Web Appendices

for

Multiclass Linear Discriminant Analysis with Ultrahigh-Dimensional Features

by

Yanming Li, Hyokyoung G. Hong and Yi Li

S1. Lemmas and Proofs

We present auxiliary lemmas, along with their proofs, that are useful in proving the main theorems. The proofs for the main theorems then follow. Theorems 1, 2 and 3 are proved for any distributions satisfying (A1), which include multivariate Gaussian and t distributions as special cases. Theorem 4 holds for multivariate Gaussian distribution.

Bickel and Levina (2008) showed that for a thresholding parameter $\alpha = O((\log C_{\max}/n)^{1/2})$, $\|\widetilde{\mathbf{\Sigma}}_l - \mathbf{\Sigma}_l\| = O_P(\rho_n)$ with $\widetilde{\mathbf{\Sigma}}_l$ and $\mathbf{\Sigma}_l$ being the principle submatrices of $\widetilde{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ corresponding to $\widehat{\mathcal{C}}_l$, respectively, $l = 1, \ldots, B$ and $\rho_n = C_t(n^{-1}C_{\max}^{t/4})$ defined in (A11). Furthermore, Bickel and Levina (2008) and Fan et al. (2011) showed that the precision component can be estimated from the thresholded sample covariance matrix with an estimation error bound $\|(\widetilde{\mathbf{\Sigma}}_l)^{-1} - \mathbf{\Sigma}_l^{-1}\| = O_P(\rho_n)$ for any $l = 1, \ldots, B$. This yields an overall precision estimation error bound

$$\|\widehat{\mathbf{\Omega}}^u - \mathbf{\Omega}^u\| \leqslant O_P(\rho_n), \tag{S5}$$

where $\mathbf{\Omega}^u$ is the principle submatrix of $\mathbf{\Omega}$ with row and column indices restricted to $\mathcal{U}$.

Lemma 1 (Bernstein's inequality, van der Vaart and Wellner (1996)): *Let $Z_1, \ldots, Z_n$ be independent random variables with mean zero. Assume $E|Z_i|^m \leqslant m!M^{m-2}v_i/2$, for every $m \geqslant 2$, $1 \leqslant i \leqslant n$ and some positive constants M and v_i. Then for any $x > 0$,*

$$P(|Z_1 + \cdots + Z_n| > x) \leqslant 2\exp\{-x^2/(2(v + Mx))\}$$

for $v \geqslant v_1 + \cdots + v_n$.

Lemma 2: *Suppose that condition (A5) holds and* $\mathbf{X}$ *is a multivariate random vector with each of its components satisfying (A1). For any* $(j, j') \in \mathcal{E}$*, there are positive constants* C_1 *and* C_2*, such that*

$$P\Big(|\mathbf{X}_j'\mathbf{X}_{j'}|/n \leqslant Cn^{(\xi-1)/2}\Big) \leqslant C_1 \exp(-C_2 n^{1+\xi}).$$

On the other hand, for any $(j, j') \notin \mathcal{E}$*, there are positive constants* $\widetilde{C}_1$ *and* $\widetilde{C}_2$*, such that*

$$P\Big(|\mathbf{X}_j'\mathbf{X}_{j'}|/n > Cn^{(\xi-1)/2}\Big) \leqslant \widetilde{C}_1 \exp(-\widetilde{C}_2 n^{1+\xi}).$$

Lemma 2 ensures that, under conditions (A1) and (A5), and with an appropriately chosen α, if (j, j') is an edge in $\mathbf{\Omega}$, the probability of (j, j') not being an edge in $\widetilde{\mathbf{\Sigma}}$ is asymptotically zero, and that for a non-edge pair (j, j') in $\mathbf{\Omega}$, the probability that it is an edge in $\widetilde{\mathbf{\Sigma}}$ is asymptotically zero.

Proof of Lemma 2.

When $(j, j') \in \mathcal{E}$, by condition (A5), $|\Sigma_{jj'}| \geqslant \min_{(j,j')\in\mathcal{E}} |\Sigma_{jj'}| = c_1 n^{(\xi-1)/2}$ for some $c_1 > 0$. Therefore, when $|\mathbf{X}_j'\mathbf{X}_{j'}|/n \leqslant Cn^{(\xi-1)/2}$ for some $0 < C < c_1$, $\Sigma_{jj'} - \mathbf{X}_j'\mathbf{X}_{j'}/n \leqslant Cn^{(\xi-1)/2} - c_1 n^{(\xi-1)/2}$ or $\Sigma_{jj'} - \mathbf{X}_j'\mathbf{X}_{j'}/n \geqslant -Cn^{(\xi-1)/2} + c_1 n^{(\xi-1)/2}$. That is, $\left|\mathbf{X}_j'\mathbf{X}_{j'}/n - \Sigma_{jj'}\right| > \widetilde{C} n^{(\xi-1)/2}$ for a positive constant $\widetilde{C} = c_1 - C$. Therefore,

$$P\left(\left|\mathbf{X}_j'\mathbf{X}_{j'}\right|/n \leqslant Cn^{(\xi-1)/2}\right) \leqslant P\left(\left|\mathbf{X}_j'\mathbf{X}_{j'}/n - \Sigma_{jj'}\right| > \widetilde{C} n^{(\xi-1)/2}\right).$$

For simplicity, assume that $\mathbf{X}$ has mean zero and a common marginal variance σ^2. Then

$$
\begin{aligned}
& P\left(\left|\mathbf{X}_j'\mathbf{X}_{j'}/n - \Sigma_{jj'}\right| > \widetilde{C}n^{-(1-\xi)/2}\right) = P\left(\left|\sum\nolimits_{i=1}^{n}(X_{ij}X_{ij'} - \Sigma_{jj'})\right| > \widetilde{C}n^{(1+\xi)/2}\right) \\
= \;& P\Big(\Big|\sum\nolimits_{i=1}^{n}\left\{(X_{ij}+X_{ij'})^2 - 2(\sigma^2+\Sigma_{jj'})\right\} - \sum\nolimits_{i=1}^{n}\left\{(X_{ij}-X_{ij'})^2 - 2(\sigma^2-\Sigma_{jj'})\right\}\Big| \\
& \geqslant 4\widetilde{C}n^{(1+\xi)/2}\Big) \\
\leqslant \;& P\Big(\Big|\sum\nolimits_{i=1}^{n}\left\{(X_{ij}+X_{ij'})^2 - 2(\sigma^2+\Sigma_{jj'})\right\}\Big| + \Big|\sum\nolimits_{i=1}^{n}\left\{(X_{ij}-X_{ij'})^2 - 2(\sigma^2-\Sigma_{jj'})\right\}\Big| \\
& \geqslant 4\widetilde{C}n^{(1+\xi)/2}\Big) \\
\leqslant \;& P\left(\left|\sum\nolimits_{i=1}^{n}\left[(X_{ij}+X_{ij'})^2 - 2(\sigma^2+\Sigma_{jj'})\right]\right| \geqslant 2\widetilde{C}n^{(1+\xi)/2}\right) + \\
& P\left(\left|\sum\nolimits_{i=1}^{n}\left[(X_{ij}-X_{ij'})^2 - 2(\sigma^2-\Sigma_{jj'})\right]\right| \geqslant 2\widetilde{C}n^{(1+\xi)/2}\right).
\end{aligned}
$$

Since $(X_{ij}+X_{ij'})^2-2(\sigma^2+\Sigma_{jj'})$ and $(X_{ij}-X_{ij'})^2-2(\sigma^2-\Sigma_{jj'})$ are both random variables with mean zero and satisfy (A1), by Lemma 1, $P(|\sum_{i=1}^{n}[(X_{ij}+X_{ij'})^2-2(\sigma^2+\Sigma_{jj'})]| \geqslant 2\widetilde{C}n^{(1+\xi)/2})$ and $P(|\sum_{i=1}^{n}[(X_{ij}-X_{ij'})^2-2(\sigma^2-\Sigma_{jj'})]| \geqslant 2\widetilde{C}n^{(1+\xi)/2})$ are bounded by $C_1\exp(-C_2n^{(1+\xi)/2})$ for some positive C_1 and C_2. Thus,

$$
P\left(\left|\mathbf{X}_j'\mathbf{X}_{j'}\right|/n \leqslant Cn^{(\xi-1)/2}\right) \leqslant C_1\exp(-C_2n^{(1+\xi)/2}).
$$

When $(j,j') \notin \mathcal{E}$, $|\Sigma_{jj'}| \leqslant \max_{(j,j')\notin\mathcal{E}}|\Sigma_{jj'}| = o(n^{(\xi-1)/2})$ is given in (A5). Therefore, when $|\mathbf{X}_j'\mathbf{X}_{j'}|/n > Cn^{(\xi-1)/2}$, we have either $\mathbf{X}_j'\mathbf{X}_{j'}/n - \Sigma_{jj'} < -Cn^{(\xi-1)/2} - o(n^{(\xi-1)/2})$ or $\mathbf{X}_j'\mathbf{X}_{j'}/n - \Sigma_{jj'} > Cn^{(\xi-1)/2} + o(n^{(\xi-1)/2})$. That is, $\left|\mathbf{X}_j'\mathbf{X}_{j'}/n - \Sigma_{jj'}\right| > \widetilde{C}n^{(\xi-1)/2}$ for a positive constant $\widetilde{C} = C + o(1)$. Therefore,

$$
P\left(\left|\mathbf{X}_j'\mathbf{X}_{j'}\right|/n > Cn^{(\xi-1)/2}\right) \leqslant P\left(\left|\mathbf{X}_j'\mathbf{X}_{j'}/n - \Sigma_{jj'}\right| > \widetilde{C}n^{(\xi-1)/2}\right).
$$

The rest of the proof is similar to the first part and is therefore omitted. □

The following Lemma 3 ensures selection consistency for the MI features. Let $\mathcal{S}_1(k,k') = \{1 \leqslant m \leqslant p : \mu_{km} - \mu_{k'm} \neq 0\}$ for any $1 \leqslant k < k' \leqslant K$.

Lemma 3: *For any pair of classes k and k', $1 \leqslant k < k' \leqslant K$, if j belongs to $\mathcal{S}_1(k,k')$, then for a thresholding parameter $\tau = O((r\log p)^s)$, $0 < s < 1/2$, and a sufficiently large n,* *there exist positive constants C_1 and C_2, such that*

$$P\left(|\overline{X}_{\cdot j}^{(k)} - \overline{X}_{\cdot j}^{(k')}| \leqslant \tau\right) < C_1 \exp(-C_2 n^{1-\varsigma/2} r \log p) \to 0$$

for ς given in (A3). On the other hand, if $j \notin \mathcal{S}_1(k,k')$, then for sufficiently large n, there exist positive constants $\widetilde{C}_1$ and $\widetilde{C}_2$, such that

$$P\left(|\overline{X}_{\cdot j}^{(k)} - \overline{X}_{\cdot j}^{(k')}| \geqslant \tau\right) < \widetilde{C}_1 \exp(-\widetilde{C}_2 n^{1-\varsigma/2} r \log p) \to 0.$$

Proof of Lemma 3. For the first statement, notice that if $j \in \mathcal{S}_1(k,k')$, then $|\mu_{kj} - \mu_{k'j}| > \tau^* = \sqrt{r \log p}$, where r is given in (A3). Then when choosing the thresholding parameter $\tau = O((r\log p)^s)$, $0 < s < 1/2$, $|\overline{X}_{\cdot j}^{(k)} - \overline{X}_{\cdot j}^{(k')}| \leqslant \tau$ and $|\mu_{kj} - \mu_{k'j}| > \tau^*$ together give that $|\overline{X}_{\cdot j}^{(k)} - \overline{X}_{\cdot j}^{(k')} - (\mu_{kj} - \mu_{k'j})| > c$, where $c = \tau^* - \tau = O(\sqrt{r\log p})$, which means that for some positive constant M_0 and sufficiently large n and p, $c \geqslant M_0\sqrt{r \log p}$. Therefore, when $j \in \mathcal{S}_1(k,k')$,

$$\begin{aligned}
&P\left(|\overline{X}_{\cdot j}^{(k)} - \overline{X}_{\cdot j}^{(k')}| \leqslant \tau\right) \leqslant P\left(|\overline{X}_{\cdot j}^{(k)} - \overline{X}_{\cdot j}^{(k')} - (\mu_{kj} - \mu_{k'j})| > c\right) \\
\leqslant\ & P\left(|\overline{X}_{\cdot j}^{(k)} - \overline{X}_{\cdot j}^{(k')} - (\mu_{kj} - \mu_{k'j})| > M_0\sqrt{r\log p}\right) \\
\leqslant\ & P\left(|\overline{X}_{\cdot j}^{(k)} - \mu_{kj}| + |\overline{X}_{\cdot j}^{(k')} - \mu_{k'j}| > M_0\sqrt{r\log p}\right) \\
\leqslant\ & P\left(|\overline{X}_{\cdot j}^{(k)} - \mu_{kj}| > M_0\sqrt{r\log p}/2\right) + P\left(|\overline{X}_{\cdot j}^{(k')} - \mu_{k'j}| > M_0\sqrt{r\log p}/2\right). \qquad \text{(S6)}
\end{aligned}$$

Then with the constants c_1 given in (A8), C given in (A1) and ς given in (A3), we have that

$$\begin{aligned}
&P\left(|\overline{X}_{\cdot j}^{(k)} - \mu_{kj}| > M_0\sqrt{r\log p}/2\right) = P\left(\frac{1}{n_k}\left|\sum\nolimits_{i:Y_i=k}(X_{ij} - \mu_{kj})\right| > M_0\sqrt{r\log p}/2\right) \\
&= P\left(\left|\sum\nolimits_{i:Y_i=k}(X_{ij} - \mu_{kj})\right| > M_0 n_k\sqrt{r\log p}/2\right) \\
&\leqslant P\left(\left|\sum\nolimits_{i:Y_i=k}(X_{ij} - \mu_{kj})\right| > \frac{M_0 c_1 n\sqrt{r\log p}}{2}\right) \\
&\leqslant 2\exp\left(\frac{-M_0^2 c_1^2 n^2 r \log p}{8(nC^2 + CM_0c_1n\sqrt{r\log p}}\right) \\
&= 2\exp\left(\frac{-M_0^2 c_1^2 n^2 r \log p}{8(nC^2 + CM_0c_1\sqrt{r}n^{1+\varsigma/2}}\right) \\
&\leqslant 2\exp\left(-C_2 n^{1-\varsigma/2} r\log p\right), \qquad \text{(S7)}
\end{aligned}$$

where $C_2 = M_0^2c_1^2/8(C^2 + CM_0c_1\sqrt{r})$ is a constant and the third to last step is from Lemma

1. With the same argument, the second term in (S6) also satisfies

$$P\left(|\overline{X}_{\cdot j}^{(k')} - \mu_{kj}| > M_0\sqrt{r\log p}/2\right) \leqslant 2\exp\left(-C_2 n^{1-\varsigma/2} r\log p\right). \tag{S8}$$

Putting together (S7) and (S8) concludes the first statement. The second statement follows by a similar argument. □

Lemma 4 (Hoeffding's inequality for dependent random variables, van de Geer (2002)): *Consider a probability triplet $(\Omega, \mathcal{F}, P)$ and let $\emptyset = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$ be an increasing sequence of σ-algebras. Let X_i be an $\mathcal{F}_i$-measurable random variables satisfying $E(X_i|\mathcal{F}_i) = 0$ a.s. Consider the martingale $S_n = \sum_{i=1}^n X_i$, $n \geqslant 1$. Let $K_i > 0$ be $\mathcal{F}_{i-1}$ random variables $i = 1, \ldots, n$. Define $B_0^2 = 0$ and for $n \geqslant 1$,*

$$B_n^2 = \sum\nolimits_{i=1}^n K_i^2 \left(1 + E\left(\Psi\left(\frac{|X_i|}{K_i}\right)|\mathcal{F}_{i-1}\right)\right)$$

with $\Psi(x) = \exp(x^2) - 1$. Then for all $a > 0$ and $b > 0$, and for some n,

$$P(\mathcal{S}_n \geqslant a \text{ and } B_n^2 \leqslant b^2) \leqslant \exp\left(-\frac{a^2}{8b^2}\right).$$

Let $\boldsymbol{\delta}_{kk'} = \boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}$, $\widehat{\boldsymbol{\mu}}_k = n_k^{-1}\sum_{i:Y_i=k}\boldsymbol{X}_i$ and $\widehat{\boldsymbol{\delta}}_{kk'} = \widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_{k'}$, $1 \leqslant k < k' \leqslant K$. Also let $\boldsymbol{\Sigma}^{\mathrm{s}}$, $\boldsymbol{\Omega}^{\mathrm{s}}$, $\boldsymbol{\mu}_k^{\mathrm{s}}$, $\boldsymbol{\delta}_{kk'}^{\mathrm{s}}$, $\widehat{\boldsymbol{\mu}}_k^{\mathrm{s}}$, $\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}}$ and $\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}$ denote the restricted true covariance matrix, true precision matrix, true mean vector, true mean difference vector, estimated mean vector, estimated mean difference vector and estimated precision matrix on $\widehat{\mathcal{S}}_0$, respectively, by keeping values of entries indexed by $\widehat{\mathcal{S}}_0$ and setting the other entries to zeros. Similarly, let $\boldsymbol{\Sigma}^0$, $\boldsymbol{\Omega}^0$, $\boldsymbol{\mu}_k^0$, $\boldsymbol{\delta}_{kk'}^0$, $\widehat{\boldsymbol{\mu}}_k^0$, $\widehat{\boldsymbol{\delta}}_{kk'}^0$ and $\widehat{\boldsymbol{\Omega}}^0$ denote the corresponding matrices/vectors restricted to $\mathcal{S}_0$.

When $\boldsymbol{X}$ follows multivariate Gaussian distribution, we have the following two lemmas.

Lemma 5: *For any pair of classes (k, k'), $1 \leqslant k < k' \leqslant K$, conditioning on the event $\{\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0\}$, we have that*

$$\begin{aligned}(\widehat{\boldsymbol{\delta}}_{kk'}^s)'\boldsymbol{\Omega}^s\widehat{\boldsymbol{\delta}}_{kk'}^s &= (\boldsymbol{\delta}_{kk'}^0)'\boldsymbol{\Omega}^0\boldsymbol{\delta}_{kk'}^0 + O_P\left(\sqrt{\Delta_p(k,k')\frac{|\mathcal{S}_0|(n_k + n_{k'})}{n_k n_{k'}}}\right) + \\ & O_P\left(\frac{|\mathcal{S}_0|(n_k + n_{k'})}{n_k n_{k'}}\right) + O_P(1). \end{aligned} \tag{S9}$$

Proof of Lemma 5. Notice that

$$(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'} = (\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{0}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'} + (\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'(\boldsymbol{\Omega}^{\mathrm{s}}-\boldsymbol{\Omega}^{0})\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}$$
$$= (\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}\boldsymbol{\delta}^{0}_{kk'} + 2(\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\boldsymbol{\delta}^{0}_{kk'}) + (\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\boldsymbol{\delta}^{0}_{kk'}) +$$
$$(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'(\boldsymbol{\Omega}^{\mathrm{s}}-\boldsymbol{\Omega}^{0})\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}. \tag{S10}$$

The third term $(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})$ on the right hand side of (S10) can be expressed as

$$(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})$$
$$= (\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'}) + 2(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\widehat{\boldsymbol{\delta}}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})$$
$$+(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\widehat{\boldsymbol{\delta}}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\widehat{\boldsymbol{\delta}}^{0}_{kk'}). \tag{S11}$$

For the term $(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})$ in (S11), assume the singular value decomposition of $\boldsymbol{\Sigma}^0 = \mathbf{D}\boldsymbol{\Lambda}_{\mathcal{S}_0}\mathbf{D}'$, where $\mathbf{D}$ is an orthogonal matrix and $\boldsymbol{\Lambda}_{\mathcal{S}_0} = \mathrm{diag}(\lambda_1,\ldots,\lambda_{|\mathcal{S}_0|})$ are the eigenvalues of $\boldsymbol{\Sigma}^0$. Since $\sqrt{\frac{n_k n_{k'}}{n_k+n_{k'}}}(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'}) \sim N(0,\boldsymbol{\Sigma}^0)$ and $\boldsymbol{\Sigma}^0\boldsymbol{\Omega}^0 = \mathbf{I}_{|\mathcal{S}_0|}$ with $\mathbf{I}_{|\mathcal{S}_0|}$ being the identity matrix of dimension $|\mathcal{S}_0|$, then

$$\boldsymbol{\epsilon} \equiv \sqrt{\frac{n_k n_{k'}}{n_k+n_{k'}}}\boldsymbol{\Lambda}_{\mathcal{S}_0}^{-\frac{1}{2}}\mathbf{D}'(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'}) \sim N(0,\mathbf{I}_{|\mathcal{S}_0|}). \tag{S12}$$

Moreover,

$$\frac{n_k n_{k'}}{n_k+n_{k'}}(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'}) = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}. \tag{S13}$$

When $n \to \infty$ and $p \to \infty$, $|\mathcal{S}_0| > p^{1-\beta}$ goes to infinity. By the law of large numbers,

$$\frac{n_k n_{k'}}{|\mathcal{S}_0|(n_k+n_{k'})}(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'}) \xrightarrow{P} E(\epsilon^2) = 1 \quad \text{as} \;\; n\to\infty, \;\; p\to\infty,$$

where ϵ^2 is a χ^2 distributed random variable with degree of freedom 1. This gives that

$$(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{0}_{kk'}-\boldsymbol{\delta}^{0}_{kk'}) = O_P\left(\frac{|\mathcal{S}_0|(n_k+n_{k'})}{n_k n_{k'}}\right). \tag{S14}$$

For the term $(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\widehat{\boldsymbol{\delta}}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\widehat{\boldsymbol{\delta}}^{0}_{kk'})$ in (S11), recall that $\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}$ and $\widehat{\boldsymbol{\delta}}^{0}_{kk'}$ are $\widehat{\boldsymbol{\delta}}_{kk'}$ restricted on $\widehat{\mathcal{S}}_0$ and $\mathcal{S}_0$, respectively, and $\boldsymbol{\Omega}^0$ is the true precision matrix restricted to $\mathcal{S}_0$. It follows that when $\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0$, $(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\widehat{\boldsymbol{\delta}}^{0}_{kk'})'\boldsymbol{\Omega}^{0}(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\widehat{\boldsymbol{\delta}}^{0}_{kk'}) = 0$, as entries of $(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}-\widehat{\boldsymbol{\delta}}^{0}_{kk'})$ equal 0 on $\mathcal{S}_0$

while entries of $\mathbf{\Omega}^0$ equal 0 on $\mathcal{S}_0^c$. For the same reason, we also have that conditioning on the event $\{\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0\}$, $(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} - \widehat{\boldsymbol{\delta}}_{kk'}^{0})'\mathbf{\Omega}^0(\widehat{\boldsymbol{\delta}}_{kk'}^{0} - \boldsymbol{\delta}_{kk'}^{0}) = 0$. Plugging these two zero terms and (S14) into (S11), we have that conditioning on the event $\{\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0\}$,

$$(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} - \boldsymbol{\delta}_{kk'}^{0})'\mathbf{\Omega}^0(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} - \boldsymbol{\delta}_{kk'}^{0}) = O_P\left(\frac{|\mathcal{S}_0|(n_k + n_{k'})}{n_k n_{k'}}\right). \tag{S15}$$

For the second term $(\boldsymbol{\delta}_{kk'}^{0})'\mathbf{\Omega}^0(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} - \boldsymbol{\delta}_{kk'}^{0})$ on the right hand sides of (S10), when conditioning on the event $\{\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0\}$, we have

$$\begin{aligned}(\boldsymbol{\delta}_{kk'}^{0})'\mathbf{\Omega}^0(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} - \boldsymbol{\delta}_{kk'}^{0}) &\leqslant \sqrt{(\boldsymbol{\delta}_{kk'}^{0})'\mathbf{\Omega}^0\boldsymbol{\delta}_{kk'}^{0}}\sqrt{(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} - \boldsymbol{\delta}_{kk'}^{0})'\mathbf{\Omega}^0(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} - \boldsymbol{\delta}_{kk'}^{0})} \\ &= \sqrt{\Delta_p^2(k,k')}\sqrt{O_P\left(\frac{|\mathcal{S}_0|(n_k + n_{k'})}{n_k n_{k'}}\right)} \\ &= O_P\left(\sqrt{\Delta_p^2(k,k')\left(\frac{|\mathcal{S}_0|(n_k + n_{k'})}{n_k n_{k'}}\right)}\right). \end{aligned} \tag{S16}$$

For the last term $(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}})'(\mathbf{\Omega}^{\mathrm{s}} - \mathbf{\Omega}^0)\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}}$ on the right hand side of (S10), notice that $(\mathbf{\Omega}^{\mathrm{s}} - \mathbf{\Omega}^0)$ only has nonzero values on the set $\widehat{\mathcal{S}}_0 \cap \mathcal{S}_0^c$. Therefore,

$$(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}})'(\mathbf{\Omega}^{\mathrm{s}} - \mathbf{\Omega}^0)\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} \leqslant 2\kappa_1^{-1}\sum\nolimits_{j\in\widehat{\mathcal{S}}_0\cap\mathcal{S}_0^c}|\widehat{\delta}_j|^2 \leqslant 2\kappa_1^{-1}|\widehat{\mathcal{S}}_0|\max_{j\in\mathcal{S}_0^c}|\widehat{\delta}_j|^2. \tag{S17}$$

According to the mLDA procedure, $|\widehat{\mathcal{S}}_0| \leqslant |\mathcal{U}| \leqslant K^2 n$ and when $j \in \mathcal{S}_0^c$, $E(\widehat{\delta}_j) = 0$ and $\mathrm{var}(\widehat{\delta}_j) = \sigma_j^2(1/n_k + 1/n_{k'})$ with σ_j^2 being the marginal variance of feature j. Therefore, $|\widehat{\delta}_j|^2 = O_P(n^{-1})$ for any j. As a result,

$$(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}})'(\mathbf{\Omega}^{\mathrm{s}} - \mathbf{\Omega}^0)\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} = O_P(\kappa_1^{-1}K^2) = O_P(1). \tag{S18}$$

Putting (S15), (S16) and (S18) into (S10), we have that conditioning on the event $\{\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0\}$,

$$\begin{aligned}(\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}})'\mathbf{\Omega}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}_{kk'}^{\mathrm{s}} &= (\boldsymbol{\delta}_{kk'}^{0})'\mathbf{\Omega}^0\boldsymbol{\delta}_{kk'}^{0} + O_P\left(\sqrt{\Delta_p^2(k,k')\frac{|\mathcal{S}_0|(n_k + n_{k'})}{n_k n_{k'}}}\right) \\ &\quad + O_P\left(\frac{|\mathcal{S}_0|(n_k + n_{k'})}{n_k n_{k'}}\right) + O_P(1).\end{aligned}$$

□

Lemma 6: *For an $\mathbf{X}$ from a class k and any pair of classes (k, k'), $1 \leqslant k < k' \leqslant K$, let $\widehat{\boldsymbol{\mu}}^s_{kk'} = (\widehat{\boldsymbol{\mu}}^s_k + \widehat{\boldsymbol{\mu}}^s_{k'})/2$. Then conditioning on the event $\{\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0\}$, we also have*

$$\begin{aligned}(\boldsymbol{\mu}^s_k - \widehat{\boldsymbol{\mu}}^s_{kk'})'\boldsymbol{\Omega}^s\widehat{\boldsymbol{\delta}}^s_{kk'} &= \frac{1}{2}(\boldsymbol{\delta}^0_{kk'})'\boldsymbol{\Omega}^0\boldsymbol{\delta}^0_{kk'} + O_P\left(\sqrt{\Delta_p^2(k,k')|\mathcal{S}_0|/n_{k'}}\right) \\ &\quad + O_P\left(\frac{|\mathcal{S}_0|(n_k + n_{k'})}{n_k n_{k'}}\right) + O_P(1).\end{aligned}$$

Proof of Lemma 6. Direct calculation shows that

$$\begin{aligned}(\boldsymbol{\mu}^{\mathrm{s}}_k - \widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'} &= \frac{1}{2}(\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}\boldsymbol{\delta}^{\mathrm{s}}_{kk'} - (\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'}) \\ &\quad -\frac{1}{2}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k)'\boldsymbol{\Omega}^{\mathrm{s}}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k) + \frac{1}{2}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'})'\boldsymbol{\Omega}^{\mathrm{s}}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'}) \\ &= \frac{1}{2}(\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}\boldsymbol{\delta}^{\mathrm{s}}_{kk'} - A_1 - \frac{1}{2}A_2 + \frac{1}{2}A_3, \end{aligned} \tag{S19}$$

where $A_1 \equiv (\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'})$, $A_2 \equiv (\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k)'\boldsymbol{\Omega}^{\mathrm{s}}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k)$ and $A_3 \equiv (\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'})'\boldsymbol{\Omega}^{\mathrm{s}}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'})$. For A_2, notice that

$$\begin{aligned}&(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k)'\boldsymbol{\Omega}^{\mathrm{s}}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k) \\ &= (\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k)'(\boldsymbol{\Omega}^{\mathrm{s}} - \boldsymbol{\Omega}^0)(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k) + (\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k)'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k).\end{aligned}$$

By the same argument as in (S17), we have that

$$(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k)'(\boldsymbol{\Omega}^{\mathrm{s}} - \boldsymbol{\Omega}^0)(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k) \leqslant \kappa_1^{-1}\sum\nolimits_{j \in \widehat{\mathcal{S}}_0 \cap \mathcal{S}_0^c} |\widehat{\delta}_j|^2 = O_P(1)$$

and $(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k)'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \boldsymbol{\mu}^{\mathrm{s}}_k) = (\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k)'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k)$. Since $\sqrt{n_k}(\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k) \sim N(0, \boldsymbol{\Sigma}^0)$, similar to (S12), define $\widetilde{\boldsymbol{\epsilon}} = \sqrt{n_k}\boldsymbol{\Lambda}_{\mathcal{S}_0}^{-1/2}\mathbf{D}'(\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k)$. Then $\widetilde{\boldsymbol{\epsilon}} \sim N(0, \mathbf{I}_{|\mathcal{S}_0|})$ and $\sqrt{n_k}(\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k)'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k) = \widetilde{\boldsymbol{\epsilon}}'\widetilde{\boldsymbol{\epsilon}}$. By the law of large numbers, $(n_k/|\mathcal{S}_0|)(\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k)'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k) \xrightarrow{P} 1$. Therefore $(\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k)'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^0_k - \boldsymbol{\mu}^0_k) = O_P(|\mathcal{S}_0|/n_k)$ and

$$A_2 = O_P\left(\frac{|\mathcal{S}_0|}{n_k}\right) + O_P(1). \tag{S20}$$

Following the same argument, we have

$$A_3 = O_P\left(\frac{|\mathcal{S}_0|}{n_{k'}}\right) + O_P(1). \tag{S21}$$

For A_1, note that $(\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'}) = (\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'}) + (\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'(\boldsymbol{\Omega}^{\mathrm{s}} - \boldsymbol{\Omega}^0)(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'})$. For the second term above, following the same argument as before, when $\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0$, $(\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'(\boldsymbol{\Omega}^{\mathrm{s}} -$

$\boldsymbol{\Omega}^0)(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'}) = 0$ as $\boldsymbol{\delta}^{\mathrm{s}}_{kk'}$ only takes non-zero values for entries indexed within $\mathcal{S}_0$, while $(\boldsymbol{\Omega}^{\mathrm{s}} - \boldsymbol{\Omega}^0)$ is zero valued for entries indexed within $\mathcal{S}_0$. Therefore, $(\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'}) = (\boldsymbol{\delta}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{k'} - \boldsymbol{\mu}^{\mathrm{s}}_{k'}) = (\boldsymbol{\delta}^0_{kk'})'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^0_{k'} - \boldsymbol{\mu}^0_{k'})$. Thus

$$\begin{aligned} A_1 &\leqslant \sqrt{(\boldsymbol{\delta}^0_{kk'})'\boldsymbol{\Omega}^0\boldsymbol{\delta}^0_{kk'}}\sqrt{(\widehat{\boldsymbol{\mu}}^0_{k'} - \boldsymbol{\mu}^0_{k'})'\boldsymbol{\Omega}^0(\widehat{\boldsymbol{\mu}}^0_{k'} - \boldsymbol{\mu}^0_{k'})} \\ &= \sqrt{\Delta^2_p(k,k')}O_p\left(\sqrt{|\mathcal{S}_0|/n_{k'}}\right). \end{aligned} \tag{S22}$$

Plugging (S22), (S20) and (S21) into (S19), we have

$$\begin{aligned} (\widehat{\boldsymbol{\mu}}^{\mathrm{s}}_k - \widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'} &= \frac{1}{2}(\boldsymbol{\delta}^0_{kk'})'\boldsymbol{\Omega}^0\boldsymbol{\delta}^0_{kk'} + O_P\left(\sqrt{\Delta^2_p(k,k')|\mathcal{S}_0|/n_{k'}}\right) \\ &+ O_P\left(\frac{|\mathcal{S}_0|(n_k + n_{k'})}{n_k n_{k'}}\right) + O_P(1). \end{aligned}$$

□

S2. Proofs of the main theorems

Proof of Theorem 1. First we show that

$$P\left(\mathcal{C}_{[j]} \subseteq \widehat{\mathcal{C}}_{[j]}\right) \geqslant 1 - C_1\exp(-C_2 n^{1+\xi} + 2n^{\xi}) \to 1 \tag{S23}$$

for some positive constants C_1 and C_2. We write

$$\begin{aligned} P\left(\mathcal{C}_{[j]} \subseteq \widehat{\mathcal{C}}_{[j]}\right) &= P\left(\bigcap\nolimits_{j^*\in\mathcal{C}_{[j]}}\{j^* \in \widehat{\mathcal{C}}_{[j]}\}\right) = 1 - P\left(\bigcup\nolimits_{j^*\in\mathcal{C}_{[j]}}\{j^* \notin \widehat{\mathcal{C}}_{[j]}\}\right) \\ &\geqslant 1 - \sum\nolimits_{j^*\in\mathcal{C}_{[j]}} P\left(j^* \notin \widehat{\mathcal{C}}_{[j]}\right). \end{aligned} \tag{S24}$$

Notice that given that $j^* \in \mathcal{C}_{[j]}$, there must exist a path consisting of a sequence of pairs $\{(j_{m-1}, j_m)\}_{m=1}^Q$ with $j_0 = j$ and $j_Q = j^*$ such that $(j_{m-1}, j_m) \in \mathcal{E}$ for $m = 1, \ldots, Q \leqslant |\mathcal{C}_{[j]}|$. If $j^* \notin \widehat{\mathcal{C}}_{[j]}$, there must exist a pair of nodes in the sequence, say (j_{m^*-1}, j_{m^*}), $1 \leqslant m^* \leqslant Q$, such that $|\mathbf{X}'_{j_{m^*-1}}\mathbf{X}_{j_{m^*}}|/n \leqslant \alpha$. Otherwise, given the training data, the path $\{(j_{m-1}, j_m)\}_{m=1}^Q$ would also connect j and j^* in $\widetilde{\boldsymbol{\Sigma}}$, which contradicts the fact that $j^* \notin \widehat{\mathcal{C}}_{[j]}$. Then by Lemma

2, with $\alpha = Cn^{(\xi-1)/2}$ for some positive constant C, for $j^* \in \mathcal{C}_{[j]}$, we have

$$\begin{aligned}
& P\left(j^* \notin \widehat{\mathcal{C}}_{[j]}\right) \\
\leqslant\ & P\left(\exists\ m^* \in \{1, \cdots, Q\}, |\mathbf{X}'_{j_{m^*-1}}\mathbf{X}_{j_{m^*}}|/n \leqslant Cn^{(\xi-1)/2} \text{ given } (j_{m^*-1}, j_{m^*}) \in \mathcal{E}\right) \\
\leqslant\ & \sum\nolimits_{m^*=1}^{|\mathcal{C}_{[j]}|} P\left(|\mathbf{X}'_{j_{m^*-1}}\mathbf{X}_{j_{m^*}}|/n \leqslant Cn^{(\xi-1)/2} \text{ given } (j_{m^*-1}, j_{m^*}) \in \mathcal{E}\right) \\
\leqslant\ & C_1 \exp(n^\xi) \exp(-C_2 n^{1+\xi})
\end{aligned}$$

for some constants $C_1 > 0$ and $C_2 > 0$. The second to last step is by (A3), which assumes that $\log |\mathcal{C}_{[j]}| = O(n^\xi)$ for all $j = 1, \ldots, p$. From (S24),

$$\begin{aligned}
& P\left(\mathcal{C}_{[j]} \subseteq \widehat{\mathcal{C}}_{[j]}\right) \geqslant 1 - \sum\nolimits_{j^* \in \mathcal{C}_{[j]}} C_1 \exp(n^\xi) \exp(-C_2 n^{1+\xi}) \\
& \geqslant 1 - C_1 \exp(2n^\xi) \exp(-C_2 n^{1+\xi}) = 1 - C_1 \exp(-C_2 n^{1+\xi} + 2n^\xi) \to 1.
\end{aligned}$$

Next we show that

$$P\left(\widehat{\mathcal{C}}_{[j]} \subseteq \mathcal{C}_{[j]}\right) \geqslant 1 - C_1 \exp(-C_2 n^{1+\xi} + 3n^\varsigma) \to 1 \tag{S25}$$

for some $\xi < \varsigma < 1$ such that $p = \exp(Cn^\varsigma)$ for some constant $C > 0$. Since $\widehat{\mathcal{C}}_{[j]} \subseteq \mathcal{C}_{[j]}$ is equivalent to $\mathcal{C}^c_{[j]} \subseteq \widehat{\mathcal{C}}^c_{[j]}$,

$$\begin{aligned}
& P\left(\widehat{\mathcal{C}}_{[j]} \subseteq \mathcal{C}_{[j]}\right) = P\left(\bigcap\nolimits_{j^* \in \mathcal{C}^c_{[j]}} \{j^* \in \widehat{\mathcal{C}}^c_{[j]}\}\right) \\
& = 1 - P\left(\bigcup\nolimits_{j^* \in \mathcal{C}^c_{[j]}} \{j^* \in \widehat{\mathcal{C}}_{[j]}\}\right) \geqslant 1 - \sum\nolimits_{j^* \in \mathcal{C}^c_{[j]}} P\left(j^* \in \widehat{\mathcal{C}}_{[j]}\right).
\end{aligned} \tag{S26}$$

Suppose that $j^* \in \mathcal{C}^c_{[j]}$. If $j^* \in \widehat{\mathcal{C}}_{[j]}$, then there must exist a pair of nodes (j_1, j_2) such that $|\mathbf{X}'_{j_1}\mathbf{X}_{j_2}|/n > Cn^{(\xi-1)/2}$ but $(j_1, j_2) \notin \mathcal{E}$. Otherwise, for any path that connects j^* and j in $\widetilde{\mathbf{\Sigma}}$, we have $|\mathbf{X}'_{j_1}\mathbf{X}_{j_2}|/n > Cn^{(\xi-1)/2}$ for any adjacent pair j_1 and j_2 in the path. If $(j_1, j_2) \in \mathcal{E}$ for all such pairs j_1 and j_2, the path will also connect j^* and j in $\mathbf{\Omega}$, which contradicts $j^* \in \mathcal{C}^c_{[j]}$. Therefore,

$$\begin{aligned}
& P\left(j^* \in \widehat{\mathcal{C}}_{[j]} \text{ given that } j^* \in \mathcal{C}^c_{[j]}\right) \leqslant P\left(\bigcup\nolimits_{(j_1,j_2)\notin\mathcal{E}} |\mathbf{X}'_{j_1}\mathbf{X}_{j_2}|/n > Cn^{(\xi-1)/2}\right) \\
& \leqslant \sum\nolimits_{(j_1,j_2)\notin\mathcal{E}} P\left(|\mathbf{X}'_{j_1}\mathbf{X}_{j_2}|/n > Cn^{(\xi-1)/2}\right) \leqslant p^2 P\left(|\mathbf{X}'_{j_1}\mathbf{X}_{j_2}|/n > Cn^{(\xi-1)/2}, (j_1, j_2) \notin \mathcal{E}\right) \\
& \leqslant \widetilde{C}_1 \exp(2n^\varsigma) \exp(-\widetilde{C}_2 n^{1+\xi}).
\end{aligned}$$

Therefore, (S26) implies that

$$\begin{aligned} & P\left(\widehat{\mathcal{C}}_{[j]} \subseteq \mathcal{C}_{[j]}\right) \geqslant 1 - p\widetilde{C}_1 \exp(2n^{\varsigma}) \exp(-\widetilde{C}_2 n^{1+\xi}) \\ & = 1 - \widetilde{C}_1 \exp(3n^{\varsigma}) \exp(-\widetilde{C}_2 n^{1+\xi}) = 1 - \widetilde{C}_1 \exp(-\widetilde{C}_2 n^{1+\xi} + 3n^{\varsigma}) \to 1. \end{aligned}$$

□

Proof of Theorem 2. By Lemma 3, we have $P(\widehat{\mathcal{S}}_1(k,k') = \mathcal{S}_1(k,k')) \to 1$ as $n \to \infty$ for any $1 \leqslant k < k' \leqslant K$. Therefore $P(\widehat{\mathcal{S}}_1 = \mathcal{S}_1) \to 1$ as $n \to \infty$.

Let $\epsilon = n^{-\widetilde{\varrho}/2}$, where $0 < \widetilde{\varrho} < \varrho$ with ϱ given in (A9). We show that if

$$P(|\mathcal{S}_0 \cap \widehat{\mathcal{S}}_0| \geqslant (1-\epsilon)|\mathcal{S}_0|) \to 1 \text{ as } n \to \infty, \tag{S27}$$

then Theorem 2 follows as $\epsilon \to 0$ when $n \to \infty$.

Let $\mathcal{S}_0(k,k') = \{1 \leqslant j \leqslant p : \sum_{j'=1}^{p} \Omega_{jj'}(\mu_{kj'} - \mu_{k'j'}) \neq 0\}$ for a given pair of classes k and k', $1 \leqslant k < k' \leqslant K$. Then $\mathcal{S}_0 = \bigcup_{1\leqslant k<k'\leqslant K} \mathcal{S}_0(k,k')$. Let $\widehat{\mathcal{S}}(k,k') = \{j \in \mathcal{U} : |\sum_{j'\in\widehat{\mathcal{C}}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})| \geqslant \nu_n\}$, and let

$$\begin{aligned} \widehat{\mathcal{S}} &= \bigcup_{1\leqslant k<k'\leqslant K} \widehat{\mathcal{S}}(k,k') \\ &= \{j \in \mathcal{U} : |\sum\nolimits_{j'\in\widehat{\mathcal{C}}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})| \geqslant \nu_n \text{ for some } 1 \leqslant k < k' \leqslant K\}. \end{aligned}$$

Since

$$\begin{aligned} \widehat{\mathcal{S}} &= \{j \in \mathcal{U} : |\sum\nolimits_{j'\in\widehat{\mathcal{C}}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})| \geqslant \nu_n \text{ for some } 1 \leqslant k < k' \leqslant K\} \\ &= \{j \in \mathcal{U} \cap \widehat{\mathcal{S}}_1 : |\sum\nolimits_{j'\in\widehat{\mathcal{C}}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})| \geqslant \nu_n \text{ for some } 1 \leqslant k < k' \leqslant K\} \cup \\ &\quad \{j \in \mathcal{U} \cap \widehat{\mathcal{S}}_1^c : |\sum\nolimits_{j'\in\widehat{\mathcal{C}}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})| \geqslant \nu_n \text{ for some } 1 \leqslant k < k' \leqslant K\} \\ &\subseteq \widehat{\mathcal{S}}_1 \cup \widehat{\mathcal{S}}_2 = \widehat{\mathcal{S}}_0, \end{aligned}$$

to show (S27), it suffices to show that

$$P(|\mathcal{S}_0 \cap \widehat{\mathcal{S}}| \geqslant (1-\epsilon)|\mathcal{S}_0|) \to 1 \text{ as } n \to \infty. \tag{S28}$$

If for any pair (k,k'), $1 \leqslant k < k' \leqslant K$, $|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0(k,k')| \geqslant (1-\epsilon)|\mathcal{S}_0(k,k')|$, then we have

$$\begin{aligned}
|\widehat{\mathcal{S}} \cap \mathcal{S}_0| &= \left|(\textstyle\bigcup_{1\leqslant k<k'\leqslant K}\widehat{\mathcal{S}}(k,k')) \cap (\bigcup_{1\leqslant k<k'\leqslant K}\mathcal{S}_0(k,k'))\right| \\
\geqslant & \left|\textstyle\bigcup_{1\leqslant k<k'\leqslant K}(\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0(k,k'))\right| \geqslant (1-\epsilon)\left|\bigcup_{1\leqslant k<k'\leqslant K}\mathcal{S}_0(k,k')\right| \\
= & (1-\epsilon)\,|\mathcal{S}_0|\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& P\left(|\widehat{\mathcal{S}} \cap \mathcal{S}_0| \geqslant (1-\epsilon)|\mathcal{S}_0|\right) \\
\geqslant\; & P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0(k,k')| \geqslant (1-\epsilon)|\mathcal{S}_0(k,k')| \text{ for all } 1 \leqslant k < k' \leqslant K\right) \\
=\; & 1 - P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0(k,k')| < (1-\epsilon)|\mathcal{S}_0(k,k')| \text{ for some } 1 \leqslant k < k' \leqslant K\right) \\
\geqslant\; & 1 - \textstyle\sum_{1\leqslant k<k'\leqslant K} P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0(k,k')| < (1-\epsilon)|\mathcal{S}_0(k,k')|\right). \qquad \text{(S29)}
\end{aligned}$$

As a result, to prove (S28), it suffices to show that for any pair k and k', when $n \to \infty$,

$$P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0(k,k')| < (1-\epsilon)|\mathcal{S}_0(k,k')|\right) \to 0.$$

For a $j \in \mathcal{U}$, denote

$$T_j(k,k') = 1\left(\left|\textstyle\sum_{j'\in\widehat{\mathcal{C}}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\right| > \nu_n\right).$$

In the following, we write $T_j = T_j(k,k')$ for short when no confusion arises. It is easy to see that

$$\begin{aligned}
& P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0(k,k')| < (1-\epsilon)|\mathcal{S}_0(k,k')|\right) \\
& \qquad = P\left(\textstyle\sum_{j\in\mathcal{S}_0(k,k')} T_j < (1-\epsilon)|\mathcal{S}_0(k,k')|\right). \qquad \text{(S30)}
\end{aligned}$$

To evaluate the probability on the right hand side of (S30), we first evaluate $E(T_j)$ for $j \in \mathcal{S}_0(k,k')$. Notice that

$$\begin{aligned}
E(T_j) &= P\left(\left|\textstyle\sum_{j'\in\widehat{\mathcal{C}}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\right| > \nu_n\right) \\
&\geqslant P\left(\left|\textstyle\sum_{j'\in\mathcal{C}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\right| > \nu_n\right) P\left(\widehat{\mathcal{C}}_{[j]} = \mathcal{C}_{[j]}\right). \qquad \text{(S31)}
\end{aligned}$$

If for all $j' \in \mathcal{C}_{[j]} \cap \mathcal{S}_1(k,k')$, $|\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'}| > \tau$, then for a $\nu_n < (\kappa_2^{-2} - O_p(\rho_n^2))^{1/2}\tau$,

$$\begin{aligned}
&\Big|\sum\nolimits_{j' \in \mathcal{C}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\Big| \\
&= \big\{(\overline{\mathbf{X}}^{(k)}_{\cdot\mathcal{C}_{[j]}} - \overline{\mathbf{X}}^{(k')}_{\cdot\mathcal{C}_{[j]}})'(\widehat{\mathbf{\Omega}}_{j\mathcal{C}_{[j]}})'\widehat{\mathbf{\Omega}}_{j\mathcal{C}_{[j]}}(\overline{\mathbf{X}}^{(k)}_{\cdot\mathcal{C}_{[j]}} - \overline{\mathbf{X}}^{(k')}_{\cdot\mathcal{C}_{[j]}})\big\}^{1/2} \\
&\geqslant (\kappa_2^{-2} - O_p(\rho_n^2))^{1/2}\big\{(\overline{\mathbf{X}}^{(k)}_{\cdot\mathcal{C}_{[j]}} - \overline{\mathbf{X}}^{(k')}_{\cdot\mathcal{C}_{[j]}})'(\overline{\mathbf{X}}^{(k)}_{\cdot\mathcal{C}_{[j]}} - \overline{\mathbf{X}}^{(k')}_{\cdot\mathcal{C}_{[j]}})\big\}^{1/2} \\
&\geqslant (\kappa_2^{-2} - O_p(\rho_n^2))^{1/2}\tau\,\big|\mathcal{C}_{[j]}\big| > \nu_n,
\end{aligned}$$

where κ_2 is from (A6), $\widehat{\mathbf{\Omega}}_{j\mathcal{C}_{[j]}}$ is a subvector of the jth row of $\widehat{\mathbf{\Omega}}$ with column entries indexed by $\mathcal{C}_{[j]}$ and $\overline{\mathbf{X}}^{(k)}_{\cdot\mathcal{C}_{[j]}}$ is the subvector of the mean vector of class k, $\overline{\mathbf{X}}^{(k)}$, with entries indexed by $\mathcal{C}_{[j]}$. The second last step comes from the fact that $\|(\widehat{\mathbf{\Omega}}_{j\mathcal{C}_{[j]}})'\widehat{\mathbf{\Omega}}_{j\mathcal{C}_{[j]}}\| \geqslant \|(\mathbf{\Omega}^u_{j\mathcal{C}_{[j]}})'\mathbf{\Omega}^u_{j\mathcal{C}_{[j]}}\| - O_p(\rho_n^2) \geqslant \lambda^2_{\min}(\mathbf{\Omega}) - O_p(\rho_n^2)$ by (A6) and (S5). Therefore,

$$\begin{aligned}
&P\Big(\Big|\sum\nolimits_{j' \in \mathcal{C}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\Big| > \nu_n\Big) \\
&> \; P\left(|\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'}| > \tau \text{ for all } j' \in \mathcal{C}_{[j]} \cap \mathcal{S}_1(k,k')\right) \\
&= \; 1 - P\left(|\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'}| \leqslant \tau \text{ for some } j' \in \mathcal{C}_{[j]} \cap \mathcal{S}_1(k,k')\right) \\
&> \; 1 - C_1\exp(-C_2 n^{1-\varsigma/2} r \log p)|\mathcal{S}_0(k,k')| > 1 - o(p^{-1}). \qquad \text{(S32)}
\end{aligned}$$

The second last step is from the first statement in Lemma 3. From (S23) and (S25), we also have that

$$P\big(\widehat{\mathcal{C}}_{[j]} = \mathcal{C}_{[j]}\big) \geqslant 1 - C_1\exp(-C_2 n^{1+\xi} + \max(2n^{\xi}, 3n^{\varsigma})) = 1 - o(p^{-1}). \qquad \text{(S33)}$$

Plug (S32) and (S33) into (S31), we have that for sufficiently large n, $E(T_j) > 1 - o(p^{-1})$.

Next, we evaluate $P\big(\sum_{j \in \mathcal{S}_0(k,k')} T_j < (1-\epsilon)|\mathcal{S}_0(k,k')|\big)$. For notational simplicity, suppose that the features in $\mathcal{S}_0(k,k')$ are indexed as $1, \ldots, |\mathcal{S}_0(k,k')|$. For the pair k and k', denote $V_j = V_j(k,k') = -(T_j - E[T_j])/\sqrt{|\mathcal{S}_0(k,k')|}$, $1 \leqslant j \leqslant |\mathcal{S}_0(k,k')|$. Then, $|V_j| \leqslant 1/\sqrt{|\mathcal{S}_0(k,k')|}$. Let $\mathcal{F}_j$ be the sigma field generated by random variables $\{V_1, \ldots, V_j\}$, $1 \leqslant j \leqslant |\mathcal{S}_0(k,k')|$. Then, $\emptyset = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_{|\mathcal{S}_0(k,k')|}$ and $E(V_j|\mathcal{F}_{j-1}) = 0$. Let $K_j = \varepsilon_0/\sqrt{|\mathcal{S}_0(k,k')|}$ for some constant $\varepsilon_0 > 1$. Then $K_j(>0)$ is a constant and is $\mathcal{F}_{j-1}$ measurable. Moreover,

$\widetilde{\Psi}(|V_j|/K_j) \leqslant e-1$. Thus,

$$\begin{aligned} B^2_{|\mathcal{S}_0(k,k')|} &= \sum\nolimits_{j=1}^{|\mathcal{S}_0(k,k')|} K_j^2 \left(1 + E\left(\widetilde{\Psi}(|V_j|/K_j)\right)|\mathcal{F}_{i-1}\right) \\ &\leqslant \frac{\varepsilon_0^2}{|\mathcal{S}_0(k,k')|} \sum\nolimits_{j=1}^{|\mathcal{S}_0(k,k')|} (1+e-1) \leqslant e\varepsilon_0^2. \end{aligned}$$

When $n \to \infty$ and $p \to \infty$,

$$\begin{aligned} & P\Big(\sum\nolimits_{j\in\mathcal{S}_0(k,k')} T_j < (1-\epsilon)|\mathcal{S}_0(k,k')|\Big) \qquad \text{(S34)} \\ = & P\Big(\sum\nolimits_{j\in\mathcal{S}_0(k,k')} -(T_j - E(T_j)) > (E(T_j)+\epsilon-1)|\mathcal{S}_0(k,k')|\Big) \\ = & P\left(\frac{1}{\sqrt{|\mathcal{S}_0(k,k')|}} \sum\nolimits_{j\in\mathcal{S}_0(k,k')} -(T_j - E(T_j)) > \sqrt{|\mathcal{S}_0(k,k')|}(\epsilon + E(T_j) - 1)\right) \\ = & P\left(\sum\nolimits_{j\in\mathcal{S}_0(k,k')} V_j > Cn^{\varrho/2}(n^{-\widetilde{\varrho}/2} - o(p^{-1})) \text{ and } B^2_{\mathcal{S}_0(k,k')} \leqslant e\varepsilon_0^2\right) \\ \leqslant & \exp\left\{\frac{-Cn^{\varrho}\left(n^{-\widetilde{\varrho}/2} - o(p^{-1})\right)^2}{8e\varepsilon_0^2}\right\} \to 0 \end{aligned}$$

for some $C > 0$. The second to last step is from the fact that $|\mathcal{S}_0(k,k')| = O(n^{\varrho})$, $\epsilon = n^{-\widetilde{\varrho}/2}$, $1 - o(p^{-1}) < E(T_j) \leqslant 1$, and $B^2_{\mathcal{S}_0(k,k')} \leqslant e\varepsilon_0^2$. The last step stems from Lemma 4 and $0 < \widetilde{\varrho} < \varrho$. Notice that T_j, and therefore V_j ($j \in \mathcal{S}_0$), are dependent random variables, so we need to apply Lemma 4 instead of the conventional Hoeffding's inequality for independent random variables. This gives that for any pair k and k', $1 \leqslant k < k' \leqslant K$, when $n \to \infty$, $P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0(k,k')| < (1-\epsilon)|\mathcal{S}_0(k,k')|\right) \to 0$. By (S29), the proof is completed. □

Proof of Theorem 3. It is clear from the definition of $\widehat{\mathcal{S}}$ that we have $\widehat{\mathcal{S}}_2 \subset \widehat{\mathcal{S}}$. Therefore, $\widehat{\mathcal{S}} \subseteq \widehat{\mathcal{S}}_0 \subseteq \widehat{\mathcal{S}} \cup \widehat{\mathcal{S}}_1$. As a result,

$$P\left(|\widehat{\mathcal{S}}_0 \cap \mathcal{S}_0^c| \leqslant \zeta^{-1}|\mathcal{S}_0^c|\right) \leqslant P\left(|\widehat{\mathcal{S}} \cap \mathcal{S}_0^c| \leqslant \zeta^{-1}|\mathcal{S}_0^c|\right) \qquad \text{(S35)}$$

and

$$\begin{aligned}
&P\left(|\widehat{\mathcal{S}}_0 \cap \mathcal{S}_0^c| \leqslant \zeta^{-1}|\mathcal{S}_0^c|\right) \geqslant P\left(|(\widehat{\mathcal{S}} \cup \widehat{\mathcal{S}}_1) \cap \mathcal{S}_0^c| \leqslant \zeta^{-1}|\mathcal{S}_0^c|\right)\\
&= 1 - P\left(|(\widehat{\mathcal{S}} \cap \mathcal{S}_0^c) \cup (\widehat{\mathcal{S}}_1 \cap \mathcal{S}_0^c)| > \zeta^{-1}|\mathcal{S}_0^c|\right)\\
&\geqslant 1 - P\left(|\widehat{\mathcal{S}} \cap \mathcal{S}_0^c| > \frac{1}{2}\zeta^{-1}|\mathcal{S}_0^c|\right) - P\left(|\widehat{\mathcal{S}}_1 \cap \mathcal{S}_0^c| > \frac{1}{2}\zeta^{-1}|\mathcal{S}_0^c|\right)\\
&\geqslant P\left(|\widehat{\mathcal{S}} \cap \mathcal{S}_0^c| \leqslant \frac{1}{2}\zeta^{-1}|\mathcal{S}_0^c|\right) - P\left(|\widehat{\mathcal{S}}_1 \cap \mathcal{S}_0^c| > \frac{1}{2}\zeta^{-1}|\mathcal{S}_0^c|\right). \qquad \text{(S36)}
\end{aligned}$$

From Lemma 3, $P(\widehat{\mathcal{S}}_1 = \mathcal{S}_1) \to 1$ as $n \to \infty$. Therefore, $P\left(|\widehat{\mathcal{S}}_1 \cap \mathcal{S}_0^c| > \frac{1}{2}\zeta^{-1}|\mathcal{S}_0^c|\right) \to 0$ as $n \to \infty$. So from (S35) and (S36), in order to show that $P\left(|\widehat{\mathcal{S}}_0 \cap \mathcal{S}_0^c| \leqslant \zeta^{-1}|\mathcal{S}_0^c|\right) \to 1$ as $n \to \infty$, it suffices to show that

$$P\left(|\widehat{\mathcal{S}} \cap \mathcal{S}_0^c| \leqslant \zeta^{-1}|\mathcal{S}_0^c|\right) \to 1 \text{ as } n \to \infty$$

for any $\zeta = o(n \log p)$.

If for any pair (k, k'), $1 \leqslant k < k' \leqslant K$, $|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0^c| \leqslant \zeta^{-1}|\mathcal{S}_0^c|$, then we have

$$\begin{aligned}
&|\widehat{\mathcal{S}} \cap \mathcal{S}_0^c| = \left|(\textstyle\bigcup_{1\leqslant k<k'\leqslant K}\widehat{\mathcal{S}}(k,k')) \cap \mathcal{S}_0^c\right| = \left|\textstyle\bigcup_{1\leqslant k<k'\leqslant K}(\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0^c)\right|\\
\leqslant\ & \textstyle\sum_{1\leqslant k<k'\leqslant K} |\mathcal{S}_0(k,k') \cap \mathcal{S}_0^c| \leqslant K^2\zeta^{-1}\,|\mathcal{S}_0^c|\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& P\left(|\widehat{\mathcal{S}} \cap \mathcal{S}_0^c| \leqslant K^2\zeta^{-1}|\mathcal{S}_0^c|\right)\\
\geqslant\ & P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0^c| \leqslant \zeta^{-1}|\mathcal{S}_0^c| \text{ for all } 1 \leqslant k < k' \leqslant K\right)\\
=\ & 1 - P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0^c| > \zeta^{-1}|\mathcal{S}_0^c| \text{ for some } 1 \leqslant k < k' \leqslant K\right)\\
\geqslant\ & 1 - \textstyle\sum_{1\leqslant k<k'\leqslant K} P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0^c| > \zeta^{-1}|\mathcal{S}_0^c|\right). \qquad \text{(S37)}
\end{aligned}$$

Therefore, to show Theorem 3, it suffices to show that for any pair k and k', $1 \leqslant k < k' < K$,

$$P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0^c| > \zeta^{-1}|\mathcal{S}_0^c|\right) \to 0 \text{ as } n \to \infty. \qquad \text{(S38)}$$

For a given pair k and k', $1 \leqslant k < k' < K$,

$$\begin{aligned} & P\left(|\widehat{\mathcal{S}}(k,k') \cap \mathcal{S}_0^c| > \zeta^{-1}|\mathcal{S}_0^c|\right) \\ = \;& P\left(\frac{1}{|\mathcal{S}_0^c|}\sum\nolimits_{j\in\mathcal{S}_0^c} 1\big(\big|\sum\nolimits_{j'\in\widehat{\mathcal{C}}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\big| > \nu_n\big) > \zeta^{-1}\right) \\ = \;& P\left(\frac{1}{|\mathcal{S}_0^c|}\sum\nolimits_{j\in\mathcal{S}_0^c} T_j > \zeta^{-1}\right). \end{aligned}$$

First we evaluate $E(T_j)$ for $j \in \mathcal{S}_0^c$.

$$\begin{aligned} E(T_j) \;&=\; P\left(\big|\sum\nolimits_{j'\in\widehat{\mathcal{C}}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\big| > \nu_n\right) \\ &\leqslant\; P\left(\big|\sum\nolimits_{j'\in\mathcal{C}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\big| > \nu_n\right) P\left(\widehat{\mathcal{C}}_{[j]} = \mathcal{C}_{[j]}\right) + P\left(\widehat{\mathcal{C}}_{[j]} \neq \mathcal{C}_{[j]}\right) \quad \text{(S39)} \end{aligned}$$

By (A11), for sufficiently large n, $\rho_n \leqslant C$ for some positive constant C. When choosing $\nu_n = \sqrt{r(\log p)\exp(n^\xi)}$,

$$\begin{aligned} & P\left(\big|\sum\nolimits_{j'\in\mathcal{C}_{[j]}} \widehat{\Omega}^u_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\big| > \nu_n\right) \\ & \leqslant P\left((1+O_P(\rho_n))\big|\sum\nolimits_{j'\in\mathcal{C}_{[j]}} \Omega_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\big| > \nu_n\right) \\ & = P\Big((1+O_P(\rho_n))\big|\sum\nolimits_{j'\in\mathcal{C}_{[j]}} \Omega_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'}) - \\ & \qquad\qquad \sum\nolimits_{j'\in\mathcal{C}_{[j]}} \Omega_{jj'}(\mu_{kj'} - \mu_{k'j'})\big| > \nu_n\Big) \qquad \text{(S40)} \\ & \leqslant \frac{(1+C)^2 \mathrm{var}\left(\sum\nolimits_{j'\in\mathcal{C}_{[j]}} \Omega_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})\right)}{\nu_n^2} \qquad \text{(S41)} \\ & \leqslant \frac{(1+C)^2\kappa_1^{-1}\kappa_2|\mathcal{C}_{[j]}|\,(1/n_k + 1/n_{k'})}{r(\log p)\exp(n^\xi)} \\ & = O((n\log p)^{-1}), \qquad \text{(S42)} \end{aligned}$$

where (S40) is from the fact that $E(\sum_{j'\in\mathcal{C}_{[j]}} \Omega_{jj'}(\overline{X}^{(k)}_{\cdot j'} - \overline{X}^{(k')}_{\cdot j'})) = \sum_{j'\in\mathcal{C}_{[j]}} \Omega_{jj'}(\mu_{kj'} - \mu_{k'j'}) = 0$ given that $j \in \mathcal{S}_0^c$, (S41) is from Markov's inequality and (S42) is from (A3) which states that, $|\mathcal{C}_{[j]}| \leqslant C_{\max} = O(\exp(n^\xi))$. Then from (S39),

$$E(T_j) = O((n\log p)^{-1}) + o(p^{-1}) = O((n\log p)^{-1}).$$

Therefore, for any $\zeta = o(n\log p)$, $\zeta^{-1} > E(T_j)$.

For the pair k and k', denote $V_j = V_j(k,k') = (T_j - E[T_j])/\sqrt{|\mathcal{S}_0^c|}$, $1 \leqslant j \leqslant |\mathcal{S}_0^c|$. Then

$|V_j| \leqslant 1/\sqrt{|\mathcal{S}_0^c|}$. Let $K_j = \varepsilon_0/\sqrt{|\mathcal{S}_0^c|}$ for some constant $\varepsilon_0 > 1$. Then $K_j(>0)$ is a constant and $\mathcal{F}_{j-1}$ measurable. Clearly $\widetilde{\Psi}(|V_j|/K_j) \leqslant e-1$. Thus, the following inequality holds.

$$\begin{aligned} B^2_{|\mathcal{S}_0^c|} &= \sum\nolimits_{j\in\mathcal{S}_0^c} K_j^2\left(1 + E\left(\widetilde{\Psi}(|V_j|/K_j)\right)|\mathcal{F}_{i-1}\right) \\ &\leqslant \frac{\varepsilon_0^2}{|\mathcal{S}_0^c|}\sum\nolimits_{j\in\mathcal{S}_0^c}(1+e-1) \leqslant e\varepsilon_0^2. \end{aligned}$$

As a result, when $n\to\infty$ and $p\to\infty$,

$$\begin{aligned} &P\left(\frac{1}{|S_0^c|}\sum\nolimits_{j\in\mathcal{S}_0^c} T_j > \zeta^{-1}\right) \\ = \;& P\left(\frac{1}{\sqrt{|S_0^c|}}\sum\nolimits_{j\in\mathcal{S}_0^c}(T_j - E(T_j)) > \sqrt{|S_0^c|}(\zeta^{-1} - E(T_j))\right) \\ = \;& P\left(\sum\nolimits_{j\in\mathcal{S}_0^c} V_j > Cp^{1/2}(1-p^{-\beta}-p^{-\gamma})^{1/2}(\zeta^{-1}-E(T_j)) \text{ and } B^2_{|S_0^c|} \leqslant e\varepsilon_0^2\right) \\ \leqslant \;& \exp\left\{\frac{-C^2p(1-p^{-\beta}-p^{-\gamma})(\zeta^{-1}-O((n\log p)^{-1}))^2}{8e\varepsilon_0^2}\right\} \to 0 \end{aligned}$$

for some constant $C>0$. We obtain $|S_0^c| = p(1-p^{-\beta}-p^{-\gamma})$ from (A10). The last step comes from the fact that when $\zeta = o(n\log p)$, $\zeta^{-1} - O((n\log p)^{-1}) > 0$. This arrives at (S38) and completes the proof. □

Proof of Theorem 4. Given the training data $\mathcal{D}$,

$$\begin{aligned} &R_{\text{mLDA}}(\mathcal{D}) \qquad\qquad \text{(S43)} \\ &= \frac{1}{K}\sum\nolimits_{k=1}^K P\Big(\underset{1\leqslant l\leqslant K}{\arg\max}\{(\mathbf{X}^{\rm s}_{\text{new}} - \widehat{\boldsymbol{\mu}}^{\rm s}_l/2)'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\mu}}^{\rm s}_l\} = k',\ k'\neq k|Y_{\text{new}} = k;\mathcal{D}\Big) \\ &\leqslant \frac{1}{K}\sum\nolimits_{k=1}^K\sum\nolimits_{k'\neq k} P\left((\mathbf{X}^{\rm s}_{\text{new}} - \widehat{\boldsymbol{\mu}}^{\rm s}_{k'}/2)'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\mu}}^{\rm s}_{k'} > (\mathbf{X}^{\rm s}_{\text{new}} - \widehat{\boldsymbol{\mu}}^{\rm s}_k/2)'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\mu}}^{\rm s}_k|Y_{\text{new}} = k;\mathcal{D}\right). \end{aligned}$$

Notice that for $k'\neq k$, $(\mathbf{X}^{\rm s}_{\text{new}} - \widehat{\boldsymbol{\mu}}^{\rm s}_{k'}/2)'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\mu}}^{\rm s}_{k'} > (\mathbf{X}^{\rm s}_{\text{new}} - \widehat{\boldsymbol{\mu}}^{\rm s}_k/2)'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\mu}}^{\rm s}_k$ if and only if $(\mathbf{X}^{\rm s}_{\text{new}} - \widehat{\boldsymbol{\mu}}^{\rm s}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'} < 0$ with $\widehat{\boldsymbol{\mu}}^{\rm s}_{kk'} = (\widehat{\boldsymbol{\mu}}^{\rm s}_k + \widehat{\boldsymbol{\mu}}^{\rm s}_{k'})/2$ and $\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'} = \widehat{\boldsymbol{\mu}}^{\rm s}_{k'} - \widehat{\boldsymbol{\mu}}^{\rm s}_k$. Direct calculation gives that $E[(\mathbf{X}^{\rm s}_{\text{new}} - \widehat{\boldsymbol{\mu}}^{\rm s}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'}|Y_{\text{new}} = k;\mathcal{D}] = (\boldsymbol{\mu}^{\rm s}_k - \widehat{\boldsymbol{\mu}}^{\rm s}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'}$, $\text{var}[(\mathbf{X}^{\rm s}_{\text{new}} - \widehat{\boldsymbol{\mu}}^{\rm s}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'}|Y_{\text{new}} = k;\mathcal{D}] = (\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\rm s}\boldsymbol{\Sigma}^{\rm s}\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'}$, and

$$P\left((\mathbf{X}^{\rm s}_{\text{new}} - \widehat{\boldsymbol{\mu}}^{\rm s}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'} < 0\,|Y_{\text{new}} = k;\mathcal{D}\right) = \Phi\left(-\frac{(\boldsymbol{\mu}^{\rm s}_k - \widehat{\boldsymbol{\mu}}^{\rm s}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'}}{\sqrt{(\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\rm s}\boldsymbol{\Sigma}^{\rm s}\widehat{\boldsymbol{\Omega}}^{\rm s}\widehat{\boldsymbol{\delta}}^{\rm s}_{kk'}}}\right). \qquad \text{(S44)}$$

First consider the denominator in (S44). By (S5),

$$(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\boldsymbol{\Sigma}^{\mathrm{s}}\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'} = (\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}(1+O_P(\rho_n)) = (\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}(1+O_P(\rho_n)).$$

Then by Lemma 5, when conditioning on the event $\{\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0\}$,

$$\begin{aligned}&(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\boldsymbol{\Sigma}^{\mathrm{s}}\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}\\ &= \Bigg\{(\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}\boldsymbol{\delta}^{0}_{kk'} + O_P\left(\sqrt{\Delta_p^2(k,k')\frac{|\mathcal{S}_0|(n_k+n_{k'})}{n_kn_{k'}}}\right) + O_P\left(\frac{|\mathcal{S}_0|(n_k+n_{k'})}{n_kn_{k'}}\right) + \\ &\quad O_P(1)\Bigg\}(1+O_P(\rho_n)). \qquad \text{(S45)}\end{aligned}$$

Next we look at the numerator in (S44). By Lemma 6, when conditioning on the event $\{\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0\}$,

$$\begin{aligned}&(\boldsymbol{\mu}^{\mathrm{s}}_{k} - \widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'} = (\boldsymbol{\mu}^{\mathrm{s}}_{k} - \widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{kk'})'\boldsymbol{\Omega}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}(1+O_P(\rho_n))\\ &= \left\{\frac{1}{2}(\boldsymbol{\delta}^{0}_{kk'})'\boldsymbol{\Omega}^{0}\boldsymbol{\delta}^{0}_{kk'} + O_P\left(\sqrt{\Delta_p^2(k,k')|\mathcal{S}_0|/n_{k'}}\right) + O_P\left(\frac{|\mathcal{S}_0|(n_k+n_{k'})}{n_kn_{k'}}\right) + \right.\\ &\left. +O_P(1)\right\}(1+O_P(\rho_n)). \qquad \text{(S46)}\end{aligned}$$

Combining (S45) and (S46), we get

$$\frac{(\boldsymbol{\mu}^{\mathrm{s}}_{k} - \widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}}{\sqrt{(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\boldsymbol{\Sigma}^{\mathrm{s}}\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}}} = \frac{A_n}{B_n}(1+O_P(\rho_n))^{1/2},$$

where

$$A_n = \frac{1}{2}\Delta_p(k,k') + O_P(\sqrt{|\mathcal{S}_0|/n_{k'}}) + O_P\left(\frac{|\mathcal{S}_0|(n_k+n_{k'})}{n_kn_{k'}\Delta_p(k,k')}\right) + O_P\left(\frac{1}{\Delta_p(k,k')}\right)$$

and

$$B_n = \left\{1 + O_P\left(\sqrt{\frac{|\mathcal{S}_0|(n_k+n_{k'})}{n_kn_{k'}\Delta_p^2(k,k')}}\right) + O_P\left(\frac{|\mathcal{S}_0|(n_k+n_{k'})}{n_kn_{k'}\Delta_p^2(k,k')}\right) + O_P\left(\frac{1}{\Delta_p^2(k,k')}\right)\right\}^{1/2}.$$

Direct calculation gives that

$$\frac{(\boldsymbol{\mu}^{\mathrm{s}}_{k} - \widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}}{\sqrt{(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\boldsymbol{\Sigma}^{\mathrm{s}}\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}}} = \frac{\Delta_p(k,k')}{2}(1+O_P(a_n(k,k')))^{1/2}(1+O_P(\rho_n))^{1/2} \qquad \text{(S47)}$$

with $a_n(k,k') = \max\{|\mathcal{S}_0|^{1/2}/(n^{1/2}\Delta_p(k,k')), |\mathcal{S}_0|/(n\Delta_p^2(k,k')), 1/\Delta_p^2(k,k')\}$. Since $\Delta_p(k,k') \geqslant \Delta_p$ and $a_n(k,k') \geqslant a_n$ for any $1 \leqslant k < k' \leqslant K$, we have that when conditioning on the event

$\{\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0\}$,

$$\frac{(\boldsymbol{\mu}^{\mathrm{s}}_k - \widehat{\boldsymbol{\mu}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}}{\sqrt{(\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'})'\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\boldsymbol{\Sigma}^{\mathrm{s}}\widehat{\boldsymbol{\Omega}}^{\mathrm{s}}\widehat{\boldsymbol{\delta}}^{\mathrm{s}}_{kk'}}} \geqslant \frac{\Delta_p}{2}(1+O_P(a_n))^{1/2}(1+O_P(\rho_n))^{1/2}. \tag{S48}$$

Equation (S48) together with (S43), (S44) and (S47) gives that

$$R_{\text{mLDA}}(\mathcal{D}) \leqslant K\Phi\left(-(1+O_P(a_n))^{1/2}(1+O_P(\rho_n))^{1/2}\Delta_p/2\right) + o_P(1), \tag{S49}$$

where the term $o_P(1)$ stems from the fact that $P\left(\mathcal{S}_0 \not\subseteq \widehat{\mathcal{S}}_0\right)$ is of order $o_P(1)$.

Furthermore, when $\Delta_p^2 \min\{n/|\mathcal{S}_0|, 1\} \to \infty$, for any $1 \leqslant k < k' \leqslant K$, we have that

$$\frac{|\mathcal{S}_0|^{1/2}}{n^{1/2}\Delta_p(k,k')} \leqslant \frac{|\mathcal{S}_0|^{1/2}}{n^{1/2}\Delta_p} \to 0 \quad \text{and} \quad \frac{1}{\Delta_p^2(k,k')} \leqslant \frac{1}{\Delta_p^2} \to 0 \quad \text{as} \quad n \to \infty.$$

Therefore

$$a_n = \min_{1\leqslant k<k'\leqslant K} \max\left\{\frac{|\mathcal{S}_0|^{1/2}}{n^{1/2}\Delta_p(k,k')}, \frac{|\mathcal{S}_0|}{n\Delta_p^2(k,k')}, \frac{1}{\Delta_p^2(k,k')}\right\} \to 0 \text{ as } n \to \infty.$$

Also by (A11), $\rho_n \to 0$ as $n \to \infty$. From (S49), we have that $R_{\text{mLDA}}(\mathcal{D}) \to 0$ as $n \to \infty$. □

S3. Additional numerical results

Figure S1 shows the correlation graphs corresponding to the correlation matrices in Eq (5) of the main manuscript.

[Figure S1 about here.]

S3.1 *Computational speed comparisons*

We conducted the following simulations to compare the computational speed of mLDA with that of the regularized classification methods, such as the regularized optimal affine discriminant (ROAD) (Fan et al., 2012), the linear programming discriminant (LPD) (Cai and Liu, 2011) and the covariance-enhanced discriminant analysis (CED) (Xu et al., 2014).

Set $K = 2$ with equal sizes $n_1 = n_2 = 100$ and $p =$200, 1,000, 10,000 and 50,000. Variables 1–20 were generated from a multivariate normal distribution with the mean structure as in Table S1. For example, Variables 5–10, 15–20 were MI features, whereas X_1–X_4 and

X_{11}–X_{14} were considered as JI features. These 20 features were divided into four blocks: X_1–X_5, X_6–X_{10}, X_{11}–X_{15}, X_{16}–X_{20}, with features from different blocks being independent of each other. Features within the same block were governed by compound symmetry (CS) covariance structure with the correlation coefficient, $\rho = 0.7$. The remaining features were independently and identically distributed from $N(0, 1)$ and were independent of the first 20 variables. Simulations were run on a 64 bit CPU Windows PC with 16GB of memory. The average computational time over 50 replications for each method under each setting was reported in Table S2.

[Table S1 about here.]

Table S2 showed that mLDA was much faster compared to the regularization-based classification methods, especially in the ultrahigh-dimensional settings.

[Table S2 about here.]

S3.2 *Tuning parameter selection*

Table S3 reported the mean prediction errors from 100 replications under Model I for different values of thresholding parameters τ and α, while fixing the other tuning parameters. The results suggested that mLDA was not extremely sensitive to different choices of α and τ in terms of prediction.

[Table S3 about here.]

S3.3 *Performance under heterogeneous covariance structures so that the classes are not linearly separable*

The simulation results for Model III are given in Table S4, which showed that mLDA performed reasonably well even under heterogeneous covariance structures, as long as the common covariance assumption was not severely violated.

[Table S4 about here.]

We also compared the classification performance of mLDA with nonlinear classification methods, including the mixed discriminant analysis (Hastie and Tibshirani, 1995; Hastie et al., 1994), the quadratic discriminant analysis (Ripley, 1996), the regularized discriminant analysis (Hastie et al., 1995), the shrunken-centroids regularized discriminant analysis (Guo et al., 2005), neural network (Ripley, 1996), kernel support vector machine (Hsu and Lin, 2002), k-nearest neighbors (Torgo, 2010) and naive Bayes (Efron, 2013).

We used the *R* package *mda* for the mixed discriminant analysis, function *qda* in package *MASS* for the quadratic discriminant analysis, package *klaR* for the regularized discriminant analysis, package *rda* for the shrunken-centroids regularized discriminant analysis, package *caret* for neural network and k-nearest neighbors, package *kernlab* for kernel support vector machine and package *naivebayes* for naivebayes. Most of these packages automatically selected the tuning parameters. Tuning parameters in *klaR* and *rda* were selected using cross-validations and $k = 5$ is used in the k-nearest neighbors.

The comparison results were given in Table S5. As many of the competing methods did not do feature selection, we investigated classification performance only. Table S5 showed that mLDA outperformed the other nonlinear classification methods in term of classification accuracy when the classes were nearly linear separable. As expected, the classification performance of mLDA deteriorated as the classes became more linearly inseparable.

[Table S5 about here.]

S3.4 *Performance with different correlation strengths*

Table S6 showed the performance of mLDA with different correlation strengths. The results suggested that if a JI feature were highly correlated with an MI feature, it was easier to be detected and hence there would be fewer false negatives.

[Table S6 about here.]

S3.5 *Comparisons with pairwiseLDA when the number of classes K is large*

This section compared the proposed mLDA with the pairwiseLDA introduced by Pan et al. (2016) when the number of classes K is large. We slightly modified the simulation setting in Example 5 (Unbalanced Case) in Pan et al. (2016), by adding jointly informative (JI) features into the true informative feature set $\mathcal{S}_0$. Specifically, we considered p =10,000, $(n, K) = (100, 10)$, $(400, 20)$ and $(1600, 40)$. Similar to Pan et al. (2016), we fixed $\pi_1 = 1/5$ and $\pi_k = 4/\{5(K-1)\}$ for $k = 2, \dots, K$, where π_k was the prior probability of class k. The means of the informative (MI and JI) features were specified in Table S7.

[Table S7 about here.]

All of the 10,000 features were generated from a multivariate normal distribution for each class. These 10,000 features were divided into 1,000 independent and equal-sized blocks. Features within the same block had a compound symmetry covariance matrix with an equal variance 1 and an equal correlation 0.7. The kth block, where $k = 1, \dots, K$, contains an MI feature, X_{10k}, and nine JI features $X_{10(k-1)+1}$–X_{10k-1} which were correlated with X_{10k}.

For each (n, K), we independently generated a training dataset for model building and a testing dataset for evaluation of classification. When K was large, a cross-validation procedure might yield a validation part with very small-sized classes and cause numerical instability. To resolve the issue, for the same (n, K) we generated an additional independent validation dataset for selecting the tuning parameters (Brownlee, 2017) in the procedures of mLDA and pairwiseLDA (Pan et al., 2016). We also selected tuning parameters for pairwiseLDA using the extended Bayes information criteria (EBIC) as in Pan et al. (2016). In this case, the pairwiseLDA was labeled as pairwiseLDA.ebic. The experiment was replicated 100 times. The results were given in Table S8, which reported the numbers of false positives (FP), false negatives (FN) and classification errors (ER).

[Table S8 about here.]

Table S8 showed that when the number of classes K was large, our proposed mLDA was still competitive. In the scenarios we have examined, the number of false negatives did increase as K increased, but it was smaller than that obtained by the pairwiseLDA methods. This is because pairwiseLDA employed an independence rule and was not able to select JI features. Moreover, mLDA gave smaller classification errors than pairwiseLDA.

S3.6 *Comparisons with pMOM-logistic*

This section compared the performance of the proposed mLDA and the Bayesian method with a non-local prior (pMOM) introduced by Johnson and Rossell (2012); Johnson (2013) and Nikooienejad et al. (2016).

The simulation set-up was similar to Model I in the main text, except that we set p =1,000, $K = 2$ and $n_1 = n_2$ =100. The first 20 features were set to be informative and were generated from a multivariate normal distribution. These 20 features were divided into four independent and equal-sized blocks. The five features within each block had a compound symmetry covariance matrix with an equal variance 1 and an equal correlation 0.7. The means of the informative features for each class were specified as in Table S1. By design, X_5–X_{10}, X_{15}–X_{20} were the MI features and X_1–X_4 and X_{11}–X_{14} were the JI features. The remaining 980 non-informative features were independently generated from $N(0, 1)$ and were independent of the 20 informative features.

We generated a training dataset and a testing dataset independently according to the same setting described above. We applied mLDA and pMOM with logistic regressions (pMOM-logistic) (Nikooienejad et al., 2016) on the training set to select variables and establish classification rules. When implementing pMOM-logistic, we used the `R` package `BVSNLP` by specifying a non-local prior density on the (non-zero) regression coefficient with the scale parameter $\tau = 3$ and the shape parameter $r = 1$; see Johnson (2013) and Nikooienejad et al. (2016). For mLDA, the optimal tuning parameters were selected by 5-fold cross-validation on the training set and the classification performance was evaluated on the testing set. For pMOM-logistic, we fitted the logistic regression on the training set, based on which we calculated the estimated probability to fall into Class 1 for each sample in the testing set. We used 0.5 probability as the cut-off when assessing the classification performance. The experiment was replicated 100 times.

The simulation results were reported in Table S9. In summary, while the mLDA gave slightly larger false positives, it incurred much fewer false negatives and much smaller classification errors compared to pMOM-logistic.

[Table S9 about here.]

S3.7 *More results for the kidney transplant data analysis*

Figure S2 gave the estimated Fisher discriminant statistics, $d_i^{(k)}$, for $i = 1, 2, \ldots, 15$ and $k \in \{\text{C, TX, AR, NR}\}$. Note that the binary classification approaches inadvertently produced ties, which made the final class membership assignment difficult. When a tie did happen, we had to randomly assign membership among the ties. However, as shown in Figure S2, no ties occur for mLDA.

[Figure S2 about here.]

The heatmap in Figure S3 illustrated the correlation matrix between the 10 selected genes. The JI genes *CEACAM8*, *RNASE3*, *TCN1*, *BPI* and *CRISP3* were all highly correlated with the MI gene *TCF12*, and the JI gene *IGHV3-23* were highly correlated with the MI gene *HLA-G*.

[Figure S3 about here.]

References

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36,** 2577–2604.

Brownlee, J. (2017). What is the difference between test and validation datasets?https://machinelearningmastery.com/difference-test-validation-datasets.

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

Efron, B. (2013). Bayes' theorem in the 21st century. *Science* **340,** 1177–1178.

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

Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. R. Statist. Soc. B* **74(4),** 745–771.

Fan, J., Liao, Y., and Min, M. (2011). High-dimensional covariance matrix estimation in approxiamte factor models. *Ann. Statist.* **39,** 3320–3356.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B* **70,** 849–911.

Gaynanova, I., Booth, J. G., and Wells, M. T. (2016). Simultaneous sparse estimation of canonical vectors in the $p \gg n$ setting. *Journal of the American Statistical Association* **111,** 696–706.

Guo, Y., Hastie, T., and Tibshirani, R. (2005). Regularized discriminant analysis and its application in microarrays. *Biostatistics* **8,** 86–100.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Statist.* **23,** 73–102.

Hastie, T. and Tibshirani, R. (1995). Discriminant analysis by Gaussian mixtures. *J. R. Statist. Soc. B* **58,** 155–176.

Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible disriminant analysis by optimal scoring. *Journal of American Statistical Association* **89,** 1255–1270.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison on methods for multi-class support vector machines. *IEEE Transactions on Neural Networks* **13,** 415–425.

Johnson, V. E. (2013). On numerical aspects of Bayesian model selection in high and ultrahigh-dimensional settings. *Bayesian Analysis* **8,** 741–758.

Johnson, V. E. and Rossell, D. (2012). Bayesian variable selection in high-dimensional settings. *Journal of the American Statistical Association* **107,** 649–660.

Nikooienejad, A., Wang, W., and Johnson, V. E. (2016). Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics* **32,** 1338–1345.

Pan, R., Wang, H., and Li, R. (2016). Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independent screening. *Journal of American Statistical Association* **111,** 169–179.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Teo, C. H., Vishwanathan, S. V. N., Smola, A., and Le, Q. V. (2010). Bundle methods for regularized risk minimization. *The Journal of Machine Learning Research* **11,** 311–365.

Torgo, L. (2010). *Data Mining using R: learning with case studies.* CRC Press.

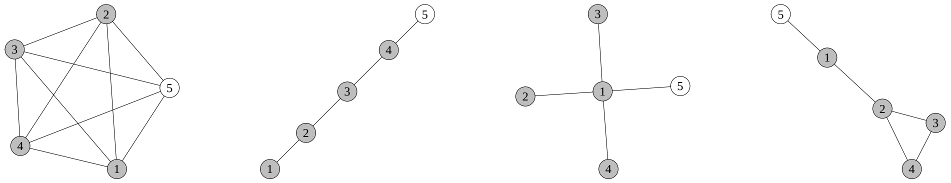
van de Geer, S. A. (2002). On Hoeffding's inequality for dependent random variables. In Dehling, H., Mikosch, T., and Sørensen, M., editors, *Empirical Process Techniques for Dependent Data*, pages 161–169. Birkhäuser, Basel.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

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

Xu, P., Zhu, J., Zhu, L., and Li, Y. (2014). Covariance-enhanced discriminant analysis. *Biometrika* **102,** 33–45.

Yu, Z., Dong, Y., and Shao, J. (2016). On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *Ann. Statist.* **44,** 2594–2623.



(a) CS and AR1 (b) Banded (c) Star (d) "Unstructured"

Figure S1: Various covariance structures within each block. The JI features are colored in gray and MI features in white. An edge between two features indicates that features are correlated.

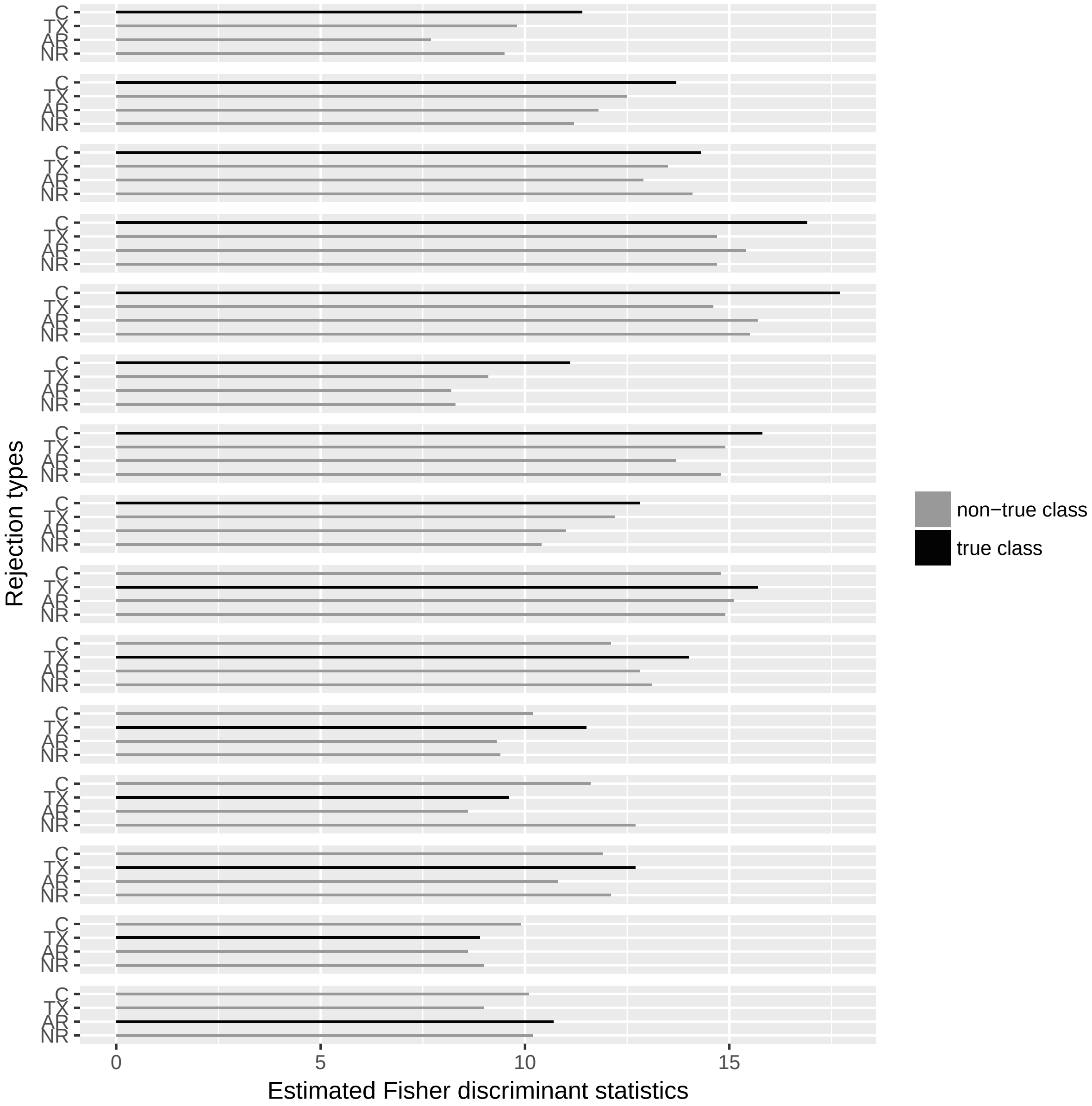


Figure S2: Estimated Fisher discriminant statistics in (4) with four rejection types for 15 random samples.

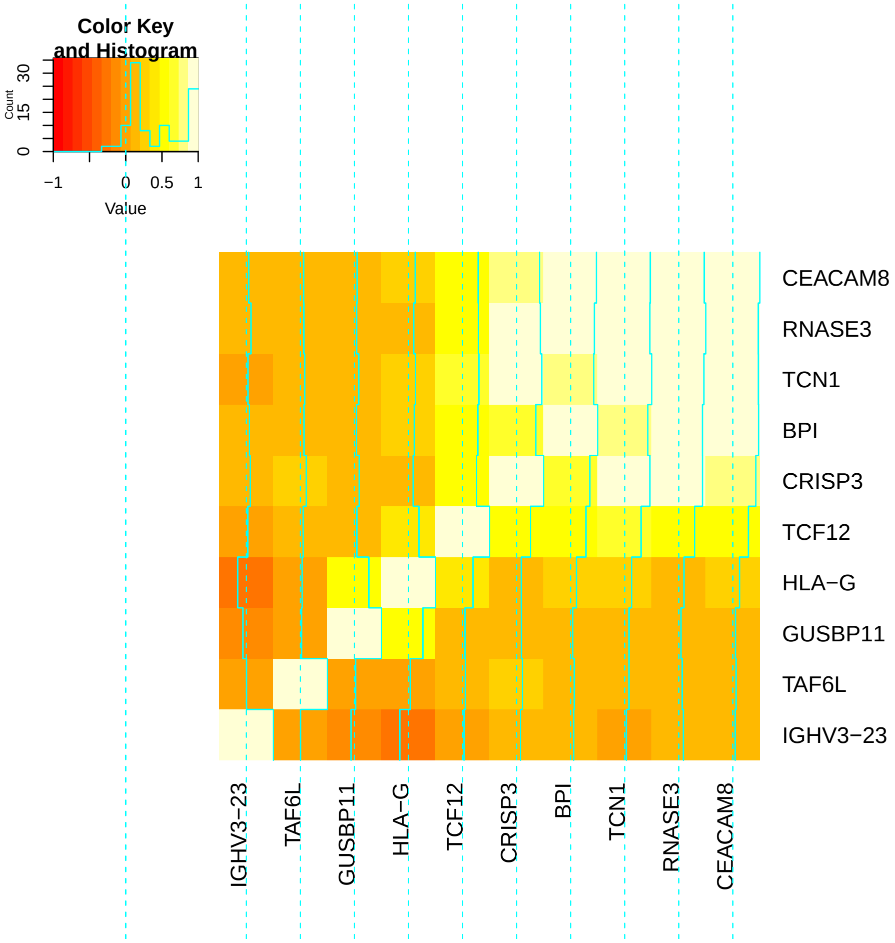


Figure S3: Heatmap of the correlation matrix between top 10 genes with the highest selection frequencies from the leave-one-out procedure.

Table S1: Means of the informative features

Features	Class 1	Class 2
X_1–X_4, X_{11}–X_{14}	0	0
X_5, X_{15}	0	2.5
X_6–X_{10}, X_{16}–X_{20}	1.5	-1.5

Table S2: Comparison of computation time (in minutes)

(n, p)	mLDA	ROAD	LPD	CED
$(200, 200)$	0.4	1.6	2.5	11.5
$(200, 1000)$	3.0	19.0	36.5	>300
$(200, 10000)$	4.5	>300	>300	>300
$(200, 50000)$	12.0	>300	>300	>300

NOTE: Competing methods include the regularized optimal affine discriminant (ROAD) (Fan et al., 2012), the linear programming discriminant (LPD) (Cai and Liu, 2011) and the covariance-enhanced discriminant analysis (CED) (Xu et al., 2014).

Table S3: Prediction error for different τ and α values under the CS structure

τ	5	10	15	20	25	30	35	40	45	50
	19.6	18.8	12.6	11.7	11.5	12.4	14.0	18.9	21.4	22.3
	(4.6)	(4.3)	(4.7)	(4.5)	(5.0)	(4.9)	(5.1)	(4.9)	(5.0)	(5.1)
α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
	14.3	9.8	9.6	9.6	9.7	9.9	9.9	10.4	14.1	
	(3.7)	(3.2)	(3.0)	(3.1)	(3.3)	(3.1)	(3.3)	(3.5)	(3.5)	

NOTE: numbers in parentheses are standard deviation.

Table S4: Comparisons with the other linear classification methods under heterogeneous "unstructured" covariance matrices

(ρ_1, ρ_2, ρ_3)		mLDA	MS	pairwiseLDA	SIR	penalizedLDA	bmrm	MGSDA	SIS
$(0.9, 0.7, 0.5)$	FP	28.3	76.4	47.1	65.9	–	35.2	66.5	5.3
		(2.9)	(4.1)	(4.0)	(9.6)		(6.3)	(11.2)	(0.3)
	FN	5.6	15.7	13.1	12.4	–	13.3	15.6	21.8
		(1.6)	(1.6)	(1.2)	(0.8)		(1.8)	(3.0)	(1.4)
	MMS	63.2	9931	–	–	–	9904	–	9974
		(4.7)	(325)				(369)		(426)
	ER	12.3	17.9	14.1	54.7	58.9	39.6	14.9	31.6
		(3.3)	(3.8)	(4.2)	(15.5)	(12.7)	(7.9)	(7.2)	(7.4)
$(0.9, 0.5, 0.1)$	FP	30.4	73.0	51.2	62.8	–	34.0	69.2	5.8
		(2.9)	(5.3)	(4.4)	(8.9)		(6.7)	(10.9)	(0.4)
	FN	10.7	14.4	13.6	13.0	–	13.7	14.9	22.5
		(1.6)	(2.7)	(1.3)	(1.0)		(1.6)	(3.2)	(2.1)
	MMS	8939.4	9916	–	–	–	9872	–	9438
		(276)	(290)				(341)		(397)
	ER	13.7	16.2	15.4	48.3	59.6	41.5	13.3	33.1
		(4.2)	(4.1)	(4.4)	(13.7)	(13.1)	(8.2)	(6.8)	(7.2)

NOTE: • The datasets were simulated using "unstructured" covariance matrices with class specific correlation coefficient specified in the first column. For example, $(0.9, 0.7, 0.5)$ means that the correlation coefficient for the first, second and third class, is 0.9, 0.7 and 0.5 respectively. • FP, average number of false positives; FN, average number of false negatives; MMS, the minimum number of features needed to include all informative features; ER, the number of misclassified cases; numbers in parentheses are interquartile ranges for MMS and standard deviations for FP, FN, and ER. MMS is reported only for methods that output the full ranks of all features. • The competing methods include penalizedLDA (Witten and Tibshirani, 2011), the regularized risk minimization package (bmrm) (Teo et al., 2010), the multi-group sparse discriminant analysis (MGSDA) (Gaynanova et al., 2016), pairwiseLDA (Pan et al., 2016), marginal sliced inverse regression (SIR) (Yu et al., 2016), the feature annealed independence rule (MS) (Fan and Fan, 2008), the sure independence screening (SIS) (Fan and Lv, 2008).

Table S5: Prediction accuracy comparisons with the non-linear classification methods when classes are not linearly separable

(ρ_1, ρ_2, ρ_3)	mLDA	mda	qda	rda	sc-rda	nn	svm	knn	nb
$(0.9, 0.8, 0.7)$	2.8	6.3	22.1	15.7	3.4	4.0	7.4	3.1	9.5
	(0.9)	(1.4)	(10.5)	(7.6)	(2.5)	(3.7)	(5.8)	(4.0)	(5.2)
$(0.9, 0.7, 0.5)$	4.1	0.7	13.2	11.8	2.3	1.9	8.2	1.5	6.6
	(1.1)	(0.9)	(9.9)	(7.1)	(2.0)	(3.6)	(6.1)	(3.9)	(5.4)
$(0.9, 0.5, 0.1)$	5.7	0.4	11.8	10.2	0.6	3.1	7.9	2.1	7.4
	(1.2)	(0.7)	(10.2)	(6.8)	(2.2)	(3.9)	(5.9)	(4.1)	(5.6)

NOTE: mda = mixed discriminant analysis (Hastie and Tibshirani, 1995; Hastie et al., 1994); qda = quadratic discriminant analysis (Ripley, 1996); rda = regularized discriminant analysis (Hastie et al., 1995); sc-rda = shrunken-centroids regularized discriminant analysis (Guo et al., 2005); nn = neural network (Ripley, 1996); svm = support vector machine (Hsu and Lin, 2002); knn = k-nearest neighbors (Torgo, 2010) and nb = naive Bayes (Efron, 2013). The datasets were simulated using "unstructured" covariance matrices with class specific correlation coefficient specified in the first column.

Table S6: Performance of mLDA with different correlation strengths

	FP	FN	MMS	ER
$\rho = 0.9$	7.8 (1.4)	0.5 (0.6)	37 (4.0)	4.9 (2.4)
$\rho = 0.5$	23.9 (2.5)	5.3 (0.7)	347 (66)	5.3 (2.7)
$\rho = 0.3$	28.7 (2.8)	11.2 (1.4)	4082 (144)	7.7 (2.6)
$\rho = 0.1$	37.8 (4.2)	12.3 (1.6)	8926 (263)	14.3 (2.9

Table S7: Means of the informative features

Features	Class 1	Class 2	...	Class K
X_1–X_9, X_{11}–X_{19}, ..., $X_{10(K-1)+1}$–X_{10K-1}	0	0	...	0
X_{10}	5	0	...	0
X_{20}	0	5	...	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_{10K}	0	0	...	5

Table S8: Comparisons with the pairwiseLDA

(n, K)	Method	FP	FN	ER
$(100, 10)$	mLDA	0 (0)	56.7 (22.3)	0.8 (0.7)
	pairwiseLDA	0 (0)	90.3 (0.4)	2.4 (1.5)
	pairwiseLDA.ebic	0 (0)	90.4 (0.1)	2.6 (1.2)
$(400, 20)$	mLDA	0 (0)	71.2 (24.6)	3.5 (3.1)
	pairwiseLDA	0 (0)	180.5 (0.3)	12.6 (3.4)
	pairwiseLDA.ebic	0 (0)	180.2 (0.1)	13.2 (4.2)
$(1600, 40)$	mLDA	0.2 (0.1)	178.2 (44.2)	5.3 (6.1)
	pairwiseLDA	0 (0)	361.0 (1.6)	86.4 (12.6)
	pairwiseLDA.ebic	0.2 (0.1)	360.3 (1.1)	91.0 (10.4)

- Numbers in parentheses are standard deviations for FP, FN, and ER.

Table S9: Comparisons with the pMOM-logistic

	FP	FN	ER
mLDA	1.4 (2.0)	3.6 (2.6)	4.7 (6.6)
pMOM-logistic	0 (0)	17.6 (1.4)	18.4 (4.2)