mathrm{MSDA}}(c_0(\lambda)\tilde{\boldsymbol{\theta}}, \lambda) \leq L^{\mathrm{MSDA}}(c_0(\lambda)\tilde{\boldsymbol{\theta}}', \lambda)$, we have (15).

On the other hand, by Theorem 1 in Mai & Zou (2013*b*), we have

$$\hat{\boldsymbol{\theta}}^{\mathrm{DSDA}}(\lambda) = c_1(\lambda) \hat{\boldsymbol{\theta}}^{\mathrm{ROAD}}(\frac{\lambda}{n|c_1(\lambda)|}) \tag{26}$$

Therefore,

$$\hat{\boldsymbol{\theta}}^{\text{ROAD}}(\frac{2\lambda}{|c_0(\lambda)|}) = \hat{\boldsymbol{\theta}}^{\text{ROAD}}\left((\frac{2n|c_1(\lambda)|\lambda}{|c_0(\lambda)|})/(n|c_1(\lambda)|)\right) \tag{27}$$

$$= \left(c_1(\frac{2n|c_1(\lambda)|\lambda}{|c_0(\lambda)|})\right)^{-1}\hat{\boldsymbol{\theta}}^{\text{DSDA}}\left(\frac{2n|c_1(\lambda)|\lambda}{|c_0(\lambda)|}\right) \tag{28}$$

$$= (c_1(a\lambda))^{-1}\hat{\boldsymbol{\theta}}^{\text{DSDA}}(a\lambda) \tag{29}$$

Combine (29) with (15) and we have (16). □

Proof of Lemma 1. We start with simplifying the first part of our objective function, $\frac{1}{2}\boldsymbol{\theta}_k^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\theta}_k - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\boldsymbol{\theta}_k$.

First, note that

$$\frac{1}{2}\boldsymbol{\theta}_k^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\theta}_k = \frac{1}{2}\sum_{l,m=1}^{p}\theta_{kl}\theta_{km}\hat{\sigma}_{lm} \tag{30}$$

$$= \frac{1}{2}\theta_{kj}^2\hat{\sigma}_{jj} + \frac{1}{2}\sum_{l\neq j}\theta_{kl}\theta_{kj}\hat{\sigma}_{lj} + \frac{1}{2}\sum_{m\neq j}\theta_{kj}\theta_{km}\hat{\sigma}_{jm} + \frac{1}{2}\sum_{l\neq j,m\neq j}\theta_{kl}\theta_{km}\hat{\sigma}_{lm} \tag{31}$$

(32)

Because $\hat{\sigma}_{lj} = \hat{\sigma}_{jl}$, we have $\sum_{l\neq j}\theta_{kl}\theta_{kj}\hat{\sigma}_{lj} = \sum_{m\neq j}\theta_{kj}\theta_{km}\hat{\sigma}_{jm}$. It follows that

$$\frac{1}{2}\boldsymbol{\theta}_k^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\theta}_k = \frac{1}{2}\theta_{kj}^2\hat{\sigma}_{jj} + \sum_{l\neq j}\theta_{kj}\theta_{kl}\hat{\sigma}_{lj} + \frac{1}{2}\sum_{l\neq j,m\neq j}\theta_{kl}\theta_{km}\hat{\sigma}_{lm} \tag{33}$$

Then recall that $\hat{\boldsymbol{\delta}}^k = \hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1$. We have

$$(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\boldsymbol{\theta}_k = \sum_{l=1}^{p}\delta_l^k\theta_{kl} = \delta_j^k\theta_{kj} + \sum_{l\neq j}\delta_l^k\theta_{kl} \tag{34}$$

Combine (33) and (34) and we have

$$\frac{1}{2}\boldsymbol{\theta}_k^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\theta}_k - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\boldsymbol{\theta}_k \tag{35}$$

$$= \frac{1}{2}\theta_{kj}^2\hat{\sigma}_{jj} + \sum_{l\neq j}\theta_{kj}\theta_{kl}\hat{\sigma}_{lj} + \frac{1}{2}\sum_{l\neq j,m\neq j}\theta_{kl}\theta_{km}\hat{\sigma}_{lm} - \delta_j^k\theta_{kj} - \sum_{l\neq j}\delta_l^k\theta_{kl} \tag{36}$$

$$= \frac{1}{2}\theta_{kj}^2\hat{\sigma}_{jj} + (\sum_{l\neq j}\hat{\sigma}_{l,j}\theta_{kl} - \hat{\delta}_j^k)\theta_{kj} + \frac{1}{2}\sum_{m\neq j,l\neq j}\theta_{kl}\theta_{km}\hat{\sigma}_{lm} - \sum_{l\neq j}\hat{\delta}_l^k\theta_{kl} \tag{37}$$

Note that the last two terms does not involve $\boldsymbol{\theta}_{.j}$. Therefore, given $\{\boldsymbol{\theta}_{.j'}, j' \neq j\}$, the solution of $\boldsymbol{\theta}_{.j}$ is defined as

$$\arg\min_{\boldsymbol{\theta}_{2,j},\ldots,\boldsymbol{\theta}_{K,j}} \sum_{k=2}^{K} \{\frac{1}{2}\theta_{kj}^2 \hat{\sigma}_{jj} + (\sum_{l\neq j} \hat{\sigma}_{lj}\theta_{kl} - \hat{\delta}_j^k)\theta_{kj}\} + \lambda \|\boldsymbol{\theta}_{.j}\|,$$

which is equivalent to (17). It is easy to get (18) from (17) (Yuan & Lin 2006). □

Proof of Lemma 2. We start with the first conclusion. If all elements in $\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}^c}$ are equal to 0, then we must have $\boldsymbol{\Sigma}_{j,\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\mathbf{t}_{k,\mathcal{D}} = 0$ and hence $\max_{j\in\mathcal{D}^c}\{\sum_{k=2}^K(\boldsymbol{\Sigma}_{j,\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\mathbf{t}_{k,\mathcal{D}})^2\}^{1/2} = 0$. It follows that Condition (C0) holds.

For the second conclusion, note that, when $\sigma_{ij} = \rho^{|i-j|}$ and $\mathcal{D} = \{1, \ldots, d\}$, for $j \in \mathcal{D}^C$, we have $\boldsymbol{\Sigma}_{j,\mathcal{D}} = \rho^{j-d}\boldsymbol{\Sigma}_{d,\mathcal{D}}$. Consequently,

$$\boldsymbol{\Sigma}_{j,\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1} = \rho^{j-d}(0_{d-1}, 1).$$

Hence,

$$\sum_{k=2}^{K}(\boldsymbol{\Sigma}_{j,\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\mathbf{t}_{k,\mathcal{D}})^2 = \rho^{2(j-d)}\sum_{k=2}^{K} t_{kd}^2 = \rho^{2(j-d)} < 1$$

which implies Condition (C0).

For the third conclusion, note that, if $\boldsymbol{\Sigma}$ is compound symmetry, then we can write $\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}} = (1 - \rho)\mathbf{I}_d + \rho 1_d 1_d^{\mathrm{T}}$. Straightforward calculation verifies that

$$\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1} = \frac{1}{1-\rho}\mathbf{I}_d - \frac{\rho}{[1+(d-1)\rho](1-\rho)} 1_d 1_d^{\mathrm{T}}.$$

Consequently, for any $j \in \mathcal{D}^{\mathcal{C}}$,

$$\boldsymbol{\Sigma}_{j,\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1} = a 1_d^{\mathrm{T}}$$

where $a = \dfrac{\rho}{1-\rho}(1 - \dfrac{d\rho}{1+(d-1)\rho})$. Therefore, by Cauchy-Schwarz inequality, we have

$$\sum_{k=2}^{K}(\mathbf{\Sigma}_{j,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\mathbf{t}_{k,\mathcal{D}})^2 = a^2\sum_{k=2}^{K}(1_d^{\mathrm{T}}\mathbf{t}_{k,\mathcal{D}})^2 \le a^2\sum_{k=2}^{K}\{(1_d^{\mathrm{T}}1_d)(\mathbf{t}_{k,\mathcal{D}}^{\mathrm{T}}\mathbf{t}_{k,\mathcal{D}}^{\mathrm{T}})\}$$
$$= a^2 d\sum_{k=2}^{K}\sum_{j\in\mathcal{D}}t_{kj}^2 = a^2 d\sum_{j\in\mathcal{D}}\sum_{k=2}^{K}t_{kj}^2 = a^2 d^2$$

where we use the fact $\sum_{k=2}^{K} t_{kj}^2 = 1$ for any $j \in \mathcal{D}$. Hence,

$$\{\sum_{k=2}^{K}(\mathbf{\Sigma}_{j,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}\mathbf{t}_{k,\mathcal{D}})^2\}^{1/2} = ad = \frac{d\rho}{1-\rho}(1-\frac{d\rho}{1+(d-1)\rho}) = \frac{d\rho}{1+(d-1)\rho} < 1$$

and we have the desired conclusion. □

In what follows we use C to denote a generic constant for convenience.

Now we define an oracle “estimator" that relies on the knowledge of $\mathcal{D}$ for a specific tuning parameter λ:

$$\hat{\boldsymbol{\theta}}_{\mathcal{D}}^{\text{oracle}} = \arg\min_{\boldsymbol{\theta}_{2,\mathcal{D}},\ldots,\boldsymbol{\theta}_{K,\mathcal{D}}}\sum_{k=2}^{K}\{\frac{1}{2}\boldsymbol{\theta}_{k,\mathcal{D}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}}\boldsymbol{\theta}_{k,\mathcal{D}} - (\hat{\boldsymbol{\mu}}_{k,\mathcal{D}} - \hat{\boldsymbol{\mu}}_{1,\mathcal{D}})^{\mathrm{T}}\boldsymbol{\theta}_{k,\mathcal{D}}\} + \lambda\sum_{j\in\mathcal{D}}\|\theta_{\cdot j}\|. \tag{38}$$

The proof of Theorem 1 is based on a series of technical lemmas. For convenience, in what follows we simply write $\boldsymbol{\theta}^{\text{Bayes}}$ as $\boldsymbol{\theta}$. This convention shall not be confused with the generic $\boldsymbol{\theta}$ in an objective function.

Lemma 4. *Define $\hat{\boldsymbol{\theta}}_{\mathcal{D}}^{\text{oracle}}(\lambda)$ as in* (38)*. Then $\hat{\boldsymbol{\theta}}_k = (\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{\text{oracle}}, 0), k = 2,\ldots,K$ is the solution to* (10) *if*

$$\max_{j\in\mathcal{D}^c}[\sum_{k=2}^{K}\{(\hat{\mathbf{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{(\text{oracle})})_j - (\hat{\mu}_{kj} - \hat{\mu}_{1j})\}^2]^{1/2} < \lambda. \tag{39}$$

Proof of Lemma 4. The proof is completed by checking that $\hat{\boldsymbol{\theta}}_k = (\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{\text{oracle}}(\lambda), 0)$ satisfies the KKT condition of (10). □

Lemma 5. *For each k, $\mathbf{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}(\boldsymbol{\mu}_{k,\mathcal{D}} - \boldsymbol{\mu}_{1,\mathcal{D}}) = \boldsymbol{\mu}_{k,\mathcal{D}^c} - \boldsymbol{\mu}_{1,\mathcal{D}^c}$.*

Proof of Lemma 5. For each k, we have $\boldsymbol{\theta}_{k,\mathcal{D}^c} = 0$. By definition, $\boldsymbol{\theta}_{\mathcal{D}^c} = (\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1))_{\mathcal{D}^c}$. Then by block inversion, we have that

$$\boldsymbol{\theta}_{k,\mathcal{D}^c} = -(\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}^c} - \boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}^c})^{-1}(\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D},\mathcal{D}}^{-1}(\boldsymbol{\mu}_{k,\mathcal{D}} - \boldsymbol{\mu}_{1,\mathcal{D}}) - (\boldsymbol{\mu}_{k,\mathcal{D}^c} - \boldsymbol{\mu}_{1,\mathcal{D}^c})),$$

and the conclusion follows. □

Proposition 2. *Under Condition (C1), there exists a constant ϵ_0 such that for any $0 < \epsilon \leq \epsilon_0$ we have*

$$\mathrm{pr}\{|(\hat{\mu}_{kj} - \hat{\mu}_{1j}) - (\mu_{kj} - \mu_{1j})| \geq \epsilon\} \leq C\exp(-C\frac{n\epsilon^2}{K}) + C\exp(-\frac{Cn}{K^2}), \quad (40)$$

$$k = 2, \ldots, K,\ j = 1, \ldots, p;$$

$$\mathrm{pr}(|\hat{\sigma}_{ij} - \sigma_{ij}| \geq \epsilon) \leq C\exp(-C\frac{n\epsilon^2}{K}) + C\exp(-\frac{Cn}{K^2}),\ i, j = 1, \ldots, p. \quad (41)$$

Proof of Proposition 2. We first show (40). We start with the fact that, conditional on $\mathbf{Y}$, $\hat{\mu}_{kj} \sim N(\mu_{kj}, \frac{\sigma_{jj}}{n_k})$. Therefore, for any $s > 0$, we have

$$\mathrm{pr}(\hat{\mu}_{kj} - \mu_{kj} \geq \epsilon \mid Y) = \mathrm{pr}(e^{s(\hat{\mu}_{kj} - \mu_{kj})} \geq e^{s\epsilon} \mid Y) \leq e^{-s\epsilon}E\left\{e^{s(\hat{\mu}_{kj} - \mu_{kj})} \mid Y\right\} = e^{-s\epsilon + \frac{\sigma_{jj}s^2}{2n_k}}$$

Let $s = \dfrac{n_k\epsilon}{\sigma_{jj}}$ and we have

$$\mathrm{pr}(\hat{\mu}_{kj} - \mu_{kj} \geq \epsilon \mid Y) \leq \exp(-\frac{n_k\epsilon^2}{2\sigma_{jj}}) \leq \exp(-Cn_k\epsilon^2),$$

where the last inequality follows from the assumption that σ_{jj} are bounded from above. Repeat these steps for $\mu_{kj} - \hat{\mu}_{kj}$ and we have

$$\mathrm{pr}(\hat{\mu}_{kj} - \mu_{kj} \leq -\epsilon \mid Y) \leq \exp(-Cn_k\epsilon^2)$$

Hence,

$$\mathrm{pr}(|\hat{\mu}_{kj} - \mu_{kj}| \geq \epsilon \mid Y) \leq C\exp(-Cn_k\epsilon^2)$$

It follows that

$$\mathrm{pr}(|\hat{\mu}_{kj} - \mu_{kj}| \geq \epsilon) \leq E(\mathrm{pr}(|\hat{\mu}_{kj} - \mu_{kj}| \geq \epsilon \mid Y)) \leq E(C\exp(-Cn_k\epsilon^2)) \tag{42}$$

$$= E\left\{C\exp(-Cn_k\epsilon^2)1(n_k > \pi_k n/2)\right\} + E\left\{C\exp(-Cn_k\epsilon^2)1(n_k < \pi_k n/2)\right\} \tag{43}$$

For the first term, note that, if $n_k > \pi_k n/2$, we must have

$$C\exp(-Cn_k\epsilon^2) \leq C\exp(-C\pi_k n\epsilon^2) \leq C\exp(-C\frac{n\epsilon^2}{K}),$$

where the last inequality follows from Condition (C1). Hence,

$$E\left\{C\exp(-Cn_k\epsilon^2)1(n_k > \pi_k n/2)\right\} \leq C\exp(-C\frac{n\epsilon^2}{K}). \tag{44}$$

For the second term, note that

$$E\left\{C\exp(-Cn_k\epsilon^2)1(n_k < \pi_k n/2)\right\} \leq C\mathrm{pr}(n_k < \pi_k n/2)),$$

Define $W^i = 1(Y^i = k)$. Then $W^i \sim$ Bernoulli(π_k) and $n_k = \sum_{i=1}^n W^i$. By Hoeffding's inequality we have that

$$\mathrm{pr}(n_k < \pi_k n/2)) = \mathrm{pr}(|\frac{1}{n}\sum_{i=1}^n W^i - E(W^i)| > \pi_k/2) \tag{45}$$

$$\leq C\exp(-Cn\pi_k^2) \leq C\exp(-C\frac{n}{K^2}), \tag{46}$$

where the last inequality again follows from Condition (C1). Combine (43),(44) and (46), and we have the desired conclusion.

A similar inequality holds for $\hat{\mu}_{1j}$, and (40) follows.

For (41), note that

$$
\begin{aligned}
\hat{\sigma}_{ij} &= \frac{1}{n-K}\sum_{k=1}^{K}\sum_{Y^m=k}(X_i^m-\hat{\mu}_{ki})(X_j^m-\hat{\mu}_{kj}) \\
&= \frac{1}{n-K}\sum_{k=1}^{K}\sum_{Y^m=k}(X_i^m-\mu_i^m)(X_j^m-\mu_j^m)+\frac{1}{n-K}\sum_{k=1}^{K}n_k(\hat{\mu}_{ki}-\mu_{ki})(\hat{\mu}_{kj}-\mu_{kj}) \\
&= \hat{\sigma}_{ij}^{(0)}+\frac{1}{n-K}\sum_{k=1}^{K}n_k(\hat{\mu}_{ki}-\mu_{ki})(\hat{\mu}_{kj}-\mu_{kj}).
\end{aligned}
$$

Now by Chernoff bound, $\mathrm{pr}(|\hat{\sigma}_{ij}^{(0)}-\sigma_{ij}|\geq\epsilon)\leq C\exp(-Cn\epsilon^2)$. Combining this fact with (40), we have the desired result.

□

Now we consider two events depending on a small $\epsilon>0$:

$$
\begin{aligned}
A(\epsilon) &= \{|\hat{\sigma}_{ij}-\sigma_{ij}|<\frac{\epsilon}{d}\text{ for any } i=1,\cdots,p \text{ and } j\in\mathcal{D}\}, \\
B(\epsilon) &= \{|(\hat{\mu}_{kj}-\hat{\mu}_{1j})-(\mu_{kj}-\mu_{1j})|<\epsilon \text{ for any } k \text{ and } j\}.
\end{aligned}
$$

By simple union bounds, we can derive Lemma 4 and Lemma 5.

Lemma 6. *There exist a constant ϵ_0 such that for any $\epsilon\leq\epsilon_0$ we have*

1. $\mathrm{pr}(A(\epsilon))\geq 1-Cpd\exp(-Cn\frac{\epsilon^2}{Kd^2})-CK\exp(-\frac{Cn}{K^2})$;

2. $\mathrm{pr}(B(\epsilon))\geq 1-Cp(K-1)\exp(-C\frac{n\epsilon^2}{K})-CK\exp(-\frac{Cn}{K^2})$;

3. $\mathrm{pr}(A(\epsilon)\cap B(\epsilon))\geq 1-\gamma(\epsilon)$, *where*

$$
\gamma(\epsilon)=Cpd\exp(-C\frac{n\epsilon^2}{d^2})+Cp(K-1)\exp(-C\frac{n\epsilon^2}{K})+2CK\exp(-\frac{Cn}{K^2}).
$$

Lemma 7. *Assume that both $A(\epsilon)$ and $B(\epsilon)$ have occurred. We have the following conclusions:*

$$\|\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{D}}\|_\infty < \epsilon;$$

$$\|\hat{\mathbf{\Sigma}}_{\mathcal{D}^c,\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\|_\infty < \epsilon;$$

$$\|(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1) - (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)\|_\infty < \epsilon;$$

$$\|(\hat{\boldsymbol{\mu}}_{k,\mathcal{D}} - \hat{\boldsymbol{\mu}}_{1,\mathcal{D}}) - (\boldsymbol{\mu}_{k,\mathcal{D}} - \boldsymbol{\mu}_{1,\mathcal{D}})\|_1 < \epsilon.$$

Lemma 8. *If both $A(\epsilon)$ and $B(\epsilon)$ have occurred for $\epsilon < \dfrac{1}{\varphi}$, we have*

$$\|\hat{\mathbf{\Sigma}}^{-1}_{\mathcal{D},\mathcal{D}} - \mathbf{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}\|_1 < \epsilon\varphi^2(1-\varphi\epsilon)^{-1},$$

$$\|\hat{\mathbf{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}(\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}})^{-1} - \mathbf{\Sigma}_{\mathcal{D}^c,\mathcal{D}}(\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}})^{-1}\|_\infty < \frac{\varphi\epsilon}{1-\varphi\epsilon}.$$

Proof of Lemma 8 . Let $\eta_1 = \|\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{D}}\|_\infty$, $\eta_2 = \|\hat{\mathbf{\Sigma}}_{\mathcal{D}^c,\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\|_\infty$ and $\eta_3 = \|(\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}})^{-1} - (\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}})^{-1}\|_\infty$. First we have

$$\eta_3 \le \|(\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}})^{-1}\|_\infty \times \|(\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D},\mathcal{D}})\|_\infty \times \|(\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}})^{-1}\|_\infty = (\varphi + \eta_3)\varphi\eta_1.$$

On the other hand,

$$\begin{aligned}
\|\hat{\mathbf{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}(\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}})^{-1} - \mathbf{\Sigma}_{\mathcal{D}^c,\mathcal{D}}(\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}})^{-1}\|_\infty &\le \|\hat{\mathbf{\Sigma}}_{\mathcal{D}^c,\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\|_\infty \times \|(\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}})^{-1} - (\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}})^{-1}\|_\infty \\
&\quad + \|\hat{\mathbf{\Sigma}}_{\mathcal{D}^c,\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\|_\infty \times \|(\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}})^{-1}\|_\infty \\
&\quad + \|\mathbf{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\|_\infty \times \|(\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}})^{-1} - (\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}})^{-1}\|_\infty \\
&\le \eta_2\eta_3 + \eta_2\varphi + \varphi\eta_3.
\end{aligned}$$

By $\varphi\eta_1 < 1$ we have $\eta_3 \le \varphi^2\eta_1(1-\varphi\eta_1)^{-1}$ and hence

$$\|\hat{\mathbf{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}(\hat{\mathbf{\Sigma}}_{\mathcal{D},\mathcal{D}})^{-1} - \mathbf{\Sigma}_{\mathcal{D}^c,\mathcal{D}}(\mathbf{\Sigma}_{\mathcal{D},\mathcal{D}})^{-1}\|_\infty < \frac{\varphi\epsilon}{1-\varphi\epsilon}.$$

$\square$

Lemma 9. *Define*

$$\hat{\boldsymbol{\theta}}^0_{k,\mathcal{D}} = \hat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{D},\mathcal{D}}(\hat{\boldsymbol{\mu}}_{k,\mathcal{D}} - \hat{\boldsymbol{\mu}}_{1,\mathcal{D}}). \tag{47}$$

Then $\|\hat{\boldsymbol{\theta}}^0_{k,\mathcal{D}} - \boldsymbol{\theta}_{k,\mathcal{D}}\|_1 \leq \dfrac{\varphi\epsilon(1+\varphi\Delta)}{1-\varphi\epsilon}$.

Proof of Lemma 9. By definition, we have

$$\begin{aligned}
&\|\hat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{D},\mathcal{D}}(\hat{\boldsymbol{\mu}}_{k,\mathcal{D}} - \hat{\boldsymbol{\mu}}_{1,\mathcal{D}}) - \boldsymbol{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}(\boldsymbol{\mu}_{k,\mathcal{D}} - \boldsymbol{\mu}_{1,\mathcal{D}})\|_1 \\
&\leq \|\hat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{D},\mathcal{D}} - \boldsymbol{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}\|_1\|(\hat{\boldsymbol{\mu}}_{k,\mathcal{D}} - \hat{\boldsymbol{\mu}}_{1,\mathcal{D}}) - (\boldsymbol{\mu}_{k,\mathcal{D}} - \boldsymbol{\mu}_{1,\mathcal{D}})\|_1 \\
&+\|\boldsymbol{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}\|_1\|(\hat{\boldsymbol{\mu}}_{k,\mathcal{D}} - \hat{\boldsymbol{\mu}}_{1,\mathcal{D}}) - (\boldsymbol{\mu}_{k,\mathcal{D}} - \boldsymbol{\mu}_{1,\mathcal{D}})\|_1 + \|\hat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{D},\mathcal{D}} - \boldsymbol{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}\|_1\|\boldsymbol{\mu}_{k,\mathcal{D}} - \boldsymbol{\mu}_{1,\mathcal{D}}\|_1 \\
\leq\ & \frac{\varphi\epsilon(1+\varphi\Delta)}{1-\varphi\epsilon}.
\end{aligned}$$

□

Lemma 10. *If $A(\epsilon)$ and $B(\epsilon)$ have occurred for $\epsilon < \min\{\frac{1}{2\varphi}, \dfrac{\lambda}{1+\varphi\Delta}\}$, then for all k*

$$\|\hat{\boldsymbol{\theta}}^{(\text{oracle})}_{k,\mathcal{D}}(\lambda) - \boldsymbol{\theta}_{k,\mathcal{D}}\|_\infty \leq 4\lambda\varphi.$$

Proof of Lemma 10. Observe $\hat{\boldsymbol{\theta}}^{\text{oracle}}_k = \hat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{D},\mathcal{D}}(\hat{\boldsymbol{\mu}}_{k,\mathcal{D}} - \hat{\boldsymbol{\mu}}_{1,\mathcal{D}}) - \lambda\hat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{D},\mathcal{D}}\hat{\mathbf{t}}_{k,\mathcal{D}}$. Therefore,

$$\begin{aligned}
&\|\hat{\boldsymbol{\theta}}^{\text{oracle}}_{k,\mathcal{D}} - \boldsymbol{\theta}_{k,\mathcal{D}}\|_\infty \\
\leq\ & \|\hat{\boldsymbol{\theta}}^0_{k,\mathcal{D}} - \boldsymbol{\theta}_{k,\mathcal{D}}\|_\infty + \lambda\|\hat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{D},\mathcal{D}} - \boldsymbol{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}\|_1\|\hat{\mathbf{t}}_{k,\mathcal{D}}\|_\infty + \lambda\|\boldsymbol{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}\|_1\|\hat{\mathbf{t}}_{k,\mathcal{D}}\|_\infty
\end{aligned}$$

where $\hat{\boldsymbol{\theta}}^0_{k,D}$ is defined as in (47). Now $\|\hat{\mathbf{t}}_{k,\mathcal{D}}\|_\infty \leq 1$ and we have

$$\|\hat{\boldsymbol{\theta}}^{\text{oracle}}_{k,\mathcal{D}} - \boldsymbol{\theta}_{k,\mathcal{D}}\|_\infty \leq \frac{\varphi\epsilon(1+\varphi\Delta)+\lambda\varphi}{1-\varphi\epsilon} < 4\varphi\lambda.$$

□

Lemma 11. *For a sets of real numbers $\{a_1,\ldots,a_N\}$, if $\sum_{i=1}^N a_i^2 \leq \kappa^2 < 1$, then $\sum_{i=1}^N (a_i+b)^2 < 1$ as long as $b < \dfrac{1-\kappa}{\sqrt{N}}$.*

Proof. By the Cauchy-Schwartz inequality, we have that

$$\begin{aligned}
\sum_{i=1}^{N}(a_i+b)^2 &= \sum_{i=1}^{N} a_i^2 + 2\sum_{i=1}^{N} a_i b + Nb^2 && (48)\\
&\le \sum_{i=1}^{N} a_i^2 + 2\sqrt{(\sum_{i=1}^{N} a_i^2)\cdot Nb^2} + Nb^2 && (49)\\
&\le \kappa^2 + 2\kappa\sqrt{Nb^2} + Nb^2 && (50)
\end{aligned}$$

which is less than 1 when $b < \dfrac{1-\kappa}{\sqrt{N}}$. $\square$

We are ready to complete the proof of Theorem 1.

Proof of Theorem 1. We first consider the first conclusion. For any $\lambda < \frac{\theta_{\min}}{8\varphi}$ and $\epsilon < \min\{\frac{1}{2\varphi}, \dfrac{\lambda}{1+\varphi\Delta}\}$, consider the event $A(\epsilon)\cap B(\epsilon)$. By Lemmas 4, 6 & 10 it suffices to verify (39).

For any $j \in \mathcal{D}^c$, by Lemma 5 we have

$$\begin{aligned}
& |(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{(\text{oracle})})_j - (\hat{\mu}_{kj}-\hat{\mu}_{1j})| \\
\le\ & |(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{(\text{oracle})})_j - (\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\boldsymbol{\theta}_{k,\mathcal{D}})_j| + |(\hat{\mu}_{kj}-\hat{\mu}_{1j}) - (\mu_{kj}-\mu_{1j})| \\
\le\ & |(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{(\text{oracle})})_j - (\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\boldsymbol{\theta}_{k,\mathcal{D}})_j| + \epsilon \\
\le\ & |(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{(0)})_j - (\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\boldsymbol{\theta}_{k,\mathcal{D}})_j| + \epsilon + \lambda|(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}\hat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}^{-1}\hat{\mathbf{t}}_{k,\mathcal{D}})_j|
\end{aligned}$$

$$\begin{aligned}
& |(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{(\text{oracle})})_j - (\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\boldsymbol{\theta}_{k,\mathcal{D}})_j| + \epsilon \\
\le\ & \|(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}})_j - (\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}})_j\|_1\|\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{0} - \boldsymbol{\theta}_{k,\mathcal{D}}\|_\infty + \|\boldsymbol{\theta}_{k,\mathcal{D}}\|_\infty\|(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}})_j - (\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}})_j\|_1 \\
& + \|(\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}})_j\|_1\|\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{0} - \boldsymbol{\theta}_{k,\mathcal{D}}\|_\infty + \epsilon \\
\le\ & C\epsilon. && (51)
\end{aligned}$$

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

$$
\begin{aligned}
|\hat{t}_{kj} - t_{kj}| &= \left|\frac{\hat{\theta}_{kj}\|\theta_{.j}\| - \theta_{kj}\|\hat{\theta}_{.j}\|}{\|\theta_{.j}\|\|\hat{\theta}_{.j}\|}\right| \\
&\leq \frac{|\hat{\theta}_{kj} - \theta_{kj}|\|\theta_{.j}\| + \theta_{\max}\|\theta_{.j} - \hat{\theta}_{.j}\|}{\|\theta_{.j}\|\|\hat{\theta}_{.j}\|} \\
&\leq \frac{C\varphi}{\theta_{\min}\sqrt{(K-1)}}\lambda.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \lambda|(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}\hat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{D},\mathcal{D}}\hat{\mathbf{t}}_{k,\mathcal{D}})_j| \\
\leq\ & \lambda|(\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\boldsymbol{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}\mathbf{t}_{k,\mathcal{D}})_j| + \lambda(\frac{C\varphi\epsilon}{1-\varphi\epsilon} + \eta^*\frac{C\varphi\lambda}{\theta_{\min}\sqrt{K-1}}) \quad (52) \\
\leq\ & \lambda|(\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\boldsymbol{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}\mathbf{t}_{k,\mathcal{D}})_j| + C\lambda^2 \quad (53)
\end{aligned}
$$

Under condition (C0), it follows from (51) and (53) that

$$
|(\hat{\boldsymbol{\Sigma}}_{\mathcal{D}^c,\mathcal{D}}\hat{\boldsymbol{\theta}}^{(\text{oracle})}_{k,\mathcal{D}})_j - (\hat{\mu}_{kj} - \hat{\mu}_{1j})| \leq \lambda|(\boldsymbol{\Sigma}_{\mathcal{D}^c,\mathcal{D}}\boldsymbol{\Sigma}^{-1}_{\mathcal{D},\mathcal{D}}\mathbf{t}_{k,\mathcal{D}})_j| + C\lambda^2 \quad (54)
$$

Combine condition (C0) with Lemma 11, we have that, there exists a generic constant $M > 0$, such that when $\lambda < M(1-\kappa)$, (39) is true. Therefore, the first conclusion is true.

Under conditions (C2)–(C4), the second conclusion directly follows from the first conclusion.

□

Lemma 12. *Under the conditions in Theorem 1, under $A(\epsilon) \cup B(\epsilon)$, we have that*

$$
\|\hat{\boldsymbol{\theta}}_k\|_1 \leq K(\Delta + \frac{\varphi\epsilon(1+\varphi\Delta)}{1-\varphi\epsilon}).
$$

Proof. Under the conditions in Theorem 1, we have that, under $A(\epsilon) \cup B(\epsilon)$, $\hat{\boldsymbol{\theta}}_k = (\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{\text{oracle}}, 0)$. It follows that

$$\begin{aligned} & \sum_{k=2}^{K}\{\frac{1}{2}(\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{\text{oracle}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{\text{oracle}} - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{\text{oracle}}\} + \lambda\sum_{j=1}^{p}\sqrt{\sum_{k=2}^{K}(\hat{\theta}_{kj}^{\text{oracle}})^2} \\ \leq & \sum_{k=2}^{K}\{\frac{1}{2}(\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{0})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{0} - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{0}\} + \lambda\sum_{j=1}^{p}\sqrt{\sum_{k=2}^{K}(\hat{\theta}_{kj}^{0})^2} \end{aligned}$$

while by the definition of $\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^0$, we must have

$$\frac{1}{2}(\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{\text{oracle}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{\text{oracle}} - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{\text{oracle}} \geq \frac{1}{2}(\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{0})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{D}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{0} - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{k,\mathcal{D}}^{0}$$

Hence,

$$\sum_{j=1}^{p}\sqrt{\sum_{k=2}^{K}(\hat{\theta}_{kj}^{\text{oracle}})^2} < \sum_{j=1}^{p}\sqrt{\sum_{k=2}^{K}(\hat{\theta}_{kj}^{0})^2} \leq \sum_{k=2}^{K}\|\hat{\boldsymbol{\theta}}_k^0\|_1 \leq K\Delta + K\frac{\varphi\epsilon(1+\varphi\Delta)}{1-\varphi\epsilon}$$

where the last inequality follows from Lemma 8. Finally, note that $\|\hat{\boldsymbol{\theta}}_k\|_1 \leq \sum_{j=1}^{p}\sqrt{\sum_{k=2}^{K}(\hat{\theta}_{kj}^{\text{oracle}})^2}$ and we have the desired conclusion. □

Proof of Theorem 2. We first show the first conclusion. Define $\hat{Y}(\boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_K)$ as the prediction by the Bayes rule and $\hat{Y}(\hat{\boldsymbol{\theta}}_2, \dots, \hat{\boldsymbol{\theta}}_K)$ as the prediction by the estimated classification rule. Also define $l_k = (\mathbf{X} - \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_1}{2})^{\mathrm{T}}\boldsymbol{\theta}_k + \log(\pi_k/\pi_1)$ and $\hat{l}_k = (\mathbf{X} - \frac{\hat{\boldsymbol{\mu}}_k + \hat{\boldsymbol{\mu}}_1}{2})^{\mathrm{T}}\hat{\boldsymbol{\theta}}_k + \log(\hat{\pi}_k/\hat{\pi}_1)$.

Define $C(\epsilon) = \{|\hat{\pi}_k - \pi_k| \leq \min\{\min_k \pi_k/2, \epsilon\}\}$. By the Bernstein inequality we have that $\Pr(C(\epsilon)) \leq C\exp(-Cn/K^2)$.

Assume that the event $A(\epsilon) \cap B(\epsilon) \cap C(\epsilon)$ for $\epsilon < \min\{\frac{1}{2\varphi}, \frac{\lambda}{1+\varphi\Delta}\}$ has happened. By Lemma 6, we have

$$\Pr(A(\epsilon)\cap B(\epsilon)\cap C(\epsilon)) \geq 1 - Cpd\exp(-Cn\frac{\epsilon^2}{Kd^2}) - CK\exp(-C\frac{n}{K^2}) - Cp(K-1)\exp(-Cn\frac{\epsilon^2}{K}) \tag{55}$$

For any $\epsilon_0 > 0$,

$$\begin{aligned}
R_n - R &\leq \Pr(\hat{Y}(\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K) \neq \hat{Y}(\hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_K)) \\
&\leq 1 - \Pr(|\hat{l}_k - l_k| < \epsilon_0/2, |l_k - l_{k'}| > \epsilon_0, \text{for any } k, k') \\
&\leq \Pr(|\hat{l}_k - l_k| \geq \epsilon_0/2 \text{ for some } k) + \Pr(|l_k - l_{k'}| \leq \epsilon_0 \text{ for some } k, k').
\end{aligned}$$

Now, for $\mathbf{X}$ in each class, $l_k - l_{k'}$ is normal with variance $(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k'})^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k''})$. Therefore,

$$\begin{aligned}
\Pr(|l_k - l_{k'}| \leq \epsilon_0 \text{ for some } k, k') &\leq \sum_{k''} \Pr(|l_k - l_{k'}| \leq \epsilon_0 \mid Y = k^{''})\pi_{k''} \\
&\leq \sum_{k,k',k''} \pi_{k''} \frac{C\epsilon_0}{\{(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k'})^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k'})\}^{1/2}} \\
&\leq CK^2\epsilon_0.
\end{aligned}$$

On the other hand, conditional on training data, $\hat{l}_k - l_k$ is normal with mean

$$u(k, k') = \boldsymbol{\mu}_{k'}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k) - \frac{(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}\hat{\boldsymbol{\theta}}_k}{2} + \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\theta}_k}{2} + \log \hat{\pi}_k/\hat{\pi}_1 - \log \pi_k/\pi_1$$

and variance $(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^{\mathrm{T}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)$ within class k'. By Markov's inequality, we have

$$\begin{aligned}
\Pr(|\hat{l}_k - l_k| \geq \epsilon_0/2 \text{ for some } k) &= \sum_{k'} \pi_{k'} \Pr(|\hat{l}_k - l_k| \geq \epsilon_0/2 \mid Y = k') \\
&\leq CE\{\frac{\max_k(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^{\mathrm{T}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)}{(\epsilon_0 - u(k, k'))^2}\}.
\end{aligned}$$

Moreover, under the event $A(\epsilon) \cap B(\epsilon) \cap C(\epsilon)$, by Lemma 12,

$$\begin{aligned}
\max_k(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^{\mathrm{T}}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k) &\leq \max_k \|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_1 \|\boldsymbol{\Sigma}\|_\infty \|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_\infty \\
&\leq \max_k(\|\hat{\boldsymbol{\theta}}_k\|_1 + \|\boldsymbol{\theta}_k\|_1)\|\boldsymbol{\Sigma}\|_\infty \|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_\infty \leq C\lambda \\
|u(k, k')| &\leq |\boldsymbol{\mu}_{k'}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)| + \frac{1}{2}|\{(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_k) - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_k)\}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)| \\
&\quad + \frac{1}{2}|\{(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_k) - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_k)\}^{\mathrm{T}}\boldsymbol{\theta}_k| + \frac{1}{2}|(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_k)^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)|
\end{aligned}$$

$$+|\log \hat{\pi}_k/\hat{\pi}_1 - \log \pi_k/\pi_1|$$
$$\leq \quad C_1\lambda$$

Hence, pick $\epsilon_0 = M_2\lambda^{1/3}$ such that $\epsilon_0 \geq C_1\lambda/2$, for C_1 in (56). Then $\Pr(|\hat{l}_k - l_k| \geq \epsilon_0/2 \text{ for some } k) \leq C\lambda^{1/3}$. It follows that $|R_n - R| \leq M_1\lambda^{1/3}$ for some positive constant M_1.

Under Conditions (C2)–(C4), the second conclusion is a direct consequence of the first conclusion. □

We need the result in the following proposition to show Lemma 3. A slightly different version of the proposition has been presented in Fukunaga (1990) (Pages 446-450), but we include the proof here for completeness.

Proposition 3. *The solution to* (4) *consists of all the right eigenvectors of* $\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_b$ *corresponding to positive eigenvalues.*

Proof. For any $\boldsymbol{\eta}_k$, set $\mathbf{u}_k = \mathbf{\Sigma}^{1/2}\boldsymbol{\eta}_k$. It follows that solving (4) is equivalent to finding

$$(\mathbf{u}_1^*, \ldots, \mathbf{u}_{K-1}^*) = \arg\max_{\mathbf{u}_k} \mathbf{u}_k^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\boldsymbol{\delta}_0\boldsymbol{\delta}_0^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\mathbf{u}_k, \text{ s.t. } \mathbf{u}_k^{\mathrm{T}}\mathbf{u}_k = 1 \text{ and } \mathbf{u}_k^{\mathrm{T}}\mathbf{u}_l = 0 \text{ for any } l < k. \tag{56}$$

and then setting $\boldsymbol{\eta}_k = \mathbf{\Sigma}^{-1/2}\mathbf{u}_k^*$. It is easy to see that $u_1^*, \ldots, u_{K-1}^*$ are the eigenvectors corresponding to positive eigenvalues of $\mathbf{\Sigma}^{-1/2}\boldsymbol{\delta}_0\boldsymbol{\delta}_0^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}$. By Proposition 4, let $\mathbf{A} = \mathbf{\Sigma}^{-1/2}\boldsymbol{\delta}_0\boldsymbol{\delta}_0^{\mathrm{T}}$, and $\mathbf{B} = \mathbf{\Sigma}^{-1/2}$ and we have that $\boldsymbol{\eta}$ consists of all the eigenvectors of $\mathbf{\Sigma}^{-1}\boldsymbol{\delta}_0\boldsymbol{\delta}_0^{\mathrm{T}}$ corresponding to positive eigenvalues. □

Proposition 4. *(Mardia et al. (1979), Page 468, Theorem A.6.2) For two matrices* $\mathbf{A}$ *and* $\mathbf{B}$*, if* $\mathbf{x}$ *is a non-trivial eigenvector of* $\mathbf{AB}$ *for a nonzero eigenvalue, then* $\mathbf{y} = \mathbf{Bx}$ *is a non-trivial eigenvector of* $\mathbf{BA}$.

Proof of Lemma 3. Set $\tilde{\boldsymbol{\delta}} = (0_p, \boldsymbol{\delta})$ and $\boldsymbol{\delta}_0 = (\boldsymbol{\mu}_1 - \bar{\boldsymbol{\mu}}, \ldots, \boldsymbol{\mu}_K - \bar{\boldsymbol{\mu}})$. Note that $\boldsymbol{\delta} 1_K = \sum_{k=2}^{K} \boldsymbol{\mu}_k - (K-1)\boldsymbol{\mu}_1 = K(\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)$. Therefore, $\boldsymbol{\delta}_0 = \tilde{\boldsymbol{\delta}} - \frac{1}{K}\tilde{\boldsymbol{\delta}} 1_K 1_K^{\mathrm{T}} = \tilde{\boldsymbol{\delta}}(\mathbf{I}_K - \frac{1}{K} 1_K 1_K^{\mathrm{T}}) = \tilde{\boldsymbol{\delta}}\boldsymbol{\Pi}$.

Then, since $\boldsymbol{\theta}_0 = \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\delta}}$, we have $\boldsymbol{\theta}_0\boldsymbol{\Pi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0$ and $\boldsymbol{\theta}_0\boldsymbol{\Pi}\boldsymbol{\delta}_0^{\mathrm{T}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_0\boldsymbol{\delta}_0^{\mathrm{T}}$. By Proposition 3, we have the desired conclusion. □

References

Bach, F. R. (2008), ‘Consistency of the group lasso and multiple kernel learning’, *Journal of Machine Learning Research* **9**, 1179–1225.

Bickel, P. J. & 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.

Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., Maganti, V., Reddy, P. S., Strahs, A., Immermann, F., Spinelli, W., Schwertschlag, U., Slager, A. M., Cotreau, M. M. & Dorner, A. J. (2006), ‘Molecular classification of crohn’s disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells’, *Journal of Molecular Diagnostics* **8**, 51–61.

Cai, T. & Liu, W. (2011), ‘A direct estimation approach to sparse linear discriminant analysis’, *J. Am. Statist. Assoc.* **106**, 1566–1577.

Candes, E. & Tao, T. (2007), ‘The Dantzig selector: Statistical estimation when p is much larger than n’, *Ann. Statist.* **35**, 2313–2351.

Clemmensen, L., Hastie, T., Witten, D. & Ersbøll, B. (2011), ‘Sparse discriminant analysis’, *Technometrics* **53**, 406–413.

Donoho, D. & Jin, J. (2008), ‘Higher criticism thresholding: optimal feature selection when useful features are rare and weak’, *Proceedings of the National Academy of Sciences* **105**, 14790–14795.

Fan, J. & Fan, Y. (2008), 'High dimensional classification using features annealed independence rules', *Ann. Statist.* **36**, 2605–2637.

Fan, J., Feng, Y. & Tong, X. (2012), 'A ROAD to classification in high dimensional space', *J. R. Statist. Soc. B* **74**, 745–771.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *J. Am. Statist. Assoc.* **96**, 1348–1360.

Fan, J. & Song, R. (2010), 'Sure independence screening in generalized linear models with NP-dimensionality', *Ann. Statist.* **38**(6), 3567–3604.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, Academic Press Professional, Inc., 2nd Edition.

Hand, D. J. (2006), 'Classifier technology and the illusion of progress', *Statistical Science* **21**, 1–14.

Hastie, T. J., Tibshirani, R. J. & Friedman, J. H. (2009), *Elements of statistical learning: data mining, inference, and prediction*, second edn, Springer Verlag.

Mai, Q. & Zou, H. (2013*a*), 'The Kolmogorov filter for variable screening in high-dimensional binary classification.', *Biometrika* **100**, 229–234.

Mai, Q. & Zou, H. (2013*b*), 'A note on the connection and equivalence of three sparse linear discriminant analysis methods', *Technometrics* **55**, 243–246.

Mai, Q., Zou, H. & Yuan, M. (2012), 'A direct approach to sparse discriminant analysis in ultra-high dimensions', *Biometrika* **99**, 29–42.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press.

Michie, D., Spiegelhalter, D. & Taylor, C. (1994), *Machine Learning, Neural and Statistical Classification*, first edn, Ellis Horwood.

Shao, J., Wang, Y., Deng, X. & Wang, S. (2011), 'Sparse linear discriminant analysis with high dimensional data', *Ann. Statist.* .

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *J. R. Statist. Soc. B* **58**, 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc. Nat. Acad. Sci.* **99**, 6567–6572.

Trendafilov, N. T. & Jolliffe, I. T. (2007), 'DALASS: Variable selection in discriminant analysis via the lasso', *Computational Statistics and Data Analysis* **51**, 3718–3736.

Witten, D. & Tibshirani, R. (2011), 'Penalized classification using fisher's linear discriminant', *J. R. Statist. Soc. B* **73**, 753–772.

Wu, M., Zhang, L., Wang, Z., Christiani, D. & Lin, X. (2008), 'Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection', *Bioinformatics* **25**, 1145–1151.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *J. R. Statist. Soc. B* **68**, 49–67.