A Locally Adaptive Shrinkage Approach to False Selection Rate Control in High-Dimensional Classification

Bowen Gang[1], Yuantao Shi[2] and Wenguang Sun[3]

[1] *Fudan University,* [2] *University of Chicago and* [3] *Zhejiang University*

Supplementary Material

This supplement contains the proofs of main theorems, propositions and corollaries (Section S1), proofs of technical lemmas (Section S2), an argument establishing the asymptotic equivalence of FSR and mFSR (Section S3), numerical illustrations of the effects of shrinkage factor (Section S4), additional simulation result on a "boarderline sparse" setting (Section S5), additional numerical results on a lung cancer data (Section S6), an example showing the advantage LASS has over LPD (Section S7) and a proof showing class-specific FSR control implies global FSR control (Section S8).

S1 Proofs of main theorems, propositions and corollaries

We swap the proofs of Theorems 4 and 3 as the latter is simpler. Other proofs are arranged according to the orders in the text.

S1.1 Proof of Theorem 1

We only need to show $\delta^j_{OR} = 2\mathbb{I}(T^j < t^2_{OR})$ controls mFSR^2 at level α_2 and maximizes $\mathbb{E}\{\sum_{i=1}^m \mathbb{I}(\theta_j = 2, \delta_j = 2)\}$. The same argument can be used to show that $\delta^j_{OR} = \mathbb{I}(1-T^j < t^1_{OR})$ controls mFSR^1 at level α_1 and maximizes $\mathbb{E}\{\sum_{i=1}^m \mathbb{I}(\theta_j = 1, \delta_j = 1)\}$. The theorem is proved when we combine these two statements.

The proof is divided into two parts. In part (a), we establish two properties of the classification rule $\boldsymbol{\delta}^2(t) = \{2\mathbb{I}(T^j < t) : 1 \le j \le m\}$. We show that it produces $\mathrm{mFSR}^2 < t$ for all $t \in (0,1)$ and that its mFSR^2 is monotonic in t. In part (b) we show that when the threshold is t^2_{OR}, the classification rule has $\mathrm{mFSR}^2 = \alpha_2$ and maximizes $\mathbb{E}\{\sum_{i=1}^m \mathbb{I}(\theta_j = 2, \delta_j = 2)\}$ amongst all valid classification rule with $\mathrm{mFSR}^2 \le \alpha_2$.

Part(a). Consider classification rule $\{2 \cdot \mathbb{I}(T^j < t) : 1 \le j \le m\}$. Let $Q^2(t) = \alpha_t$ be its the mFSR^2 level. We first show that $\alpha_t < t$. Since $T^j = P(\theta_j = 1|\boldsymbol{W}_j)$, then

$$\mathbb{E}\left\{\sum_{j=1}^m \mathbb{I}(\theta_j = 1, \delta_j = 2)\right\} = \mathbb{E}_{\boldsymbol{W}}\left[\left\{\sum_{j=1}^m \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{W}}\mathbb{I}(\theta_j = 1, \delta_j = 2)\right\}\right] = \mathbb{E}_{\boldsymbol{W}}\left\{\sum_{i=1}^m T^j \mathbb{I}(\delta_j = 2)\right\},$$

where $\mathbb{E}$ is the expectation over $(\boldsymbol{W}, \boldsymbol{\theta})$, E_W is the expectation over $\boldsymbol{W}$, and $E_{\boldsymbol{\theta}|\boldsymbol{W}}$ is the expectation over $\boldsymbol{\theta}$ holding $\boldsymbol{W}$ fixed. Recall $Q^2_{OR}(t) = \alpha_t$ is the

FSR^2 level of the classification rule $\{2 \cdot \mathbb{I}(T^j < t) : 1 \le i \le m\}$, we have

$$\mathbb{E}_{\boldsymbol{W}} \left\{ \sum_{j=1}^{m} (T^j - \alpha_t) \mathbb{I}(T^j < t) \right\} = 0. \tag{S1.1}$$

This implies that $\alpha_t < t$. To see this, if $\alpha_t \ge t$, then $(T^j - \alpha_t)\mathbb{I}(T^j < t) < 0$, which contradicts the right hand side.

Next, we show that $Q^2(t)$ is nondecreasing in t. That is, letting $Q^2(t_j) = \alpha_{t_j}$, if $t_1 < t_2$, then $\alpha_{t_1} \le \alpha_{t_2}$. We argue by contradiction. Suppose that $t_1 < t_2$ but $\alpha_{t_1} > \alpha_{t_2}$. Then

$$\begin{aligned} & (T^j - \alpha_{t_2})\mathbb{I}(T^j < t_2) \qquad\qquad (\text{S1.2}) \\ =& (T^j - \alpha_{t_1})\mathbb{I}(T^j < t_1) + (\alpha_{t_1} - \alpha_{t_2})\mathbb{I}(T^j < t_1) + (T^j - \alpha_{t_2})\mathbb{I}(t_1 \le T^j < t_2) \\ \ge& (T^j - \alpha_{t_1})\mathbb{I}(T^j < t_1) + (\alpha_{t_1} - \alpha_{t_2})\mathbb{I}(T^j < t_1) + (T^j - \alpha_{t_1})\mathbb{I}(t_1 \le T^j < t_2). \end{aligned}$$

By (S1.1) we have $\mathbb{E}\left\{\sum_{j=1}^{m}(T^j - \alpha_{t_1})\mathbb{I}(T^j < t_1)\right\} = 0$, together with the fact that $\alpha_{t_1} < t_1$ we have $\mathbb{E}\left\{\sum_{j=1}^{m}(T^j - \alpha_{t_2})\mathbb{I}(T^j < t_2)\right\} > 0$, contradicting (S1.1).

Part(b). Define $t^2_{OR} = \sup_t\{t \in (0,1) : Q^2(t) \le \alpha\}$. By part (a), $Q^2(t)$ is non–decreasing in t. By continuity, $Q^2(T) = \alpha_2$. Next, consider the oracle rule $\boldsymbol{\delta}^2_{OR} = (\delta^{2,1}_{OR}, \ldots, \delta^{2,m}_{OR}) = \{2\mathbb{I}(T^j < t^2_{OR}) : \ 1 \le j \le m\}$ and an arbitrary rule $\boldsymbol{\delta}_* = (\delta^1_*, \cdots, \delta^m_*)$ such that $\text{mFSR}^2_{\boldsymbol{\delta}_*} \le \alpha_2$. Using the result in part

(a), we have

$$\mathbb{E}\left\{\sum_{j=1}^{m}(T^j-\alpha)\mathbb{I}(\delta_{OR}^{2,j}=2)\right\}=0 \quad \text{and} \quad \mathbb{E}\left\{\sum_{j=1}^{m}(T^j-\alpha)\mathbb{I}(\delta_{*}^{j}=2)\right\}\leq 0. \tag{S1.3}$$

Taking the difference of the two equations in (S1.3), we have

$$\mathbb{E}\left\{\sum_{j=1}^{m}(T^j-\alpha)\mathbb{I}(\delta_{OR}^{2,j}=2)-(T^j-\alpha)\mathbb{I}(\delta_{*}^{j}=2)\right\}\geq 0. \tag{S1.4}$$

Next consider the transformation $f(x)=(x-\alpha)/(1-x)$. Note that $f'(x)=(1-\alpha)/(1-x)^2>0$, $f(x)$ is monotonically increasing, the order is preserved by this transformation: if $T^i<t_{OR}^2$ then $f(T^i)<f(t_{OR}^2)$. This means we can rewrite the oracle rule as

$$\delta_{OR}^{2,i}=2\mathbb{I}\left[\left\{(T^i-\alpha)/(1-T^i)\right\}<\lambda_{OR}^2\right],$$

where $\lambda_{OR}^2=(t_{OR}^2-\alpha)/(1-t_{OR}^2)$. It will be useful to note that, from part (a), we have $\alpha_{t_2}<t_{OR}^2<1$, which implies that $\lambda_{OR}^2>0$. Note that

$$\mathbb{E}\left[\sum_{j=1}^{m}\left\{\mathbb{I}(\delta_{OR}^{2,j}=2)-\mathbb{I}(\delta_{*}^{j}=2)\right\}\left\{(T^j-\alpha)-\lambda_{OR}^2(1-T^j)\right\}\right]\leq 0 \tag{S1.5}$$

To see this, consider that if $\mathbb{I}(\delta_{OR}^{2,j}=2)-\mathbb{I}(\delta_{*}^{j}=2)\neq 0$, then either (i) $\mathbb{I}(\delta_{OR}^{2,j}=2)-\mathbb{I}(\delta_{*}^{j}=2)>0$ or (ii) $\mathbb{I}(\delta_{OR}^{2,j}=2)-\mathbb{I}(\delta_{*}^{j}=2)<0$ holds. If

(i) holds, then $\delta_{OR}^{2,j} = 2$ and it follows that $\{(T^j - \alpha)/(1 - T^j)\} < \lambda_{OR}^2$. If (ii) holds, then $\delta_{OR}^{j} \neq 2$ and $\{(T^j - \alpha)/(1 - T^j)\} \geq \lambda_{OR}^2$. In both cases, we have

$$\left\{\mathbb{I}(\delta_{OR}^{2,j} = 2) - \mathbb{I}(\delta_*^j = 2)\right\}\left\{(T^j - \alpha) - \lambda_{OR}^2(1 - T^i)\right\} \leq 0.$$

Summing over all m terms and taking the expectation yields (S1.5).

Combining (S1.4) and (S1.5), we obtain

$$0 \leq \lambda_{OR}^2 \mathbb{E}\left[\sum_{j=1}^{m}\left\{\mathbb{I}(\delta_{OR}^{2,j} = 2) - \mathbb{I}(\delta_*^j = 2)\right\}(1 - T^j)\right]$$

Finally, since $\lambda_{OR}^2 > 0$, it follows that

$$\mathbb{E}\left[\sum_{i=1}^{m}\left\{\mathbb{I}(\delta_{OR}^{j} = 2) - \mathbb{I}(\delta_*^j = 2)\right\}(1 - T^j)\right] \geq 0.$$

□

S1.2 Proof of Theorem 2

As in the proof of Theorem 1, we will show $\boldsymbol{\delta}_{OR}^{*2}$=$\{\delta_{OR}^{*2,j} = 2\mathbb{I}(T^j < \hat{t}_{OR}^2),\ j = 1, \ldots, m\}$ controls mFSR^2 and FSR^2 at level α_2. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. Note

that

$$\mathrm{FSR}^2 = \mathbb{E}\left[\frac{\sum_{j=1}^m \delta_{OR}^{*2,j}\mathbb{I}(\theta_j \neq 2)}{\{\sum_{j=1}^m \delta_{OR}^{*2,j}\} \vee 1}\right] = \mathbb{E}_{\boldsymbol{\delta}_{OR}^{*2}}\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{\delta}_{OR}^{*2}}\left[\frac{\sum_{j=1}^m \delta_{OR}^{*2,j}\mathbb{I}(\theta_j \neq 2)}{\{\sum_{j=1}^m \delta_{OR}^{*2,j}\} \vee 1}\middle|\boldsymbol{\delta}_{OR}^{*2}\right].$$

By definition of the rule $\boldsymbol{\delta}_{OR}^{*2}$, if $\sum_{j=1}^m \delta_{OR}^{*2,j} = k \geq 1$ then

$$\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{\delta}_{OR}^{*2}} \sum_{j=1}^m \delta_{OR}^{*2,j}\mathbb{I}(\theta_j \neq 2) = \sum_{i=1}^m \mathbb{P}(\theta_j \neq 2)\delta_{OR}^{*2,j} \leq k\alpha_2.$$

It follows that $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{\delta}_{OR}^{*2}}\left[\frac{\sum_{j=1}^m \delta_{OR}^{*2,j}\mathbb{I}(\theta_j \neq 2)}{\{\sum_{j=1}^m \delta_{OR}^{*2,j}\} \vee 1}\middle|\boldsymbol{\delta}_{OR}^{*2}\right] \leq \alpha$, and $\mathrm{FSR}^2 \leq \alpha_2$.

Let $\mathcal{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_m)$. For mFSR^2 we have

$$\begin{aligned}
\frac{\mathbb{E}\left(\sum_{j=1}^m \mathbb{I}(\theta_j \neq 2)\delta_{OR}^{*2,j}\right)}{\mathbb{E}\left(\sum_{j=1}^m \delta_{OR}^{*2,j}\right)} &= \frac{\mathbb{E}_{\mathcal{W}}\mathbb{E}_{\boldsymbol{\theta}|\mathcal{W}}\left\{\sum_{j=1}^m \mathbb{I}(\theta_j \neq 2)\delta_{OR}^{*2,j}\right\}}{\mathbb{E}\left(\sum_{j=1}^m \delta_{OR}^{*2,j}\right)} \\
&= \frac{\mathbb{E}_{\mathcal{W}}\mathbb{E}_{\boldsymbol{\theta}|\mathcal{W}}\left\{(\sum_{j=1}^m \delta_{OR}^{*2,j})\left(\sum_{j=1}^m \mathbb{I}(\theta_j \neq 2)\delta_{OR}^{*2,j}/\sum_{j=1}^m \delta_{OR}^{*2,j}\right)\right\}}{\mathbb{E}\left(\sum_{j=1}^m \delta_{OR}^{*2,j}\right)} \\
&= \frac{\mathbb{E}_{\mathcal{W}}\left(\sum_{j=1} \delta_{OR}^{*2,j}\right)\mathbb{E}_{\mathcal{W}}\left\{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{W}}\left(\sum_{j=1}^m \mathbb{I}(\theta_j \neq 2)\delta_{OR}^{*2,j}/\sum_{j=1} \delta_{OR}^{*2,j}\right)\right\}}{\mathbb{E}_{\mathcal{W}}\left(\sum_{j=1}^m \delta_{OR}^{*2,j}\right)} \\
&= \mathbb{E}_{\mathcal{W}}\left\{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{W}}\left(\sum_{j=1}^m \mathbb{I}(\theta_j \neq 2)\delta_{OR}^{*2,j}/\sum_{j=1} \delta_{OR}^{*2,j}\right)\right\}.
\end{aligned}$$

The fourth equality holds because $\boldsymbol{\delta}_{OR}^{*2}$ depends on $\boldsymbol{\theta}$ only through $\mathcal{W}$. Now

given $\mathcal{W}$ and that $\sum_i \delta_{OR}^{*2,j} = k$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{W}}\left\{E_{\boldsymbol{\theta}|\mathcal{W}}\left(\sum_{j=1}^{m}\mathbb{I}(\theta_j \neq 2)\delta_{OR}^{*2,j}/\sum_{j=1}^{m}\delta_{OR}^{*2,j}\right)\right\} &= \mathbb{E}_{\mathcal{W}}\left\{\frac{\mathbb{E}_{\boldsymbol{\theta}|\mathcal{W}}\left(\sum_{j=1}^{m}\mathbb{I}(\theta_j \neq 2)\right)\delta_{OR}^{*2,j}}{k}\right\} \\ &= \mathbb{E}_{\mathcal{W}}\left(\frac{\sum_{j=1}^{m}T^j\delta_{OR}^{*2,j}}{k}\right).\end{aligned}$$

By the definition of $\boldsymbol{\delta}_{OR}^{*2}$, we have

$$\mathbb{E}_{\mathcal{W}}\left(\frac{\sum_{j=1}^{m}T^j\delta_{OR}^{*2,j}}{k}\right) \leq \alpha_2.$$

Hence $\boldsymbol{\delta}_{OR}^{*2}$ satisfies $\text{mFSR}^2 \leq \alpha_2$. □

S1.3 Proof of Proposition 1

Let $q_k(x) := \tilde{g}_{1k}(|x|)/\{g_0(|x|) + \tilde{g}_{1k}(|x|)\}$, where g_0 and $\tilde{g}_{1k}$ are the density function of $\mathcal{N}\left(0, \frac{n_1+n_2}{n_1n_2}\right)$ and $\mathcal{N}\left(\left\{(2+b)\sqrt{\sigma_{kk}} + \sqrt{(2+b)^2\sigma_{kk}+4}\right\}\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}, \frac{n_1+n_2}{n_1n_2}\right)$ respectively. We will assume without loss of generality that $d_k > 0$.

We shall first prove the result when the true variance σ_{kk} is known and then argue that with probability greater than $1 - pe^{-O(n)}$ the result still holds when σ_{kk} is replaced by $\hat{\sigma}_{kk}$. Let g_{2k} be the density function of $\mathcal{N}\left(d_k, \sigma_{kk}\frac{n_1+n_2}{n_1n_2}\right)$. Assume d_k is strong, we investigate the asymptotic

behavior of

$$
\begin{aligned}
1-\mathbb{E}\{q_k(X)|X\sim g_{2k}\} &= \int_{-\infty}^{\infty}\frac{g_0(|x|)g_{2k}(x)}{g_0(|x|)+\tilde{g}_{1k}(|x|)}dx \qquad \text{(S1.6)}\\
&= \int_{-\infty}^{d_k-\frac{\epsilon}{4}\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}}\frac{g_0(|x|)g_{2k}(x)}{g_0(|x|)+\tilde{g}_{1k}(|x|)}dx\\
&\quad+\int_{d_k-\frac{\epsilon}{4}\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}}^{\infty}\frac{g_0(|x|)g_{2k}(x)}{g_0(|x|)+\tilde{g}_{1k}(|x|)}dx.
\end{aligned}
$$

Consider the first term, it is easy to see that

$$
\int_{-\infty}^{d_k-\frac{\epsilon}{4}\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}}\frac{g_0(|x|)g_{2k}(x)}{g_0(|x|)+\tilde{g}_{1k}(|x|)}dx\le\int_{-\infty}^{d_k-\frac{\epsilon}{4}\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}}g_{2k}(x)dx=O(p^{-\epsilon^2/(64\sigma_{kk})}).
$$

Since $d_k\in\mathcal{G}_1$, we have $x>(a_k/2+3\epsilon/4)\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}$ if $\left(d_k-(\epsilon/4)\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p},\infty\right)$; on this interval we have

$$
\begin{aligned}
\frac{\tilde{g}_{1k}(|x|)}{g_0(|x|)} &= \exp\left(\frac{2a_k|x|\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}-a_k^2\frac{(n_1+n_2)}{2n_1n_2}\log p}{2\frac{n_1+n_2}{n_1n_2}}\right)\\
&\ge \exp\left(\frac{1}{2}a_k(a_k/2+3\epsilon/4)\log p-\frac{1}{4}a_k^2\log p\right)\\
&\ge \exp\left(\frac{3a_k\epsilon}{8}\log p\right)\\
&\ge p^{3a_k\epsilon/8}.
\end{aligned}
$$

It follows that

$$\begin{aligned}\int_{(d_k-\epsilon/4)\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}}^{\infty} \frac{g_0(|x|)g_{2k}(x)}{g_0(|x|)+\tilde{g}_{1k}(|x|)}dx &\leq \sup_{x>(d_k-\epsilon/4)\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}} \frac{1}{1+\tilde{g}_{1k}(x)/g_0(x)} \\ &\leq p^{-3a_k\epsilon/8}.\end{aligned}$$

Using (S1.6), we have $\mathbb{E}(q_k|\theta_k = 1) = O(p^{-\epsilon_1})$, where $\epsilon_1 > 0$ is some constant.

Next, when $d_k \in \mathcal{G}_3$, we can similarly show that

$$\mathbb{E}\{q_k(X)|X \sim g_{2k}\} = \left|1 - \int_{-\infty}^{\infty} \frac{g_0(|x|)g_{2k}(x)}{g_0(|x|)+\tilde{g}_{1k}(|x|)}dx\right| = O(p^{-\epsilon_2}).$$

Finally, consider the the case when $d_k \in \mathcal{G}_2$. We have

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{g_0(|x|)g_{2k}(x)}{g_0(|x|)+\tilde{g}_{1k}(|x|)}dx &= \int_{-\infty}^{\infty} \frac{g_{2k}(x)}{1+\exp\left(\frac{2a_k\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}|x| - {a_k}^2\frac{(n_1+n_2)}{2n_1n_2}\log p}{2\frac{n_1+n_2}{n_1n_2}}\right)}dx \\ &= \int_{-\infty}^{\infty} \frac{g_{2k}(x)}{1+\exp\left(\frac{1}{\sqrt{2}}a_k|x|\sqrt{\frac{n_1n_2}{n_1+n_2}\log p} - \frac{{a_k}^2}{4}\log p\right)}dx.\end{aligned}$$

Let $\mathcal{A}_t = \left\{x : -t/\sqrt{\frac{2n_1n_2}{n_1+n_2}} + d_k \le x \le t/\sqrt{\frac{2n_1n_2}{n_1+n_2}} + d_k\right\}$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{g_{2k}(x)}{1+\exp\left(\frac{1}{\sqrt{2}}a_k|x|\sqrt{\frac{n_1n_2}{n_1+n_2}\log p} - \frac{{a_k}^2}{4}\log p\right)} - g_{2k}(x)dx \\
=&\int_{x\in\mathcal{A}_t^c} \frac{g_{2k}(x)}{1+\exp\left(\frac{1}{\sqrt{2}}a_k|x|\sqrt{\frac{n_1n_2}{n_1+n_2}\log p} - \frac{{a_k}^2}{4}\log p\right)} - g_{2k}(x)dx \\
+&\int_{x\in\mathcal{A}_t} \frac{g_{2k}(x)}{1+\exp\left(\frac{1}{\sqrt{2}}a_k|x|\sqrt{\frac{n_1n_2}{n_1+n_2}\log p} - \frac{{a_k}^2}{4}\log p\right)} - g_{2k}(x)dx.
\end{aligned}$$

Some algebra shows that

$$\begin{aligned}
&\left|\int_{x\in\mathcal{A}_t^c} \frac{g_{2k}(x)}{1+\exp\left(\frac{1}{\sqrt{2}}a_k|x|\sqrt{\frac{n_1n_2}{n_1+n_2}\log p} - \frac{{a_k}^2}{4}\log p\right)} - g_{2k}(x)dx\right| \\
\le\ & \int_{x\in\mathcal{A}_t^c} g_{2k}(x)dx = O\left(e^{-\frac{t^2}{4\sigma_{kk}}}\right), \text{ and} \qquad \text{(S1.7)} \\
&\left|\int_{x\in\mathcal{A}_t} \frac{g_{2k}(x)}{1+\exp\left(\frac{1}{\sqrt{2}}a_k|x|\sqrt{\frac{n_1n_2}{n_1+n_2}\log p} - \frac{{a_k}^2}{4}\log p\right)}dx\right| \\
\le\ & \left|\int_{x\in\mathcal{A}_t} \frac{\phi_{2k}(x)}{1+\exp\left(\frac{1}{2}a_kt\sqrt{\log p} + o(\log p) - \frac{a_k^2}{4}\log p\right)} - g_{2k}(x)dx\right|. \qquad \text{(S1.8)}
\end{aligned}$$

Take $t = (2+b/2)\sqrt{\sigma_{kk}\log p}$, we can see that (S1.7) is bounded by $o(1/p^{1+b})$. For (S1.8) to be bounded by $o(1/p^{1+a})$ for some constant $a > 0$, we need ${a_k}^2/4 - (1+b/4)a_k\sqrt{\sigma_{kk}} > 1+c$ for some constant $c > 0$. Some computation shows that $a_k > (2+b/2)\sqrt{\sigma_{kk}} + \sqrt{(2+b/2)^2\sigma_{kk}+4}$ can satisfy both requirements. Since we take $a_k = (2+b)\sqrt{\sigma_{kk}} + \sqrt{(2+b)^2\sigma_{kk}+4}$ we have

(S1.8) is bounded by:

$$\left|\int_{x\in\mathcal{A}_t}\frac{g_{2k}(x)}{1+o(1/p^{1+\epsilon_3})}-g_{2k}(x)dx\right|=O(1/p^{1+\epsilon_3})\int g_{2k}(x)dx=O(1/p^{1+\epsilon_3}),$$

where $\epsilon_3>0$ is some constant. Use the fact that $\int\phi_{2k}dx=1$ we conclude that

$$\mathbb{E}\{q_k(X)|X\sim g_{2k}\}=\left|1-\int_{-\infty}^{\infty}\frac{g_0(|x|)g_{2k}(x)}{g_0(|x|)+\tilde{g}_{1k}(|x|)}dx\right|=O(1/p^{1+\epsilon_3}).$$

Take $\gamma=\min(\epsilon_1,\epsilon_2,\epsilon_3)$, the results follow.

Define $\hat{a}_k=(2+b)\sqrt{\hat{\sigma}_{kk}}+\sqrt{(2+b)^2\hat{\sigma}_{kk}+4}$. From the above proof, we can see that if we would like the results to hold when we replace a_k by $\hat{a}_k$, we need $x>(\hat{a}_k/2+3\epsilon'/4)\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}$ on the interval $(d_k-(\epsilon/4)\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p},\infty)$ when d_k is strong and $\hat{a}_k>(2+b/2)\sqrt{\sigma_{kk}}+\sqrt{(2+b/2)^2\sigma_{kk}+4}$ when d_i is weak. Here ϵ' is some positive constant less than ϵ and depends solely on ϵ. As a_k satisfies these two conditions, we can choose a constant d which depends solely on ϵ' and b such that $\hat{a}_k$ satisfies these two conditions when $|\hat{a}_k-a_k|<d$.

When the true variance is unknown, we use the pooled sample variance

$$\hat{\sigma}_{kk}=\frac{n_1-1}{n_1+n_2-2}\sum_{i=1}^{n_1}(X_{ik}-\bar{X}_k)^2+\frac{n_2-1}{n_1+n_2-2}\sum_{i=1}^{n_2}(Y_{ik}-\bar{Y}_k)^2$$

to estimate σ_{kk}, and use $\hat{a}_k$ to estimate a_k. Since $\hat{a}_k$ is a continuous function of $\hat{\sigma}_{kk}$ and $\hat{a}_k = a_k$ when $\hat{\sigma}_{kk} = \sigma_{kk}$, we can choose a constant λ small enough such that $|\hat{a}_k - a_k| < d$ when $|\hat{\sigma}_{kk} - \sigma_{kk}| < \epsilon_0^{-1}\lambda$. (Recall that ϵ_0^{-1} is the upper bound of σ_{kk}.) The constant λ here depends solely on d and ϵ_0, which are fixed and independent of n_1, n_2 and p. Recall that since X_{ik} are $i.i.d$ for $1 \le i \le n_1$ and Y_{ik} are $i.i.d$ for $1 \le i \le n_2$, we have $(n_1+n_2-2)\hat{\sigma}_{kk}/\sigma_{kk} \sim \chi^2_{n_1+n_2-2}$. Lemma 1 in Foygel and Drton (2010) and Lemma 4 in Cai (2002) proved the following concentration inequality for χ^2 random variable. For $n \ge 4\lambda^{-2}+1$, we have

$$\begin{aligned} P\left\{\chi^2_n > n(1+\lambda)\right\} &\le \frac{1}{\lambda\sqrt{\pi n}} e^{-\frac{n}{2}(\lambda - \log(1+\lambda))}; \\ P\left\{\chi^2_n < n(1-\lambda)\right\} &\le \frac{1}{\lambda\sqrt{\pi(n-1)}} e^{-\frac{n-1}{2}(\lambda + \log(1-\lambda))}. \end{aligned}$$

Noting that λ is a constant independent of n_1, n_2 and p, we apply the above results to conclude that

$$\begin{aligned} & P\left(\left|\frac{\hat{\sigma}_{kk}}{\sigma_{kk}} - 1\right| < \lambda\right) \\ \ge & 1 - \frac{1}{\lambda\sqrt{\pi(n_1+n_2-2)}} e^{-\frac{n_1+n_2-2}{2}(\lambda - \log(1+\lambda))} - \frac{1}{\lambda\sqrt{\pi(n_1+n_2-3)}} e^{-\frac{n_1+n_2-3}{2}(\lambda + \log(1-\lambda))} \\ = & 1 - e^{-\frac{n_1+n_2-2}{2}(\lambda - \log(1+\lambda)) + O(\log(n_1+n_2))} - e^{-\frac{n_1+n_2-3}{2}(\lambda + \log(1-\lambda)) + O(\log(n_1+n_2))}. \\ = & 1 - e^{-O(n_1+n_2)} \end{aligned}$$

Since $P\left(\bigcup_{i=1}^{p}|\hat{\sigma}_{kk}-\sigma_{kk}|>\epsilon_0^{-1}\lambda\right)\leq\sum_{i=1}^{p}P\left(|\hat{\sigma}_{kk}-\sigma_{kk}|>\epsilon_0^{-1}\lambda\right)\leq\sum_{i=1}^{p}P\left(|\hat{\sigma}_{kk}-\sigma_{kk}|>\sigma_{kk}\lambda\right)$, we have with probability greater than $1-pe^{-O(n_1+n_2)}$, $|\hat{\sigma}_{kk}-\sigma_{kk}|<\epsilon_0^{-1}\lambda$ for all $1\leq i\leq p$.

Combining with the fact that $\mathbb{E}(q_k)$ is bounded, the theorem follows. □

S1.4 Proof of Theorem 4

Define $\mathrm{ECC}^2_{\boldsymbol{\delta}}=\mathbb{E}\left\{\sum_{j=1}^{m}\mathbb{I}(\delta_j=2,\theta_j=2)\right\}$. We will prove $\mathrm{mFSR}^2_{\hat{\boldsymbol{\delta}}}=\mathrm{mFSR}^2_{\boldsymbol{\delta}_{OR}}+o(1)$ and $\mathrm{ECC}^2_{\hat{\boldsymbol{\delta}}}/\mathrm{ECC}^2_{\boldsymbol{\delta}_{OR}}=1+o(1)$. Then by the same argument one can show $\mathrm{mFSR}^1_{\hat{\boldsymbol{\delta}}}=\mathrm{mFSR}^1_{\boldsymbol{\delta}_{OR}}+o(1)$ and $\mathrm{ECC}^1_{\hat{\boldsymbol{\delta}}}/\mathrm{ECC}^1_{\boldsymbol{\delta}_{OR}}=1+o(1)$, then the theorem follows.

We begin with a summary of notation used throughout the proof:

- $Q(t)=m^{-1}\left\{\sum_{j=1}^{m}(T^j-\alpha_2)\mathbb{I}(T^j<t)\right\}$.
- $\widehat{Q}(t)=m^{-1}\left\{\sum_{j=1}^{m}(\hat{T}^j-\alpha_2)\mathbb{I}(\hat{T}^j<t)\right\}$.
- $Q_\infty(t)=\mathbb{E}\{(T-\alpha_2)\mathbb{I}(T<t)\}$.
- $t_\infty=\min\{\sup\{t:Q_\infty(t)\leq 0\},0.5\}$ is the "ideal" threshold.

Note that T^j is a function of $\boldsymbol{W}_j$ only. Since each $\boldsymbol{W}_j$ are iid, it follows that T^j are also iid. Without loss of generality, assume the first s_1 signals are strong, the next s_2 signals are moderate and the rest are weak. We break the proof into 2 cases:

Case 1: $\sum_{k=1}^{s_1} d_k^2 < \infty$ as $n, p \to \infty$.

Case 2: $\sum_{k=1}^{s_1} d_k^2 \to \infty$ as $n, p \to \infty$.

Proof of case 1

We first show that $\mathrm{mFSR}^2_{\hat{\boldsymbol{\delta}}} = \mathrm{mFSR}^2_{\boldsymbol{\delta}_{OR}} + o(1)$. We define a continuous version of $\widehat{Q}(t)$ using the following procedure: If t_1 and t_2 are two adjacent points of discontinuity, on the interval $[t_1, t_2]$,

$$\widehat{Q}_C(t) = \frac{t - t_2}{t_1 - t_2}\widehat{Q}(t_1) + \frac{t - t_1}{t_2 - t_1}\widehat{Q}(t_2).$$

As t_∞ is proved to be larger than α_1 in section S1.1, it is easy to verify that $\widehat{Q}_C(t)$ is continuous and monotone on the interval $[\alpha_1, 1)$. Hence, its inverse $\widehat{Q}_C^{-1}$ is well–defined, continuous, and monotone.

Next, we shall show the following two results in turn: (i) $\widehat{Q}(t) \xrightarrow{p} Q_\infty(t)$ and (ii) $\widehat{Q}_C^{-1}(0) \xrightarrow{p} t_\infty$. We shall need the following two lemmas, which are proved later:

Lemma 1. $\mathbb{E}(\hat{T} - T)^2 \to 0$.

Lemma 2. *Let* $V_j = (T^j - \alpha_2)\mathbb{I}(T^j < t)$ *and* $\hat{V}_j = (\hat{T}^j - \alpha_2)\mathbb{I}(\hat{T}^j < t)$. *Then* $\mathbb{E}\left(\hat{V}_j - V_j\right)^2 = o(1)$.

To show (i), note that $Q(t) \xrightarrow{p} Q_\infty(t)$ by the WLLN, so that we only

need to establish that $\widehat{Q}(t) \xrightarrow{p} Q(t)$. Let $S_m = \sum_{j=1}^{m} \left(\hat{V}_j - V_j\right)$. By Lemma 2 and the Cauchy-Schwartz inequality, $\mathbb{E}\left\{\left(\hat{V}_i - V_i\right)\left(\hat{V}_j - V_j\right)\right\} = o(1)$. It follows that

$$\begin{aligned} Var\left(m^{-1}S_m\right) =& m^{-2}Var(S_m) \leq m^{-2}\sum_{j=1}^{m}\mathbb{E}\left\{\left(\hat{V}_j - V_j\right)^2\right\} \\ &+ O\left(\frac{1}{m^2}\sum_{i,j:i\neq j}\mathbb{E}\left\{\left(\hat{V}_i - V_i\right)\left(\hat{V}_j - V_j\right)\right\}\right) \\ =& o(1). \end{aligned}$$

By Lemma 2, $\mathbb{E}(m^{-1}S_m) \to 0$, applying Chebyshev's inequality, we obtain $m^{-1}S_m = \widehat{Q}(t) - Q(t) \xrightarrow{p} 0$. Hence (i) is proved.

Next, we show (ii). Since $\widehat{Q}_C(t)$ is continuous, for any $\varepsilon > 0$, we can find $\eta > 0$ such that $\left|\widehat{Q}_C^{-1}(Q_\infty(t_\infty)) - \widehat{Q}_C^{-1}\left\{\widehat{Q}_C\left(t_\infty\right)\right\}\right| < \varepsilon$ if $\left|\widehat{Q}_C\left(t_\infty\right) - Q_\infty(t_\infty)\right| < \eta$. It follows that

$$P\left\{\left|\widehat{Q}_C\left(t_\infty\right) - Q_\infty(t_\infty)\right| > \eta\right\} \geq P\left\{\left|\widehat{Q}_C^{-1}(Q_\infty(t_\infty)) - \widehat{Q}_C^{-1}\left\{\widehat{Q}_C\left(t_\infty\right)\right\}\right| > \varepsilon\right\}.$$

Lemma 1 and the WLLN imply that $\widehat{Q}_C(t) \xrightarrow{p} Q_\infty(t)$.Then,

$$P\left(\left|\widehat{Q}_C\left(t_\infty\right) - Q_\infty(t_\infty)\right| > \eta\right) \to 0.$$

Hence, we have

$$\widehat{Q}_C^{-1}(Q_\infty(t_\infty)) \xrightarrow{p} \widehat{Q}_C^{-1}\left\{\widehat{Q}_C\left(t_\infty\right)\right\} = t_\infty, \tag{S1.1}$$

completing the proof of (ii). Notice that $Q_\infty(t)$ is continuous by construction, by (i) we also have $\widehat{Q}(t) \xrightarrow{p} \widehat{Q}_C(t)$.

We can similarly define the continuous version of $Q(t)$ as $Q_C(t)$ and the corresponding threshold as $Q_C^{-1}(0)$. Write $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}^1 + \hat{\boldsymbol{\delta}}^2$, where $\hat{\boldsymbol{\delta}}^1$ is of the form $\mathbb{I}\left\{1 - \hat{T}^j \le \beta_1\right\}$ and $\hat{\boldsymbol{\delta}}^2$ is of the form $2\mathbb{I}\left\{\hat{T}^j \le \beta_2\right\}$ for some $\beta_1, \beta_2 > 0$. Similarly, write $\boldsymbol{\delta} = \boldsymbol{\delta}^1 + \boldsymbol{\delta}^2$, where $\boldsymbol{\delta}^1$ is of the form $\mathbb{I}\{1 - T^j \le t_{OR}^1\}$ and $\boldsymbol{\delta}_{OR}^2$ is of the form $2\mathbb{I}\{T^j \le t_{OR}^2\}$ for some $\beta_1, \beta_2 > 0$. Then by construction, we have

$$\hat{\boldsymbol{\delta}}^2 = \left[2\mathbb{I}\left\{\hat{T}^j \le \widehat{Q}_C^{-1}(Q_\infty(t_\infty))\right\} : 1 \le j \le m\right]$$

$$\text{and} \quad \boldsymbol{\delta}_{OR}^2 = \left[2\mathbb{I}\left\{T^j \le Q_C^{-1}(Q_\infty(t_\infty))\right\} : 1 \le i \le m\right].$$

Also, following the previous arguments, we can show that

$$Q_C^{-1}(Q_\infty(t_\infty)) \xrightarrow{p} t_\infty. \tag{S1.2}$$

According to (S1.1) and (S1.2), we have

$$\widehat{Q}_C^{-1}(Q_\infty(t_\infty)) = Q_C^{-1}(Q_\infty(t_\infty)) + o_p(1). \tag{S1.3}$$

Note that the mFSR^2 level of $\boldsymbol{\delta}_{OR}$ and $\hat{\boldsymbol{\delta}}$ are

$$\text{mFSR}^2_{\boldsymbol{\delta}_{OR}} = \frac{P_{\theta_j=1}\left\{T^j \le Q_C^{-1}(Q_\infty(t_\infty))\right\}}{P\left\{T^j \le Q_C^{-1}(Q_\infty(t_\infty))\right\}} \quad \text{and} \quad \text{mFSR}^2_{\hat{\boldsymbol{\delta}}} = \frac{P_{\theta_j=1}\left\{\hat{T}^j \le Q_C^{-1}(Q_\infty(t_\infty))\right\}}{P\left\{\hat{T}^j \le Q_C^{-1}(Q_\infty(t_\infty))\right\}}.$$

From Lemma 1, $\hat{T}^j \xrightarrow{p} T^j$, (S1.2), and by the continuous mapping theorem, $\text{mFSR}^2_{\hat{\boldsymbol{\delta}}} = \text{mFSR}^2_{\boldsymbol{\delta}_{OR}} + o(1) \le \alpha_2 + o(1)$. By the asymptotic equivalence between mFSR and FSR (Supplement S3), the desired result follows.

Next we show that $\text{ECC}^2_{\hat{\boldsymbol{\delta}}}\ /\text{ECC}^2_{\boldsymbol{\delta}_{OR}} = 1 + o(1)$. By definition,

$$\text{ECC}^2_{\hat{\boldsymbol{\delta}}} = \mathbb{E}\left\{\sum_{j=1}^{m} \mathbb{I}(\hat{T}^j \le \widehat{Q}_C^{-1}(Q_\infty(t_\infty)))(1 - T^i)\right\}$$

$$\text{and} \quad \text{ECC}^2_{\boldsymbol{\delta}_{OR}} = \mathbb{E}\left\{\sum_{j=1}^{m} \mathbb{I}(T^j \le Q_C^{-1}(Q_\infty(t_\infty)))(1 - T^j)\right\}.$$

Using the fact that $\hat{T}^j \xrightarrow{p} T^j$ and $\widehat{Q}_C^{-1}(Q_\infty(t_\infty)) \xrightarrow{p} Q_C^{-1}(Q_\infty(t_\infty))$, the result follows.

Proof of case 2

From the proof of Lemma 1, we have:

$$\mathbb{E}\left(S_j^\pi - \hat{S}_j\right)^2 = O(s_1\frac{n_1+n_2}{n_1n_2}) + o(\sum_{k=1}^{s_1} d_k^2) + O(\sum_{k=1}^{s_1} d_k\frac{n_1+n_2}{n_1n_2}) + o(1). \quad \text{(S1.4)}$$

Since $\sum_{k=1}^{s_1} d_k^2 \geq cs_1 \frac{n_1+n_2}{n_1 n_2} \log p$ for some $c > 0$ and $\mathbb{E}(S_j^\pi) \propto \sum_{k=1}^{s_1} d_k^2$, by (S1.4)

$$\mathbb{E}\left(S_j^\pi - \hat{S}_j\right)^2 \Big/ \sum_{k=1}^{s_1} d_k^2 = O\left(\frac{1}{\log p}\right).$$

It follows that $\mathbb{E}\sum_{k=1}^{s_1}\{q_k(\bar{X}_k - \bar{Y}_k)\}^2 \propto \mathbb{E}(|S_j^\pi|) \propto \sum_{k=1}^{s_1} d_k^2$. If $\sum_{k=1}^{s_k} d_i^2 \to \infty$, then $\hat{T}^j \to 0$ or $\hat{T}^j \to \infty$. In either case, we have perfect separation between the two classes asymptotically. The theorem follows. □

S1.5 Proof of Theorem 3

Again, without loss of generality, assume the first s_1 signals are strong, the next s_2 signals are moderate and the rest are weak. Consider the same model as described in section 2, but replace $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ by $\tilde{\boldsymbol{\mu}}_1 = (\mu_{11}, ..., \mu_{1s_1}, 0, ..., 0)^t$ and $\tilde{\boldsymbol{\mu}}_2 = (\mu_{21}, , ..., \mu_{2s_1}, 0, ..., 0)^t$ respectively. Let $\tilde{\boldsymbol{W}}_j = \boldsymbol{W}_j - \boldsymbol{\mu}_1 + \tilde{\boldsymbol{\mu}}_1$ when $\theta_j = 1$ and $\tilde{\boldsymbol{W}}_j = \boldsymbol{W}_j - \boldsymbol{\mu}_2 + \tilde{\boldsymbol{\mu}}_2$ when $\theta_j = 2$. It is then clear that

$$\tilde{\boldsymbol{W}}_j \sim \mathbb{I}(\theta_j = 1)\mathcal{N}(\tilde{\boldsymbol{\mu}}_1, \Sigma) + \mathbb{I}(\theta_j = 2)\mathcal{N}(\tilde{\boldsymbol{\mu}}_2, \Sigma).$$

Denote $\tilde{\boldsymbol{d}} = \tilde{\boldsymbol{\mu}}_1 - \tilde{\boldsymbol{\mu}}_2$, define

$$Z_j := Z(\tilde{\boldsymbol{W}}_j) := \left(\tilde{\boldsymbol{W}}_j - \frac{\tilde{\boldsymbol{\mu}}_1 + \tilde{\boldsymbol{\mu}}_2}{2}\right)^\top \Sigma^{-1}\tilde{\boldsymbol{d}},$$

$Z_{OR}^j := Z_{OR}(\tilde{\boldsymbol{W}}_j) = \exp(Z_j)/(\exp(Z_j)+1)$. Consider the decision rule:

$$\delta_Z^j = 2\mathbb{I}\{Z_{OR}^j \leq \min(Z_{OR}^{(k_4)}, 0.5)\} + \mathbb{I}\{1 - Z_{OR}^j \leq \min(1 - Z_{OR}^{(m-k_3)}, 0.5)\},$$

where

$$k_3 = \inf\left\{j : \frac{1}{j+1}\sum_{i=0}^{j}(1 - Z_{OR}^{(m-i)}) \leq \alpha_1\right\} \quad \text{and} \quad k_4 = \sup\left\{j : \frac{1}{j}\sum_{i=1}^{j} Z_{OR}^{(i)} \leq \alpha_2\right\}.$$

By the proof of theorem 1 and theorem 3, $\boldsymbol{\delta}_Z = (\delta_Z^1, \ldots, \delta_Z^m)$ controls mFSR[1] and mFSR[2] at level α_1 and α_2 respectively. Now notice that

$$\begin{aligned}
&\mathbb{E}\left(Z_j - \hat{S}_j\right)^2 \\
=&\mathbb{E}\left\{\left(\tilde{\boldsymbol{W}}_j - \frac{\tilde{\boldsymbol{\mu}}_1 + \tilde{\boldsymbol{\mu}}_2}{2}\right)^\top \Sigma^{-1}\tilde{\boldsymbol{d}} - \left(\boldsymbol{W}_j - \frac{\bar{\boldsymbol{X}} + \bar{\boldsymbol{Y}}}{2}\right)^\top \hat{\Sigma}^{-1}\hat{\boldsymbol{d}}\right\}^2 \\
=&O\left[\mathbb{E}\left\{\left(\boldsymbol{W}_i - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}(\tilde{\boldsymbol{d}} - \hat{\boldsymbol{d}})\right\}^2 + \mathbb{E}\left\{\left(\tilde{\boldsymbol{W}}_j - \boldsymbol{W}_j + \frac{\bar{\boldsymbol{X}} + \bar{\boldsymbol{Y}}}{2} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}\hat{\boldsymbol{d}}\right\}^2\right. \\
&\left. + \mathbb{E}\left\{\left(\boldsymbol{W}_i - \frac{\bar{\boldsymbol{X}} + \bar{\boldsymbol{Y}}}{2}\right)^\top (\Sigma - \hat{\Sigma}^{-1})\hat{\boldsymbol{d}}\right\}^2\right] \\
=&I + II + III
\end{aligned}$$

For II, note that

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\tilde{\boldsymbol{W}}_j - \boldsymbol{W}_j + \frac{\tilde{\boldsymbol{\mu}}_1 + \tilde{\boldsymbol{\mu}}_2}{2} - \frac{\bar{\boldsymbol{X}} + \bar{\boldsymbol{Y}}}{2}\right)^\top \Sigma^{-1}\hat{\boldsymbol{d}}\right\}^2 \\
=&O\left[\mathbb{E}\left\{\hat{\boldsymbol{d}}^\top (\Sigma^{-1})^t \frac{1}{4}\frac{n_1+n_2}{n_1 n_2}\Sigma\Sigma^{-1}\hat{\boldsymbol{d}}\right\} + \left\{\lambda_{max}(\Sigma^{-1})p^{-\gamma}\log p\right\}^2\right] \\
=&O\left[\mathbb{E}\left\{\hat{\boldsymbol{d}}^\top (\Sigma^{-1})^\top \frac{1}{4}\frac{n_1+n_2}{n_1 n_2}\hat{\boldsymbol{d}}\right\} + \left\{\lambda_{max}(\Sigma^{-1})p^{-\gamma}\log p\right\}^2\right] \\
=&O\left[\mathbb{E}\left(\frac{n_1+n_2}{n_1 n_2}\lambda_{max}(\Sigma^{-1})||\boldsymbol{d}-\hat{\boldsymbol{d}}||^2\right) + \mathbb{E}\left(\frac{n_1+n_2}{n_1 n_2}\lambda_{max}(\Sigma^{-1})||\boldsymbol{d}||^2\right) + \left\{\lambda_{max}(\Sigma^{-1})p^{-\gamma}\log p\right\}^2\right] \\
=&O\left[\mathbb{E}\frac{n_1+n_2}{n_1 n_2}\lambda_{max}(\Sigma^{-1})(||\boldsymbol{d}-\hat{\boldsymbol{d}}||^2 + ||\boldsymbol{d}||^2) + \left\{\lambda_{max}(\Sigma^{-1})p^{-\gamma}\log p\right\}^2\right].
\end{aligned}
$$

By (A1) $\left\{\lambda_{max}(\Sigma^{-1})p^{-\gamma}\log p\right\}^2 \to 0$, use the same computation in the proof of lemma 1, we have

$$
\mathbb{E}\left(Z_j - \hat{S}_j\right)^2 = O\left(\mathbb{E}\sum_{k=1}^{p}\{q_k(\bar{X}_k - \bar{Y}_k) - \tilde{d}_k\}^2\right) + o(1),
$$

$$
\mathbb{E}\sum_{k=1}^{s_1}\{q_k(\bar{X}_k - \bar{Y}_k) - \tilde{d}_k\}^2 = O(s_1\frac{n_1+n_2}{n_1 n_2}) + o(\sum_{k=1}^{s_1} d_k^2) + O(\sum_{k=1}^{s_1} d_k\frac{n_1+n_2}{n_1 n_2}) + o(1).
$$

We know that

$$
\mathbb{E}\sum_{i=s_1+1}^{s_1+s_2}\left\{q_k(\bar{X}_k - \bar{Y}_k)\right\}^2 = O(s_2)O(p^{-\gamma})O\left(\frac{n_1+n_2}{n_1 n_2}\log p\right) = O(p^{-\gamma}\log p) \to 0,
$$

$$
\mathbb{E}\sum_{k=s_2+1}^{p}\{q_k(\bar{X}_k - \bar{Y}_k)\}^2 = O(p)O(p^{-1-\gamma})o\left(\frac{n_1+n_2}{n_1 n_2}\log p\right) \to 0.
$$

Thus,

$$\mathbb{E}\left(Z_j - \hat{S}_j\right)^2 = O(s_1\frac{n_1+n_2}{n_1n_2}) + o(\sum_{k=1}^{s_1} d_k^2) + O(\sum_{k=1}^{s_1} d_k\frac{n_1+n_2}{n_1n_2}) + o(1). \quad \text{(S1.5)}$$

Since $\sum_{k=1}^{s_1} d_k^2 \geq cs_1 \log p\frac{n_1+n_2}{n_1n_2}$ for some $c > 0$ and $\mathbb{E}(Z_j) \propto \sum_{k=1}^{s_1} d_k^2$, by (S1.5)

$$\mathbb{E}\left(Z_j - \hat{S}_j\right)^2 \Big/ \sum_{k=1}^{s_1} d_k^2 = O\left(\frac{1}{\log p}\right)$$

It follows that $\mathbb{E}\sum_{k=1}^{s_1}\{q_k(\bar{X}_k - \bar{Y}_k)\}^2 \propto \mathbb{E}(|Z_j|) \propto \sum_{k=1}^{s_1} d_k^2$. If $\sum_{k=1}^{s_1} d_k^2 \to \infty$, then $\hat{T}^j \to 0$ or $\hat{T}^j \to 1$. In either case, we have perfect separation between the two classes asymptotically. The theorem follows trivially. If $\sum_{k=1}^{s_1} d_k^2 < \infty$, then

$$\left(\mathbb{E}\left|\hat{T}^j - Z_{OR}^j\right|\right)^2 = e^{O(\sum_{k=1}^{s_1} d_k^2)}\left\{O\left(s_1\frac{n_1+n_2}{n_1n_2}\right) + O(\sum_{k=1}^{s_1} d_k^2 p^{-2\gamma}) + O(p^{-\gamma}\log p)\right\}.$$

Since $\sum_{k=1}^{s_1} d_k^2 \geq cs_1 \log p\frac{n_1+n_2}{n_1n_2}$ and $\sum_{k=1}^{s_1} d_k^2 < \infty$, it follows that $e^{O(\sum_{k=1}^{s_1} d_k^2)} < \infty$

and $O\left(s_1\frac{n_1+n_2}{n_1n_2}\right) + O(\sum_{k=1}^{s_1} d_k^2 p^{-2\gamma}) + O(p^{-\gamma}\log p) \to 0$. Thus, $\mathbb{E}\left(\hat{T}^j - Z_{OR}^j\right)^2 \to 0$. The theorem follows by using the same argument presented in the proof of theorem 4. □

S1.6 Proof of Corollary 1

Note that $\min\{T^j, 1-T^j\} \leq 0.5$. Hence, if we choose $\alpha_1 = \alpha_2 = 0.5$, $\boldsymbol{\delta}_{OR}$ makes no indecision. Since $\boldsymbol{\delta}_{OR}$ has the highest ECC among all rules that controls mFSR^1 and mFSR^2 at level 0.5, and $\boldsymbol{\delta}^F$ also controls mFSR^1 and mFSR^2 at level 0.5, it follows that $\mathrm{ECC}_{\boldsymbol{\delta}_{OR}} \geq \mathrm{ECC}_{\boldsymbol{\delta}^F}$. On the other hand, $\boldsymbol{\delta}^F$ has the lowest risk among all rules that do not make indecision and risk equals to $1 - \mathrm{ECC}/m$. It follows that $\mathrm{ECC}_{\boldsymbol{\delta}_{OR}} \leq \mathrm{ECC}_{\boldsymbol{\delta}^F}$. Hence, $\mathrm{ECC}_{\boldsymbol{\delta}_{OR}} = \mathrm{ECC}_{\boldsymbol{\delta}^F}$. By Theorem 4, we have $\mathrm{ECC}_{\boldsymbol{\delta}_{OR}}/\mathrm{ECC}_{\hat{\boldsymbol{\delta}}} \to 1$, it follows that $R(\hat{\boldsymbol{\delta}}) \to R(\boldsymbol{\delta}^F)$. □

S2 Proof of Technical Lemmas

Supplement S2 contains proofs of techincal lemmas. We will assume without loss of generality, the first s_1 signals are strong, the next s_2 signals are moderate and the rest are weak.

S2.1 Proof of Lemma 1

Let $\hat{\boldsymbol{d}} = (q_1(\bar{X}_1 - \bar{Y}_1), ..., q_p(\bar{X}_p - \bar{Y}_p))$. We have

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{d}} - \boldsymbol{d}\|^2 &= \sum_{k=1}^{p} \left\{ \mathbb{E}(q_k^2(\bar{X}_k - \bar{Y}_k)^2) - 2d_k \mathbb{E}\{q_k(\bar{X}_k - \bar{Y}_k)\} + d_k^2 \right\} \\
&= \sum_{k=1}^{p} \left\{ \mathrm{Cov}\{q_k^2, (\bar{X}_k - \bar{Y}_k)^2\} + \mathbb{E}(q_k^2)\mathbb{E}(\bar{X}_k - \bar{Y}_k)^2 \right\} \\
&\quad - \sum_{k=1}^{p} \left\{ 2d_k \left[\mathrm{Cov}\left\{ q_k, (\bar{X}_k - \bar{Y}_k) \right\} + d_k \mathbb{E} q_k \right] + d_k^2 \right\}.
\end{aligned}$$

Since q_k is bounded between 0 and 1, $\mathrm{Var}(q_k^2)$ and $\mathrm{Var}(q_i)$ are both bounded. For $1 \le k \le s_1$, note that $(\bar{X}_k - \bar{Y}_k) \sim N(d_k, \sigma_{kk}\frac{n_1+n_2}{n_1 n_2})$, and $d_k^2 < \infty$ we have $\mathrm{Var}\{(\bar{X}_k - \bar{Y}_k)^2\} = O(\frac{n_1+n_2}{n_1 n_2})$. Use Cauchy-Schwarz inequality, for $1 \le k \le s_1$, each summand has absolute value bounded by

$$\begin{aligned}
&O(\frac{n_1+n_2}{n_1 n_2}) + \mathbb{E}\left\{ q_k^2 \left(\frac{n_1+n_2}{n_1 n_2}\sigma_{kk} + d_k^2 \right) \right\} - 2d_k^2 \mathbb{E}(q_k) + d_k^2 - 2d_k \mathrm{Cov}(q_k, \bar{X}_k - \bar{Y}_k) \\
=&O(\frac{n_1+n_2}{n_1 n_2}) + d_k^2 \mathbb{E}(q_k^2) + d_k^2 - 2d_k^2 \mathbb{E}(q_k) + O(d_k \frac{n_1+n_2}{n_1 n_2}) + O(\frac{n_1+n_2}{n_1 n_2}) \\
=&O(\frac{n_1+n_2}{n_1 n_2}) + O(d_k \frac{n_1+n_2}{n_1 n_2}) + O\left\{ d_k^2 (1 - \mathbb{E}(q_k))^2 \right\} + O\left\{ d_k^2 (E(q_k^2) - (Eq_k)^2) \right\} \\
=&O(\frac{n_1+n_2}{n_1 n_2}) + O(d_k \frac{n_1+n_2}{n_1 n_2}) + o(d_k^2).
\end{aligned}$$

For $s_1 + 1 \le i \le p$,

$$\mathbb{E}\{q_k(\bar{X}_k - \bar{Y}_k) - d_k\}^2 \le d_k^2.$$

Hence, by (A4)

$$\mathbb{E}\|\hat{\boldsymbol{d}}-\boldsymbol{d}\|^2 = O(s_1\frac{n_1+n_2}{n_1n_2}) + o(\sum_{k=1}^{s_1} d_k^2) + O(\sum_{k=1}^{s_1} d_k\frac{n_1+n_2}{n_1n_2}) + o(1). \quad \text{(S2.6)}$$

Since $s_1\frac{n_1+n_2}{n_1n_2} = O(\sum_{k=1}^{s_1} d_k^2)$ and by assumption of case 1, $\sum_{k=1}^{s_1} d_k^2 = O(1)$ and $\sum_{k=1}^{s_1} d_k\frac{n_1+n_2}{n_1n_2} = o(1)$ we have $\mathbb{E}\|\hat{\boldsymbol{d}}-\boldsymbol{d}\|^2 = o(1)$. Now, since $\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}$ and $\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}$ are independent, and using Basu's Theorem we know that $\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}$ and $\hat{\sigma}_{kk}$ are independent, thus $\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}$ and $\hat{d}$ are independent and we have

$$\begin{aligned}
&\mathbb{E}(S_j^\pi - \hat{S}_j)^2 \\
=&\mathbb{E}\left\{\left(\boldsymbol{W}_j - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}\boldsymbol{d} - \left(\boldsymbol{W}_j - \frac{\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}}{2}\right)^\top \hat{\Sigma}^{-1}\hat{\boldsymbol{d}}\right\}^2 \\
=&\mathbb{E}\left\{\left(\boldsymbol{W}_j - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}(\boldsymbol{d}-\hat{\boldsymbol{d}}) + \left(\frac{\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}}{2} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}\hat{\boldsymbol{d}}\right. \\
&\left. + \left(\boldsymbol{W}_j - \frac{\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}}{2}\right)^\top (\Sigma^{-1}-\hat{\Sigma}^{-1})\hat{\boldsymbol{d}}\right\}^2 \\
=&O\left[\mathbb{E}\left\{\left(\boldsymbol{W}_j - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}(\boldsymbol{d}-\hat{\boldsymbol{d}})\right\}^2 + \mathbb{E}\left\{\left(\frac{\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}}{2} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}\hat{\boldsymbol{d}}\right\}^2\right. \\
&\left. + \mathbb{E}\left\{\left(\boldsymbol{W}_j - \frac{\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}}{2}\right)^\top (\Sigma-\hat{\Sigma}^{-1})\hat{\boldsymbol{d}}\right\}^2\right] \\
=&I + II + III.
\end{aligned}$$

For term I, as in case 1 we have $||\boldsymbol{d}||^2$ bounded:

$$\mathbb{E}\left\{\left(\boldsymbol{W}_j - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}(\boldsymbol{d}-\hat{\boldsymbol{d}})\right\}^2 = O\left\{\mathbb{E}\left(\lambda_{max}(\Sigma^{-1})^2||\boldsymbol{d}-\hat{\boldsymbol{d}}||^2\right)\right\},$$

where $\lambda_{max}(\Sigma^{-1})$ is the largest eigenvalue of Σ^{-1}, which in our case is bounded by some constant.

For term II, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left(\frac{\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}}{2} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}\right)^\top \Sigma^{-1}\hat{\boldsymbol{d}}\right\}^2 \\
=&O\left\{\mathbb{E}\left(\frac{n_1+n_2}{n_1n_2}\lambda_{max}(\Sigma^{-1})||\boldsymbol{d}-\hat{\boldsymbol{d}}||^2\right) + \mathbb{E}\left(\frac{n_1+n_2}{n_1n_2}\lambda_{max}(\Sigma^{-1})||\boldsymbol{d}||^2\right)\right\} \\
=&O\left\{\frac{n_1+n_2}{n_1n_2}\mathbb{E}\left(\lambda_{max}(\Sigma^{-1})(||\boldsymbol{d}-\hat{\boldsymbol{d}}||^2 + ||\boldsymbol{d}||^2)\right)\right\}.
\end{aligned}$$

For term III, we have

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\boldsymbol{W}_j - \frac{\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}}{2}\right)^\top (\Sigma^{-1}-\hat{\Sigma}^{-1})\hat{\boldsymbol{d}}\right\}^2 \\
=&\mathbb{E}\left[\hat{\boldsymbol{d}}^\top(\Sigma-\hat{\Sigma}^{-1})^\top\left\{\left(1+\frac{1}{4}\frac{n_1+n_2}{n_1n_2}\right)\Sigma^{-1}+\frac{1}{4}\boldsymbol{d}\boldsymbol{d}^\top\right\}(\Sigma^{-1}-\hat{\Sigma}^{-1})\hat{\boldsymbol{d}}\right] \\
\leq&\mathbb{E}\left[\left\{\left(1+\frac{1}{4}\frac{n_1+n_2}{n_1n_2}\right)\frac{1}{\epsilon_0}+\frac{1}{4}\max\{d_id_j, 1\leq i,j\leq p\}\right\}\hat{\boldsymbol{d}}^\top(\Sigma^{-1}-\hat{\Sigma}^{-1})^\top(\Sigma^{-1}-\hat{\Sigma}^{-1})\hat{\boldsymbol{d}}\right] \\
=&O\left[\mathbb{E}\left\{||(\Sigma^{-1}-\hat{\Sigma}^{-1})||_2^2||\hat{\boldsymbol{d}}||^2\right\}\right] \\
=&O\left[\mathbb{E}\left\{||(\Sigma^{-1}-\hat{\Sigma}^{-1})||_2^2||\boldsymbol{d}||^2\right\}\right] \\
=&o(1).
\end{aligned}
$$

The last equality follows from assumption (A2) and the fact that $||\boldsymbol{d}||^2<\infty$ in case 1. Hence,

$$
\mathbb{E}\left(S_j^\pi-\hat{S}_j\right)^2 = O\left\{\mathbb{E}\left(\|\hat{\boldsymbol{d}}-\boldsymbol{d}\|^2\right)\right\}+O(\frac{n_1+n_2}{n_1n_2})+o(1)=o(1).
$$

Next we ask under what conditions do we have $\mathbb{E}\left|\hat{T}^j-T^j\right|^2\to 0$. Let $\delta>0$ be some constant, applying Chebyshev's inequality, we have $P\left(|\hat{S}_j-S_j^\pi|>\delta\right)=O\left\{\mathbb{E}\left(|S_j^\pi-\hat{S}_j|\right)^2\right\}\to 0$. When $|\hat{S}_j-S_j^\pi|\leq\delta$, apply Cauchy-Schwartz in-

equality, we have:

$$\begin{aligned}
&\mathbb{E}_{\left|S_j^\pi - \hat{S}_j\right| < \delta} \left|\exp(\hat{S}_j) - \exp(S_j^\pi)\right|^2 \\
\leq &\mathbb{E} e^{2S_j^\pi + 2\delta} \mathbb{E}\left(S_j^\pi - \hat{S}_j\right)^2 \\
= &e^{O(\sum_{k=1}^{s_1} d_k^2)} \cdot o(1).
\end{aligned}$$

Under the assumption of case 1 the above goes to 0. Therefore, as $\hat{T}^j, T^j$ are bounded above by 1, we have:

$$\begin{aligned}
\mathbb{E}(\hat{T}^j - T^j)^2 &\leq \mathbb{E}\left|\hat{T}^j - T^j\right| \\
&\leq 2P\left(\left|\hat{S}_j - S_j^\pi\right| > \delta\right) + E_{\left|\hat{S}_j - S_j^\pi\right| < \delta}\left|\exp(\hat{S}_j) - \exp(S_j^\pi)\right|.
\end{aligned}$$

The lemma follows. □

S2.2 Proof of Lemma 2

Using the definitions of $\hat{V}_j$ and V_j, we can show that

$$\begin{aligned}
\frac{1}{2}\left(\hat{V}_j - V_j\right)^2 \leq &\left(\hat{T}^j - T^j\right)^2 \mathbb{I}\left(\hat{T}^j \leq t, T^j \leq t\right) + \left(\hat{T}^j - \alpha_2\right)^2 \mathbb{I}\left(\hat{T}^j \leq t, T^j > t\right) \\
&+ \left(T^j - \alpha_2\right)^2 \mathbb{I}\left(\hat{T}^j > t, T^j \leq t\right).
\end{aligned}$$

Let us refer to the three summands on the right hand as (1), (2) and (3) respectively. By Lemma 1, $(1) = o(1)$. Then let $\varepsilon > 0$, and consider that

$$\begin{aligned}
&P\left(\hat{T}^j \le t, T^j > t\right) \\
\le &P\left(\hat{T}^j \le t, T^j \in (t, t+\varepsilon)\right) + P\left(\hat{T}^j \le t, T^j \ge t+\varepsilon\right) \\
\le &P\left\{T^j \in (t, t+\varepsilon)\right\} + P\left(\left|T^j - \hat{T}^j\right| > \varepsilon\right).
\end{aligned}$$

The first term on the right hand is vanishingly small as $\varepsilon \to 0$ because T^j is a continuous random variable. The second term converges to 0 by Lemma 1. Noting that $0 \le T^j \le 1$, we conclude $(2) = o(1)$. In a similar fashion, we can show that $(3) = o(1)$, thus proving the lemma. □

S3 Asymptotic Equivalence of FSR and mFSR

We show $\mathrm{FSR}^c_{\boldsymbol{\delta}}$ and $\mathrm{mFSR}^c_{\boldsymbol{\delta}}$ are asymptotically equivalent. Let $\mathcal{X}^c_{\boldsymbol{\delta}} = \frac{1}{m}\sum_{j=1}^m \mathbb{I}(\delta_j = c, \theta_j \ne c)$ and $\mathcal{Y}^c_{\boldsymbol{\delta}} = \frac{1}{m}\sum_{j=1}^m \mathbb{I}(\delta_j = c)$. The goal is to show $|\mathrm{FSR}^c_{\boldsymbol{\delta}} - \mathrm{mFSR}^c_{\boldsymbol{\delta}}| = o(1)$.

$$|\mathrm{FSR}^c_{\boldsymbol{\delta}} - \mathrm{mFSR}^c_{\boldsymbol{\delta}}| \le \mathbb{E}\left\{\left|\frac{\mathcal{X}^c_{\boldsymbol{\delta}}}{\mathcal{Y}^c_{\boldsymbol{\delta}}} - \frac{\mathcal{X}^c_{\boldsymbol{\delta}}}{\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}}\right| \mathbb{I}(\mathcal{Y}^c_{\boldsymbol{\delta}} > 0)\right\} = \mathbb{E}\left\{\frac{\mathcal{X}^c_{\boldsymbol{\delta}}}{\mathcal{Y}^c_{\boldsymbol{\delta}}}\mathbb{I}(\mathcal{Y}^c_{\boldsymbol{\delta}} > 0)\frac{|\mathcal{Y}^c_{\boldsymbol{\delta}} - \mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}|}{\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}}\right\}$$

Since $\mathcal{X}^c_{\boldsymbol{\delta}} \le \mathcal{Y}^c_{\boldsymbol{\delta}}$, we have

$$\mathbb{E}\left\{\frac{\mathcal{X}^c_{\boldsymbol{\delta}}}{\mathcal{Y}^c_{\boldsymbol{\delta}}}\mathbb{I}(\mathcal{Y}^c_{\boldsymbol{\delta}}>0)\frac{|\mathcal{Y}^c_{\boldsymbol{\delta}}-\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}|}{\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}}\right\} \le \mathbb{E}\left\{\frac{|\mathcal{Y}^c_{\boldsymbol{\delta}}-\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}|}{\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}}\right\} \le \frac{(\mathbb{E}|\mathcal{Y}^c_{\boldsymbol{\delta}}-\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}|^2)^{1/2}}{\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}} = \frac{(\mathrm{Var}\mathcal{Y}^c_{\boldsymbol{\delta}})^{1/2}}{\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}}}$$

By assumption (A3), we have $m\mathcal{Y}^c_{\boldsymbol{\delta}_{OR}} \sim \mathrm{Binom}(m,\eta)$ with $\eta > 0$. Therefore, $\mathbb{E}\mathcal{Y}^c_{\boldsymbol{\delta}_{OR}} = \eta$ and $\mathrm{Var}(\mathcal{Y}^c_{\boldsymbol{\delta}_{OR}})^{1/2} = \sqrt{\eta(1-\eta)/m}$, $|\mathrm{FSR}^c_{\boldsymbol{\delta}_{OR}} - \mathrm{mFSR}^c_{\boldsymbol{\delta}_{OR}}| = O(m^{-1/2})$. In the setting of Theorem 4, the data driven procedure is mimicking $\boldsymbol{\delta}_{OR}$. It can be seen from the proof of Theorem 4 that mFSR and FSR of the data driven procedure are also asymptotically equivalent. Similarly, for the setting of Theorem 3, consider the rule $\boldsymbol{\delta}_Z$ defined in the proof of Theorem 3, by assumption (A3) we also have $|\mathrm{FSR}^c_{\boldsymbol{\delta}_Z} - \mathrm{mFSR}^c_{\boldsymbol{\delta}_Z}| = O(m^{-1/2})$. In the setting of Theorem 3, the data driven procedure is mimicking $\boldsymbol{\delta}_Z$. It can be seen from the proof of Theorem 3 that mFSR and FSR of the data driven procedure are also asymptotically equivalent. □

S4 Adaptation to unknown sparsity: illustrations

We present numerical examples to provide insights on why the shrinkage rule (3.1) works well across the sparse and dense regimes.

We start with the sparse case. Suppose $n_1 = n_2 = n$. Let $n = 5$, $p = 625$, $\boldsymbol{d} = (d_1, \ldots, d_p)$, where $d_k = 2.5$ for $k \in \{1, \ldots, 50\}$ and $d_k = 0$ for

$k \in \{51, \ldots, 625\}$. The observations are generated as $\boldsymbol{X}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{2} I_p)$ and $\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{d}, \frac{1}{2} I_p)$, $i = 1, \ldots, n$, where $\mathbf{0}$ is a p-dimensional vector of zeros and I_p a p-dimensional identity matrix. It follows that $\bar{X}_k - \bar{Y}_k \sim \mathcal{N}(d_k, 1/n)$. We contrast the proposed shrinkage rule with hard/soft thresholding rules in Figure 1. Panels (a) and (b) plot d_k and observed $\bar{X}_k - \bar{Y}_k$, respectively, with coordinates corresponding to zero/nonzero d_k being marked in blue/red. Panel (c) plots the shrinkage estimates $(\bar{X}_k - \bar{Y}_k)q_k$, where q_k is defined in (3.2) with $b = 0$. Panels (d)–(f) show the hard/soft thresholding estimates with different thresholds λ, where $\rho_h(t, \lambda, \sigma) \coloneqq t\mathbb{I}(|t| > \lambda\sqrt{\sigma})$ and $\rho_s(t, \lambda, \sigma) \coloneqq \operatorname{sgn}(t) \cdot \max(|t| - \lambda\sqrt{\sigma}, 0)$ are the hard and soft thresholding functions, respectively.

We can see from Panel (c) that almost all blue points are pulled towards and centered around the line of zero by our proposed shrinkage rule. The multiplicative factor q_k is as effective as existing thresholding methods for noise reduction in the sense that the patterns in Panel (c) is qualitatively similar to those in Panels (d) to (f), except that in (d) and (e) too many blue points have survived whereas in (f) too many nonzero signals have been killed. Panel (c) shrinks most noisy entries to zero while being capable of preserving a significant portion of nonzero signals. The bottom row compares the shrinkage functions ρ_h and ρ_s with q_k by setting $\hat{\sigma}_{kk} = 0.5$

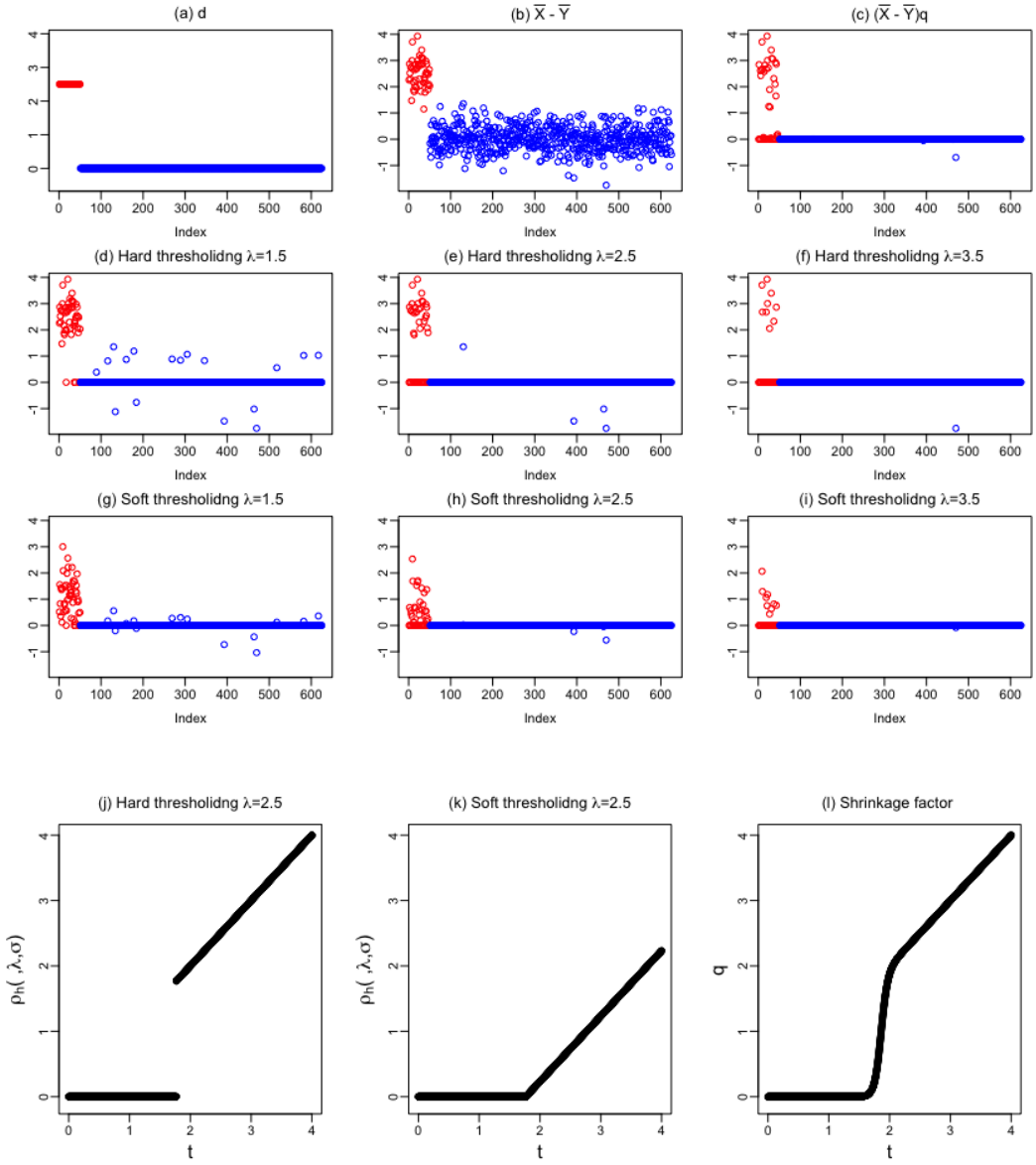

Figure 1: A comparison of our shrinkage rule (3.2) with hard/soft thresholding rules in the sparse case. Panels (a), (b) and (c) respectively plot the true d_k, observed $\bar{X}_k - \bar{Y}_k$ and $(\bar{X}_k - \bar{Y}_k)q_k$. Panels (d)-(f) [((g)-(i)] present results of hard-thresholding (soft-thresholding) rules. The effects of different shrinkage functions are provided at the bottom row.

and $\lambda = 2.5$. We can see that our shrinkage function is continuous, nearly unbiased for large signals and yields similar effects as that of the hard-thresholding function. The proposed shrinkage rule is desirably tuning-free in the sense that one can simply set $b = 0$ in practice; this merit is justified in our theoretical analysis and corroborated by our numerical results. By contrast, the performance of existing thresholding rules depends critically on the value of λ, which is nontrivial to choose.

Next we turn to the dense case. The data are generated in a similar way as before except that we let $d_k = 2.5(k-1)/624$, $k = 1, \ldots, 625$, i.e. d_k increases linearly from 0 to 2.5. We mark the coordinates of $\boldsymbol{d}$ in three colors: red if $d_k > (\sqrt{0.5} + \sqrt{0.5+1})\sqrt{\log p/n} \approx 2.19$, black if $1.5 < d_k < (\sqrt{0.5} + \sqrt{0.5+1})\sqrt{\log p/n}$ and blue otherwise[1]. We plot our shrinkage estimate and thresholding estimates in Figure 2. In this high-dimensional "dense" regime, the boundaries between weak, moderate and strong signals are blurred. Therefore the working assumptions (sparsity and dichotomy of $\boldsymbol{d}$) underpinning the use of thresholding rules can be problematic. In contrast with the (roughly) linear patterns of the surviving points in Panels (d) to (i), the curved pattern in Panel (c) provides differential amount of shrinkage for weak and strong signals, achieving a more desirable balance

[1]There are no clearcut boundaries; the colors in this example are only chosen for illustration purpose.

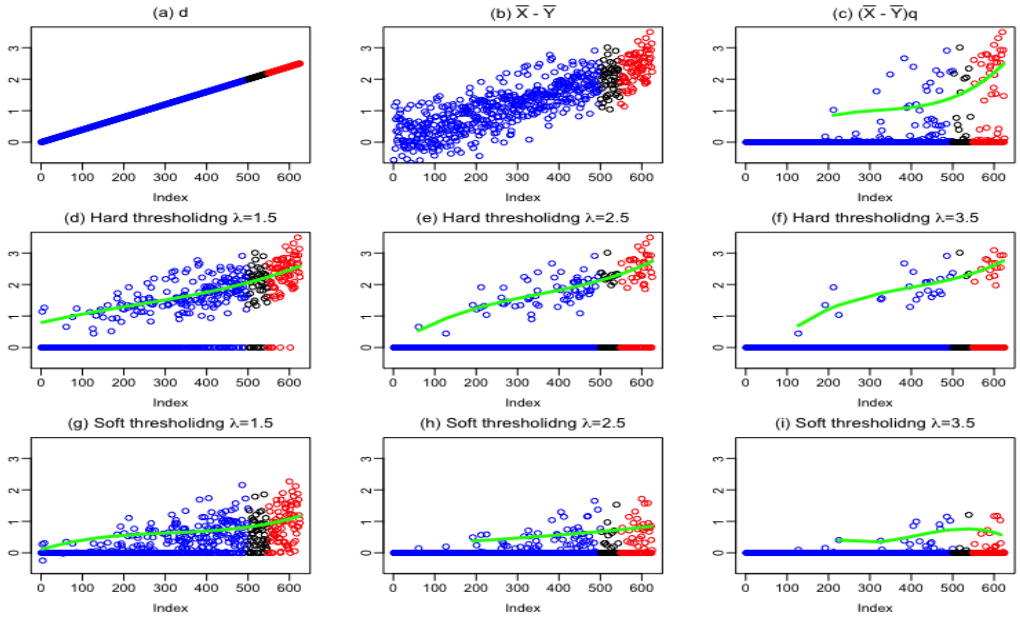

Figure 2: Comparison of the proposed shrinkage rule with thresholding rules. The proposed shrinkage rule exhibits a curved pattern [Panel (c)], which is more effective in eliminating weak signals and keeping strong signals. Green lines are cubic polynomial fits of the points above the horizontal line of 0.2.

between reducing the uncertainties and maintaining useful signals.

S5 Additional Simulations

Let $\boldsymbol{\mu}_1 = (0, ..., 0)^\top \in \mathbb{R}^p$, and $\boldsymbol{\mu}_2$ be a vector with the first $p^{1/2}$ entries being 0.5 and the rest being 0. Consider the following three correlation structures (same as in the main text).

Model 1: Band graph. Let $\Sigma^{-1} = \Omega = (\omega_{ij})_{p\times p}$, where $\omega_{ii} = 1$, $\omega_{i,i+1} = \omega_{i+1,i} = 0.35$, $\omega_{i,i+2} = \omega_{i+2,i} = 0.175$, and $\omega_{ij} = 0$ if $|i - j| > 2$.

Model 2: AR(1) structure. Let $\Sigma^{-1} = \Omega = (\omega_{ij})_{p\times p}$, where $\omega_{ij} = 0.3^{|i-j|}$.

Model 3: Block structure. Let $\Sigma^{-1} = \Omega = (\mathbf{B} + \delta I_p)/(1 + \delta)$, where $b_{ij} = b_{ji} = 0.05 \cdot \text{Bernoulli}(0.1)$ for $1 \le i \le p/2, i < j \le p$, $b_{ij} = b_{ji} = 0.05$ for $p/2 + 1 \le i < j \le p$, $b_{ii} = 1$ for $1 \le i \le p$, and $\delta = \max\{-\lambda_{min}(\mathbf{B}), 0\} + 0.1$.

The size of the training set is $n = 400$, with p varying from 500 to 1000. The mis-classification rate is computed based on $m = 2000$ test points generated from $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma)$ or $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma)$ with equal probability. We repeat the experiment for 100 times, and report the misclassification rates (in percentage) in Table 1. As p varies from 500 to 1000, the proportion of non-zero entries of $\boldsymbol{d}$ varies from 0.045 to 0.031, so this can still be viewed as a sparse setting. It can bee seen that LASS remains competitive.

p	**Oracle**	**Naive**	**LASS**	**LPD**	**AdaLDA**	**Lasso**	**Ebay**
			Model 1				
500	4.97	17.97	**5.32**	8.01	5.73	6.91	5.70
600	4.17	22.11	**4.58**	7.48	5.02	6.21	4.86
700	3.58	28.44	**3.91**	6.78	4.29	5.43	4.16
800	3.01	45.61	**3.41**	8.14	4.83	4.96	3.63
900	2.67	32.09	**3.08**	8.84	3.45	4.49	3.24
1000	2.48	30.05	**2.85**	13.12	3.27	4.23	2.97
			Model 2				
500	5.82	19.32	**6.21**	8.62	6.58	7.84	6.27
600	4.98	23.44	**5.41**	8.11	5.85	7.03	5.50
700	4.29	29.81	**4.65**	7.37	5.00	6.17	4.70
800	3.72	46.62	**4.11**	8.30	4.56	5.76	4.18
900	3.20	32.67	**3.68**	8.24	4.02	5.26	3.73
1000	3.12	30.47	**3.44**	10.93	3.87	4.93	3.47
			Model 3				
500	11.07	26.61	**13.13**	15.27	14.32	14.76	13.21
600	9.79	29.58	**12.29**	14.94	13.63	14.05	12.33
700	9.01	35.20	**11.76**	14.59	13.21	13.48	11.78
800	8.03	47.23	**11.19**	14.42	12.62	13.08	11.25
900	7.27	36.62	**10.80**	15.67	12.40	12.56	10.85
1000	6.90	33.69	**10.88**	14.72	12.63	12.82	11.04

Table 1: Comparison of average misclassification rate in percentage. The smallest error rate (next to that by the oracle) in each setting has been boldfaced.

S6 Lung cancer data

We analyze the lung cancer data (Gordon et al., 2002, available at `https://leo.ugr.es/elvira/DBCRepository/LungCancer/LungCancer-Harvard2.html`), a benchmark data set in high-dimensional classification problems. The data set collects expression levels for $p = 12,533$ genes on 181 tissue samples. Among the 181 samples, 31 are from malignant pleural mesothelioma (MPM) group and 150 are from adenocarcinoma (ADCA) group. The training set contains 16 samples from MPM and 16 samples from ADCA. The testing set contains 134 samples from MPM and 15 samples from ADCA.

We follow the same data pre-processing steps in Cai and Liu (2011). First, the sample variances of individual genes are obtained based on the training data. Next, we drop 195 genes whose sample variances are greater than 10^2 or less than 10^{-2} after rescaling by a factor of 10^4. Finally, to reduce the computational cost (which mainly comes from estimating the precision matrix), only top 200 genes (with the largest absolute values of the two sample t-statistics) are used for constructing different classification rules. Figure 3 illustrates the scores $\hat{T}^j = \frac{\exp(\hat{S}_j)}{1+\exp(\hat{S}_j)}$ estimated by LASS, Naive, Ebay and Lasso. For better illustration, the first 134 testing points are from ADCA group and the next 15 are from MPM group.

To achieve better separation, the values of the first 134 points (the

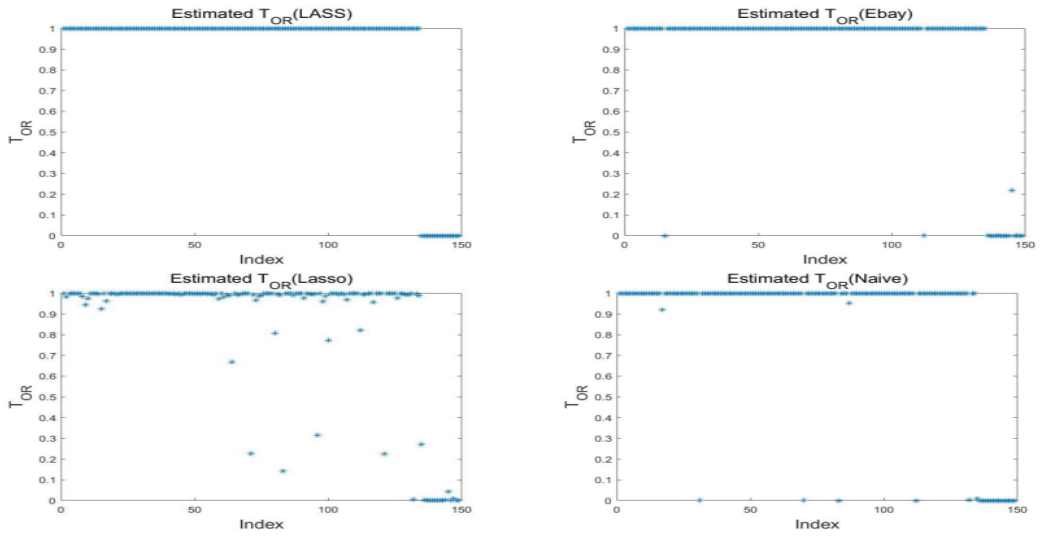

Figure 3: Comparison of estimated T: LASS separates the two classes almost perfectly. Ebay does better than Naive and Lasso but worse than LASS.

Table 2: Number of misclassifications by each method.

	LASS	Naive	LPD	AdaLDA	Lasso	Ebay
Misclassification	**0** (**0**%)	5 (3.36%)	1 (0.67%)	1 (0.67%)	5 (3.36%)	3 (2.01%)

next 15 points) should be high (low). We can see that LASS shows perfect separation of the two classes. While Ebay provides a clearer separation than Naive and Lasso, it is less effective than LASS. For this particular data set, if we choose to control the FSR at level $\alpha = 0.1$, then LASS makes no indecision. Thus, it makes sense to compare the misclassification rates under the classical setup. The results are presented in Table 2, from which we can see that LASS has the best performance.

S7 An example comparing LASS and LPD

In this section we give an example to contrast our theory with existing theories in high-dimensional LDA, where the assumption on the sparsity of $\boldsymbol{d}$ is commonly believed to be "indispensable". For example, Cai and Liu (2011) requires $|\Sigma^{-1}\boldsymbol{d}|_0 = o(\sqrt{(n_1+n_2)/\log p})$ to achieve risk consistency. This sparsity condition seems to be necessary if the scope of analysis is limited to the class of thresholding rules. It is not an artifact of the theoretical analysis, as we can see from the numerical results in Section 5.1.2 where the disadanvantages of existing works are reflected. However, LASS still achieves risk consistency even when the condition in Cai and Liu (2011) is violated. We illustrate this through the next example.

Example 1. Consider an asymptotic setup where $n_1 = n_2 = n,\ p = n^2$, $\Sigma = I_p$ and $\boldsymbol{\mu}_1 = (0,\ldots,0)^\top$. Let $\boldsymbol{\mu}_2 = (1,\ldots,1,0,\ldots,0)^\top$ be a vector with the first k entries being 1 and the rest 0. Let $k = n$ and $\Delta_p = \boldsymbol{d}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{d}$, then in this setting $\Delta_p = n$. It is clear that $|\boldsymbol{\Sigma}^{-1}\boldsymbol{d}|_0 = n$ and $|\boldsymbol{\Sigma}^{-1}\boldsymbol{d}|_1 = n$. Denote $\hat{\boldsymbol{\delta}}_{LPD}$ the LPD rule in Cai and Liu (2011). To guarantee that $R(\hat{\boldsymbol{\delta}}_{LPD}) - R(\boldsymbol{\delta}^F) \to 0$, the theory in Cai and Liu (2011) requires that $|\boldsymbol{\Sigma}^{-1}\boldsymbol{d}|_0 = o(\sqrt{n/\log p})$ or $\frac{|\boldsymbol{\Sigma}^{-1}\boldsymbol{d}|_1}{\Delta_p^{1/2}} + \frac{|\boldsymbol{\Sigma}^{-1}\boldsymbol{d}|_1^2}{\Delta_p^2} = o(\sqrt{n/\log p})$, both of which are violated. By contrast, note that all nonzero elements in $\boldsymbol{d}$ are in $\mathcal{G}_1$ (strong signals), the conditions of Corollary 1 are satisfied for any $k \geq 1$,

which guarantees that LASS achieves risk consistency across sparse and dense settings.

As opposed to existing works that produce approximately the same amount of shrinkage for elements in both $\mathcal{G}_1$ (strong) and $\mathcal{G}_3$ (moderate), LASS adopts an adaptive strategy that makes it possible to provide differential amounts of shrinkage for the elements in $\mathcal{G}_1$ and G_3 (Proposition 1 in Section 3.1 and Figure 2c in Section S4). The resulting shrinkage rule is more effective in keeping strong signals and eliminating noisy elements; this explains the superiority of LASS in both theory and numerical performance.

S8 Class-specific FSR control implies asymptotic global FSR control

We will show if mFSR^c is controlled for $c = 1, 2$ then mFSR is also controlled. The result then follows from Section S3. Recall

$$\text{mFSR}^c = \frac{\mathbb{E}\left\{\sum_{j=1}^m \mathbb{I}(\delta_j = c, \theta_j \neq c)\right\}}{\mathbb{E}\left\{\sum_{j=1}^m \mathbb{I}(\delta_j = c)\right\}}, \quad \text{mFSR} = \frac{\mathbb{E}\left\{\sum_{j=1}^m \mathbb{I}(\delta_j \neq \theta_j, \delta_j \neq 0)\right\}}{\mathbb{E}\left\{\sum_{j=1}^m \mathbb{I}(\delta_j \neq 0)\right\}}.$$

Denote

$$a^c = \mathbb{E}\left\{\sum_{j=1}^m \mathbb{I}(\delta_j = c, \theta_j \neq c)\right\}, \quad b^c = \mathbb{E}\left\{\sum_{j=1}^m \mathbb{I}(\delta_j = c)\right\}.$$

Then $\text{mFSR}^c = a^c/b^c$, and

$$\text{mFSR} = \frac{a^1 + a^2}{b^1 + b^2} = \frac{a^1}{b^1}\frac{b^1}{b^1 + b^2} + \frac{a^2}{b^2}\frac{b^2}{b^2 + b^1} \leq \alpha\frac{b^1}{b^1 + b^2} + \alpha\frac{b^2}{b^1 + b^2} \leq \alpha.$$

S9 Discussion on condition (A2)

For all numerical experiments, we used the ACLIME estimator to estimate precision matrix Σ^{-1}. Suppose we have samples $X_1, \ldots, X_n \sim N(\mu, \Sigma)$. In Cai et al. (2016) the authors proved that if the precision matrix $\Sigma^{-1} \in \mathbb{R}^{p\times p}$ belongs to $\mathcal{G}_q(c_{n,p}, M_{n,p})$:

$$\mathcal{G}_q(c_{n,p}, M_{n,p}) = \left\{ \begin{array}{l} \Omega = (\omega_{ij})_{1\leq i,j\leq p} : \max\limits_j \sum\limits_{i=1}^p |\omega_{ij}|^q \leq c_{n,p}, \\ \\ ||\Omega||_1 \leq M_{n,p}, \lambda_{max}(\Omega)/\lambda_{min}(\Omega) \leq M_1, \Omega \succ 0 \end{array} \right\}.$$

Under some regularity conditions on p, n, $c_{n,p}$ the ACLIME estimator satisfies

$$\sup_{\mathcal{G}_q(c_{n,p}, M_{n,p})} \mathbb{E}\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_w^2 = M_{n,p}^{2-2q} c_{n,p} (\log p/n)^{1-q}$$

for all $1 \leq w \leq \infty$ and $0 \leq q < 1$. Hence if $M_{n,p}$ and $c_{n,p}$ are chosen such that

$$M_{n,p}^{2-2q} c_{n,p} (\log p/n)^{1-q} = o(1),$$

then the ACLIME estimator satisfies condition (A2).

We emphasize that our Condition (A2) is strictly weaker than the condition $\Sigma^{-1} \in \mathcal{G}_q(c_{n,p}, M_{n,p})$ considered in Cai et al. (2016). Moreover, LASS does not depend on any particular class of precision matrix estimators nor do we regard the parameter space $\mathcal{G}_q(c_{n,p}, M_{n,p})$ to be of any specific type. Instead, a range of estimators, as proposed in Yuan (2010), Liu and Luo (2015),Cai et al. (2016) and Avella-Medina et al. (2018), all have good empirical performances and are consistent under various conditions. These estimators can be employed to construct LASS classifiers too.

In practice we often have some prior knowledge on the covariance structure, and we can choose which estimator to use according to this prior knowledge. For example, if we know that the covariance matrix has an AR(1) structure (i.e. $\Sigma = (\sigma_{ij})$, with $\sigma_{ij} = \rho^{|i-j|}$) then the inverse of the MLE of the covariance matrix $\hat{\Sigma}_{mle}^{-1}$ satisfies (A2). Hence, we do not need to assume Σ^{-1} belongs to any particular parameter space.

Estimating precision matrix is still an active area of research. It is conceivable (and hopeful) that future works on this topic can further relax the sufficient conditions under which an existing estimator is consistent. It would also be of interest to propose new estimators that are consistent in a wider range of parameter spaces.

Bibliography

Avella-Medina, M., H. S. Battey, J. Fan, and Q. Li (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika 105*(2), 271–284.

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association 106*(496), 1566–1577.

Cai, T. T. (2002). On block thresholding in wavelet regression: Adaptivity, block size, and threshold level. *Statistica Sinica*, 1241–1273.

Cai, T. T., W. Liu, H. H. Zhou, et al. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics 44*(2), 455–488.

Foygel, R. and M. Drton (2010). Extended bayesian information criteria for gaussian graphical models. *Advances in neural information processing systems 23*.

Gordon, G. J., R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno (2002). Translation of microarray data into clinically relevant cancer diagnostic

tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research 62*(17), 4963–4967.

Liu, W. and X. Luo (2015). Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of multivariate analysis 135*, 153–162.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research 11*, 2261–2286.