Biometrika (2012), **99**, 1, *pp.* 1–12

Printed in Great Britain
Advance Access publication on 31 July 2012

Supplementary material for "Covariance-Enhanced discriminant analysis"

BY PEIRONG XU
Department of Mathematics, Southeast University, Nanjing, 211189, China
xupeirong@seu.edu.cn

JI ZHU
Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
jizhu@umich.edu

LIXING ZHU
Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
lzhu@hkbu.edu.hk

AND YI LI
Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
yili@umich.edu

1. TECHNICAL PROOFS

Proof of Theorem 1. The proof is summarized in the following three steps. First, we prove $Q_n(\omega^*, \mu^*, \Omega^*) \geq Q_n(\omega, \mu^*, \Omega^*)$ for $\|\omega_{(1)} - \omega^*_{(1)}\|_2^2 = O_p(n^{-1})$. In Step 2, we show that $Q_n(\omega, \mu^*, \Omega^*) \geq Q_n(\omega, \mu^*, \Omega)$ for $\|\Omega - \Omega^*\|_F^2 = O_p\{(p_n + a_n)\log p_n/n\}$. In Step 3, we prove that $Q_n(\omega, \mu^*, \Omega) \geq Q_n(\omega, \mu, \Omega)$ for $\|\mu - \mu^*\|_2^2 = O_p(p_n \log p_n/n)$. The following are the details.

Step 1. Let $\Delta_{\omega_{(1)}} = \omega_{(1)} - \omega^*_{(1)}$, and $h(\omega_{(1)}) = \sum_{i=1}^n \sum_{k=1}^K \tau_{ik} \log \omega_k$, where $\omega_K = 1 - \sum_{k=1}^{K-1} \omega_k$. We denote by $J_\omega = (\delta_1, \ldots, \delta_K)^T$ the Jacobian matrix, where $\delta_k (1 \leq k < K)$ is a $(K-1)$-dimensional unit vector with the kth component being 1, and δ_K is a $(K-1)$-dimensional vector of ones. An application of Taylor expansion yields

$$\begin{aligned} &Q_n(\omega, \mu^*, \Omega^*) - Q_n(\omega^*, \mu^*, \Omega^*) \\ &= \frac{1}{n} J_\omega^T \frac{\partial h(\omega^*_{(1)})}{\partial \omega} \Delta_{\omega_{(1)}} - \frac{1}{2} \Delta^T_{\omega_{(1)}} J_\omega^T \left\{ -\frac{1}{n} \frac{\partial^2 h(\omega^*_{(1)})}{\partial \omega \partial \omega^T} \right\} J_\omega \Delta_{\omega_{(1)}} \\ &\quad + o_p \left\{ \Delta^T_{\omega_{(1)}} J_\omega^T \left(-\frac{1}{n} \frac{\partial^2 h(\omega^*_{(1)})}{\partial \omega \partial \omega^T} \right) J_\omega \Delta_{\omega_{(1)}} \right\}. \end{aligned} \quad (1)$$

Note that $n^{-1} \sum_{i=1}^n \{\tau_{ik} \omega_k^{*-1} - \tau_{iK} \omega_K^{*-1}\} = o_p(1)$ because $\mathrm{E}\tau_{ik} = \omega^*_k$ for $k = 1, \ldots, K$. Consequently, we have

$$\frac{1}{n} J_\omega^T \frac{\partial h(\omega^*_{(1)})}{\partial \omega} \Delta_{\omega_{(1)}} \leq n^{-1/2} O_p(1) \|\Delta_{\omega_{(1)}}\|_1 \leq (K-1)^{1/2} O_p(n^{-1/2}) \|\Delta_{\omega_{(1)}}\|_2.$$

Further, since $n^{-1}\sum_{i=1}^{n}\tau_{ik}\omega_k^{*-2}\xrightarrow{P}\omega_k^{*-1}$ for $k=1,\dots,K$, we have

$$J_\omega^T\left\{-\frac{1}{n}\frac{\partial^2 h(\omega^*_{(1)})}{\partial\omega\partial\omega^T}\right\}J_\omega\xrightarrow{P}J_\omega^T H J_\omega>0,$$

where H is a $K\times K$ diagonal matrix with the kth element ω_k^{*-1}. Hence,

$$\frac{1}{2}\Delta^T_{\omega_{(1)}}J_\omega^T\left\{-\frac{1}{n}\frac{\partial^2 h(\omega^*_{(1)})}{\partial\omega\partial\omega^T}\right\}J_\omega\Delta_{\omega_{(1)}}\geq\frac{1}{2}O_p(1)\|\Delta_{\omega_{(1)}}\|_2^2,$$

implying that it dominates both the first and third terms in (1) uniformly in $\|\omega_{(1)}-\omega^*_{(1)}\|_2^2=O_p(n^{-1})$. Therefore, $Q_n(\omega^*,\mu^*,\Omega^*)\geq Q_n(\omega,\mu^*,\Omega^*)$ for $\|\omega_{(1)}-\omega^*_{(1)}\|_2^2=O_p(n^{-1})$.

Step 2. Let $\Delta_\Omega=\Omega-\Omega^*$ and $S=S(\mu^*)$. Consider the difference

$$Q_n(\omega,\mu^*,\Omega)-Q_n(\omega,\mu^*,\Omega^*)=B_1-B_2-B_3,$$

where

$$\begin{aligned}B_1&=2^{-1}\left(\log|\Omega|-\log|\Omega^*|\right)-2^{-1}\mathrm{tr}(S\Delta_\Omega),\\B_2&=\lambda_{2n}\sum_{(j,l)\in\mathcal{A}^c,j\neq l}(|\Omega_{jl}|-|\Omega^*_{jl}|),\\B_3&=\lambda_{2n}\sum_{(j,l)\in\mathcal{A}}(|\Omega_{jl}|-|\Omega^*_{jl}|).\end{aligned}$$

An application of Taylor expansion with the integral remainder yields that

$$\log|\Omega|-\log|\Omega^*|=\mathrm{tr}(\Sigma^*\Delta_\Omega)-\vec{\Delta}_\Omega^T\left\{\int_0^1(1-v)\Omega_v^{-1}\otimes\Omega_v^{-1}dv\right\}\vec{\Delta}_\Omega,$$

where $\Omega_v=\Omega^*+v\Delta_\Omega$ with $0\leq v\leq 1$, $\vec{\Delta}_\Omega$ is the vectorization of Δ_Ω, and $\otimes$ is the Kronecker product. Therefore, B_1 can be written as $B_1=-2^{-1}(I_1+I_2)$, where

$$\begin{aligned}I_1&=\mathrm{tr}\left\{(S-\Sigma^*)\Delta_\Omega\right\},\\I_2&=\vec{\Delta}_\Omega^T\left\{\int_0^1(1-v)\Omega_v^{-1}\otimes\Omega_v^{-1}dv\right\}\vec{\Delta}_\Omega.\end{aligned}$$

First consider I_1. Let s_{jl}, σ^*_{jl}, and $\Delta_{\Omega jl}$ be respectively the (j,l)th element of S, Σ^* and Δ_Ω. Denote by $\mathcal{C}=\{(j,j):j=1,\dots,p_n\}$. Then, it is clear that $|I_1|\leq I_{11}+I_{12}$, where

$$\begin{aligned}I_{11}&=|\sum_{(j,l)\in\mathcal{A}\cup\mathcal{C}}(s_{jl}-\sigma^*_{jl})\Delta_{\Omega jl}|,\\I_{12}&=|\sum_{(j,l)\in\mathcal{A}^c,j\neq l}(s_{jl}-\sigma^*_{jl})\Delta_{\Omega jl}|.\end{aligned}$$

Let $z_i=\sum_{k=1}^K\tau_{ik}(x_i-\mu_k^*)$ for $i=1,\dots,n$. By the assumption, $z_i=(z_{i1},\dots,z_{ip})^T$'s are i.i.d. p-variate normal random variables with mean 0 and covariance matrix Σ^*. Note that

$s_{jl} = n^{-1}\sum_{i=1}^{n} z_{ij}z_{il}$. Using Lemma 3 in Bickel & Levina (2008), we have

$$\begin{aligned} I_{11} &\le (p_n + a_n)^{1/2} \max_{(j,l)\in\mathcal{A}\cup\mathcal{C}} |s_{jl} - \sigma^*_{jl}| \cdot \|\Delta_\Omega\|_F \\ &\le O_p[\{(p_n+a_n)\log p_n/n\}^{1/2}] \cdot \|\Delta_\Omega\|_F \\ &= O_p\{(p_n+a_n)\log p_n/n\}. \end{aligned}$$

Consider $B_2 - I_{12}$ for penalties. Note that $\Delta_{\Omega jl} = \Omega_{jl}$ for all $(j,l) \in \mathcal{A}^c$, $j \neq l$. Invoking Lemma 3 in Bickel & Levina (2008) again, we have

$$\begin{aligned} B_2 - I_{12} &\ge \lambda_{2n} \sum_{(j,l)\in\mathcal{A}^c, j\neq l} |\Omega_{jl}| - \max_{(j,l)} |s_{jl} - \sigma^*_{jl}| \sum_{(j,l)\in\mathcal{A}^c, j\neq l} |\Delta_{\Omega jl}| \\ &\ge \sum_{(j,l)\in\mathcal{A}^c, j\neq l} [\lambda_{2n} - O_p\{(\log p_n/n)^{1/2}\}] |\Omega_{jl}| \\ &\ge 0 \end{aligned}$$

for $\lambda_{2n}^2 = O(\log p_n/n)$. For the term B_3, we have

$$\begin{aligned} B_3 &= \lambda_{2n} \sum_{(j,l)\in\mathcal{A}} (|\Omega_{jl}| - |\Omega^*_{jl}|) \\ &\le \lambda_{2n} \sum_{(j,l)\in\mathcal{A}} |\Delta_{\Omega jl}| \\ &\le \lambda_{2n} a_n^{1/2} \|\Delta_\Omega\|_F \\ &= O_p\{(p_n+a_n)\log p_n/n\}. \end{aligned}$$

Finally, we bound I_2. Recall that $\lambda_{\min}(M) = \min_{\|x\|=1} x^T M x$ for any symmetric matrix M. Then, under condition (A), we have

$$\begin{aligned} I_2 &\ge \int_0^1 (1-v) \min_{0\le v\le 1} \lambda_{\min}(\Omega_v^{-1} \otimes \Omega_v^{-1}) dv \cdot \|\vec{\Delta}_\Omega\|_2^2 \\ &= \frac{1}{2}\|\vec{\Delta}_\Omega\|_2^2 \min_{0\le v\le 1} \lambda_{\max}^{-2}(\Omega_v) \\ &\ge \frac{1}{2}\|\vec{\Delta}_\Omega\|_2^2 \{\kappa_1 + o(1)\}^{-2} \\ &= C_1 (p_n + a_n)\log p_n/n, \end{aligned}$$

for a large constant C_1. To derive the above inequality, we have used $\|\Delta_\Omega\| \le \|\Delta_\Omega\|_F = O\{(\log p_n)^{(1-m)/2}\} = o(1)$ by our assumption. Therefore, I_2 dominates both I_{11} and B_3 with a large constant C_1. With $B_2 - I_{12} \ge 0$, this completes the proof of the Step 2.

Step 3. Let $\Delta_{\mu_k} = (\Delta_{\mu_{k1}}, \ldots, \Delta_{\mu_{kp_n}})^T = \mu_k - \mu_k^*$, for $k = 1, \ldots, K$, and $\Delta_\mu = \mu - \mu^*$. Then, for each $1 \le k \le K$, $\Delta_{\mu_k} = (I_{p_n} \otimes e_k^T)\Delta_\mu$, where I_{p_n} is a $p_n \times p_n$ identity matrix and e_k is a K-dimensional unit vector with kth component 1. For the sake of simplicity, let $z_i = \sum_{k=1}^K \tau_{ik}(x_i - \mu_k^*)$ and $E_i = \sum_{k=1}^K \tau_{ik}(I_{p_n} \otimes e_k^T)$, for $i = 1, \ldots, n$. Consider the difference

$$Q_n(\omega, \mu, \Omega) - Q_n(\omega, \mu^*, \Omega) = I_1' - I_2' + I_3'$$

where

$$I_1' = \frac{1}{n}\sum_{i=1}^{n} z_i^T \Omega E_i \Delta_\mu,$$

$$I_2' = \frac{1}{2n}\sum_{i=1}^{n} \Delta_\mu^T E_i^T \Omega E_i \Delta_\mu^T,$$

$$I_3' = -\lambda_{1n}\sum_{j=1}^{p_n}\sum_{1\le k<k'\le K}\left\{|\mu_{kj}-\mu_{k'j}| - |\mu_{kj}^* - \mu_{k'j}^*|\right\}.$$

Let $\Delta_\mu^{(s)}$ be the sth component of Δ_μ, and δ_s' be a (Kp_n)-dimensional unit vector with sth component 1, for $s = 1, \dots, Kp_n$. Then, it can be seen that $|I_1'| = \sum_{s=1}^{Kp_n} \eta_s \Delta_\mu^{(s)}$, where

$$\eta_s = \frac{1}{n}\sum_{i=1}^{n} z_i^T \Omega E_i \delta_s',$$

for $s = 1, \dots, Kp_n$. Now, consider the event $\mathcal{F} = \bigcap_{s=1}^{Kp_n}\{|\eta_s| \le \lambda_{1n}\}$. Since $\|\Omega - \Omega^*\| = o_p(1)$, we have $\|\Omega\Sigma^* - I_{p_n}\| = o_p(1)$ by condition (A). Thus, $\|\Omega\Sigma^*\Omega - \Omega^*\| = \|(\Omega\Sigma - I_{p_n})(\Omega - \Omega^*)\| = o_p(1)$. Consequently,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\delta_s'^T E_i^T \Omega\Sigma^*\Omega E_i \delta_s' &= \frac{1}{n}\sum_{i=1}^{n}\delta_s'^T E_i^T \Omega^* E_i \delta_s' + o_p(1)\\ &\triangleq M_s + o_p(1).\end{aligned}$$

Therefore, using the probability bound on the tail of the standard Gaussian distribution, we know that

$$\begin{aligned}\Pr(\mathcal{F}^c) &\le \sum_{s=1}^{Kp_n}\Pr(n^{1/2}|\eta_s| > n^{1/2}\lambda_{1n})\\ &\le O_p(1)\cdot\sum_{s=1}^{Kp_n}\exp\left(-\frac{n\lambda_{1n}^2}{2M_s}\right)\\ &\le O_p(Kp_n)\exp\left[-\frac{n\lambda_{1n}^2}{2\max_s\{M_s\}}\right]\end{aligned}$$

which tends to 0 when $\lambda_{1n} = [2\max_s\{M_s\}\log p_n/n]^{1/2}$. Consequently, by considering the event $\mathcal{F}$, we have

$$|I_1'| \le \sum_{s=1}^{Kp_n}|\eta_s||\Delta_\mu^{(s)}| \le \lambda_{1n}\|\Delta_\mu\|_1$$

with a probability tending to one. Note that $|I_3'| \le \lambda_{1n}\sum_{j=1}^{p_n}\sum_{1\le k<k'\le K}|\Delta_{\mu_{kj}} - \Delta_{\mu_{k'j}}| \le (K-1)\lambda_{1n}\|\Delta_\mu\|_1$. Thus, with a probability tending to one, we have

$$\begin{aligned}|I_1'| + |I_3'| &\le K\lambda_{1n}\|\Delta_\mu\|_1\\ &\le K^{3/2}p_n^{1/2}\lambda_{1n}\|\Delta_\mu\|_2\\ &= O_p(p_n\log p_n/n).\end{aligned}$$

The proof can be concluded from proving that $I_2' \geq C_2 p_n \log p_n / n$ for some constant C_2.

Since $\|\Omega - \Omega^*\| = o_p(1)$, we have

$$\begin{aligned} I_2' &= \frac{1}{2n} \sum_{i=1}^{n} \Delta_\mu^T E_i^T \Omega^* E_i \Delta_\mu^T + o_p(1) \\ &\geq \frac{1}{2\kappa_2} \left\{ \frac{1}{n} \sum_{k=1}^{K} n_k \|\Delta_{\mu_k}\|_2^2 \right\} \\ &\geq \frac{1}{2\kappa_2} \min_{1 \leq k \leq K} \frac{n_k}{n} \cdot \|\Delta_\mu\|_2^2 \\ &= C_2 p_n \log p_n / n \end{aligned}$$

with a probability tending to one. This finishes the proof. □

Before proving Theorem 2, we first prove the following lemma.

LEMMA 1. *Let* $\|\cdot\|_{FP} : R^K \to R$ *be the fused penalty* $\|x\|_{FP} = \sum_{1 \leq k < k' \leq K} |x_k - x_{k'}|$. *Then,* $\|\cdot\|_{FP}$ *is convex and, for any* $x \in R^K$*, the subdifferential* $\partial \|x\|_{FP}$ *is the set of all vectors* $s \in R^K$ *such that*

$$s_i = \sum_{j \neq i} \text{sgn}(x_i - x_j),$$

for $i = 1, \ldots, K$.

Proof. For each $j = 1, \ldots, K-1$, let $H^{(j)}$ be a $(K-j) \times K$ matrix with $H_{ii}^{(j)} = -1$, $H_{i,i+j}^{(j)} = 1$ for $i = 1, \ldots, K-j$ and 0 otherwise. Denote by H the $K(K-1)/2 \times K$ matrix with jth row block matrix $H^{(j)}$. Then, for any $x \in R^K$, $\|x\|_{FP} = \|Hx\|_1$. Note that the l_1 norm $\|\cdot\|_1$ is convex and $\|\cdot\|_{FP}$ is the composition of a linear functional by the l_1 norm. Hence, $\|\cdot\|_{FP}$ is convex. Further, by the definition of the subdifferential of the l_1 norm, for any $y \in R^K$,

$$\|Hy\|_1 \leq \|Hx\|_1 + < H(y - x), \upsilon > \tag{2}$$

holds if and only if $\upsilon \in \mathcal{W}_\upsilon \subset R^{K(K-1)/2}$, where $\mathcal{W}_\upsilon$ is the set of all vectors $\upsilon = \text{sgn}(Hx)$. Note that

$$\begin{aligned} < H(y-x), \text{sgn}(Hx) > &= \sum_{1 \leq k < k' \leq K} \{(y_{k'} - x_{k'}) - (y_k - x_k)\} \text{sgn}(x_{k'} - x_k) \\ &= 2^{-1} \sum_{k' \neq k} \{(y_{k'} - x_{k'}) - (y_k - x_k)\} \text{sgn}(x_{k'} - x_k) \\ &= \sum_{k=1}^{K} (y_k - x_k) \left\{ \sum_{k' \neq k} \text{sgn}(x_k - x_{k'}) \right\}. \end{aligned}$$

Thus, equation (2) is equivalent to

$$\|y\|_{FP} \leq \|x\|_{FP} + < y - x, s >,$$

where s is a K-dimensional vector with ith component $s_i = \sum_{j \neq i} \text{sgn}(x_i - x_j)$. The set of all such vectors s is, therefore, $\partial \|x\|_{FP}$. □

Proof of Theorem 2. First, we prove the sparsistency of the precision matrix estimator $\hat{\Omega}$. The derivative of $Q_n(\omega, \mu, \Omega)$ w.r.t. Ω_{jl} for $(j, l) \in \mathcal{A}^c, j \neq l$ at $(\hat{\omega}, \hat{\mu}, \hat{\Omega})$ is

$$\frac{\partial Q_n(\hat{\omega}, \hat{\mu}, \hat{\Omega})}{\partial \Omega_{jl}} = \hat{\sigma}_{jl} - s_{jl} - 2\lambda_{2n}\text{sgn}(\hat{\Omega}_{jl}),$$

where s_{jl} is the (j, l)th element of $S = S(\hat{\mu})$ and sgn(a) denotes the sign of a. Note that

$$\begin{aligned} S &= S(\mu^*) - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\Delta_{\mu_k}(x_i - \mu_k^*)^T \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}(x_i - \mu_k^*)\Delta_{\mu_k}^T + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\Delta_{\mu_k}\Delta_{\mu_k}^T \\ &\triangleq I_1 - I_2 - I_3 + I_4. \end{aligned}$$

Then, we decompose $\hat{\sigma}_{jl} - s_{jl} = A_1 + A_2 + A_3$, where

$$A_1 = \hat{\sigma}_{jl} - \sigma_{jl}^*,\ A_2 = \sigma_{jl}^* - I_{1jl},\ A_3 = I_{2jl} + I_{3jl} - I_{4jl},$$

where B_{jl} denotes the (j, l)th element of matrix B. Now, consider the order of A_1. Under condition (A), we have $\|\Sigma^*\| = O(1)$ and $\|\hat{\Sigma}\| \leq \{\lambda_{\min}(\hat{\Omega} - \Omega^*) + \lambda_{\min}(\Omega^*)\}^{-1} = O_p(1)$. Thus,

$$\begin{aligned} |A_1| &\leq \|\hat{\Sigma} - \Sigma^*\| \\ &\leq \|\hat{\Sigma}\| \cdot \|\hat{\Omega} - \Omega^*\| \cdot \|\Sigma^*\| \\ &= O_p(\rho_{n2}^{1/2}). \end{aligned}$$

By Lemma 3 in Bickel & Levina (2008), we have $|A_2| = O_p\{(\log p_n/n)^{1/2}\}$. Now, we estimate the order of A_3. Since $\max_{1\leq j\leq p_n} \|\hat{\mu}_{(j)} - {\mu^*}_{(j)}\|_2^2 = O_p(\rho_{n1})$ for a sequence $\rho_{n1} \to 0$, we have

$$\begin{aligned} |I_{2jl}| &= \left|\frac{1}{n}\sum_{i=1}^{n} z_{il}\left(\sum_{k=1}^{K}\tau_{ik}\Delta_{\mu_{kj}}\right)\right| \\ &\leq O_p(1) \cdot \left(\frac{1}{n}\sum_{k=1}^{K} n_k\Delta_{\mu_{kj}}^2\right)^{1/2} \\ &\leq O_p(1) \cdot \left(\sum_{k=1}^{K}\Delta_{\mu_{kj}}^2\right)^{1/2} = O_p(\rho_{n1}^{1/2}). \end{aligned}$$

Similarly, we have $|I_{3jl}| \leq O_p(\rho_{n1}^{1/2})$ and $|I_{4jl}| \leq O_p(\rho_{n1})$. Thus, $|A_3| \leq O_p(\rho_{n1}^{1/2})$. Combining above results yields that

$$\max_{j,l} |\hat{\sigma}_{jl} - s_{jl}| = O_p\{(\log p_n/n)^{1/2} + \rho_{n1}^{1/2} + \rho_{n2}^{1/2}\}.$$

Hence, we need to have $\log p_n/n + \rho_{n1} + \rho_{n2} = O(\lambda_{2n}^2)$ in order to have the sign of $\partial Q_n(\hat{\omega}, \hat{\mu}, \hat{\Omega})/\partial\Omega_{jl}$ that depends on sgn$(\hat{\Omega}_{jl})$ with a probability tending to one. This completes the proof of Theorem 2(i).

Next, we prove the second result of Theorem 2. The main idea of the proof is inspired by Rinaldo (2009). Let $\bar{\tau}_k = n^{-1}\sum_{i=1}^n \tau_{ik}$, $k = 1, \ldots, K$. Then, by Lemma 1, we know that

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^n \tau_{ik}x_i - \lambda_{1n}\bar{\tau}_k^{-1}\hat{\Sigma}\hat{s}_k$$

where $\hat{s}_k = (\hat{s}_{k1}, \ldots, \hat{s}_{kp_n})^T$ with jth element $\hat{s}_{kj} = \sum_{t\neq k}\text{sgn}(\hat{\mu}_{kj} - \hat{\mu}_{tj})$. Hence, for $k, k' = 1, \ldots, K$ and $k < k'$,

$$\hat{\mu}_{k'j} - \hat{\mu}_{kj} = \sum_{i=1}^n \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k}\right)x_{ij} - \lambda_{1n}e_j^T\hat{\Sigma}(\bar{\tau}_{k'}^{-1}\hat{s}_{k'} - \bar{\tau}_k^{-1}\hat{s}_k)$$

where e_k is a p_n-dimensional unit vector with the kth component 1. Since $\lambda_{\max}(\hat{\Sigma}) = \|\hat{\Sigma}\| \leq \{\lambda_{\min}(\hat{\Omega} - \Omega^*) + \lambda_{\min}(\Omega^*)\}^{-1} \leq \kappa_2$ and $|\bar{\tau}_{k'}^{-1}\hat{s}_{k'l} - \bar{\tau}_k^{-1}\hat{s}_{kl}| \leq 2(K-1)$ for $l = 1, \ldots, p_n$, we have

$$\begin{aligned}\|e_j^T\hat{\Sigma}(\bar{\tau}_{k'}^{-1}\hat{s}_{k'} - \bar{\tau}_k^{-1}\hat{s}_k)\|_2 &\leq \lambda_{\max}(\hat{\Sigma})\|\bar{\tau}_{k'}^{-1}\hat{s}_{k'} - \bar{\tau}_k^{-1}\hat{s}_k\|_2 \\ &\leq 2p_n^{1/2}\kappa_2(K-1). \qquad (3)\end{aligned}$$

As a result, the event $\{\hat{\mathcal{B}} = \mathcal{B}\}$ occurs in probability if both

$$\max_{\mathcal{B}}\left|\sum_{i=1}^n \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k}\right)x_{ij}\right| < 2\lambda_{1n}p_n^{1/2}\kappa_2(K-1) \qquad (4)$$

and

$$\min_{\mathcal{B}^c}\left|\sum_{i=1}^n \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k}\right)x_{ij} - \lambda_{1n}e_j^T\hat{\Sigma}(\bar{\tau}_{k'}^{-1}\hat{s}_{k'} - \bar{\tau}_k^{-1}\hat{s}_k)\right| > 0 \qquad (5)$$

hold with a probability tending to 1 and $n \to \infty$.

We first consider (4). For the sake of simplicity, let $M = 2\kappa_2(K-1)$ and $a_{kk'i} = \tau_{ik'}/n_{k'} - \tau_{ik}/n_k$, $i = 1, \ldots, n$. Then, by condition (C)(i), we know that

$$\max_{\mathcal{B}}\left|\sum_{i=1}^n \left(\frac{\tau_{ik'}}{n_{k'}} - \frac{\tau_{ik}}{n_k}\right)x_{ij}\right| \leq \max_{\mathcal{B}}\left|\sum_{i=1}^n a_{kk'i}\epsilon_{ij}\right| + o_p(\lambda_{1n}p_n^{1/2}),$$

where $\epsilon_{ij} = x_{ij} - \sum_{k=1}^K \tau_{ik}\mu_{kj}^*$, which follows normal distribution with mean 0 and variance σ_{jj}^*. Let $\xi_j^{kk'} = \sum_{i=1}^n a_{kk'i}\epsilon_{ij}$, $k, k' = 1, \ldots, K$, $k < k'$ and $j = 1, \ldots, p_n$. It is easy to show that $\mathrm{E}\xi_j^{kk'} = 0$, $\text{var}(\xi_j^{kk'}) = \sum_{i=1}^n a_{kk'i}^2\sigma_{jj}^* \leq 2\sigma_{jj}^*$, and $\text{cov}(\xi_j^{kk'}, \xi_t^{ll'}) = \sum_{i=1}^n a_{kk'i}a_{ll't}\sigma_{jt}^*$ for each $(k, k', j) \neq (l, l', t)$. For $(k, k', j) \in \mathcal{B}$, let $\zeta_j^{kk'} \sim N(0, \sum_{i=1}^n a_{kk'i}^2\sigma_{jj}^*)$ such that

$$\begin{aligned}\mathrm{E}(\zeta_j^{kk'})^2 &= \mathrm{E}(\xi_j^{kk'})^2, &&\text{for all } (k, k', j) \in \mathcal{B}, \\ \mathrm{E}(\zeta_j^{kk'}\zeta_t^{ll'}) &\geq \mathrm{E}(\xi_j^{kk'}\xi_t^{ll'}), &&\text{for all } (k, k', j), (l, l', t) \in \mathcal{B} \text{ and } j \neq t.\end{aligned}$$

Then, by Slepian's inequality (Ledoux & Talagrand, 1991) and Chernoff's bound for standard Gaussian variables, we have

$$\begin{aligned}\Pr(\max_{\mathcal{B}} |\xi_j^{kk'}| \geq \lambda_{1n} p_n^{1/2} M) &\leq \Pr(\max_{\mathcal{B}} |\zeta_j^{kk'}| \geq \lambda_{1n} p_n^{1/2} M) \\ &\leq \sum_{\mathcal{B}} \Pr(|\zeta_j^{kk'}| \geq \lambda_{1n} p_n^{1/2} M) \\ &\leq \sum_{\mathcal{B}} 2\exp\left\{-\frac{\lambda_{1n}^2 p_n M^2}{4b^*_{\max}}\right\} \\ &= 2\exp\left\{-\frac{\lambda_{1n}^2 p_n M^2}{4b^*_{\max}} + \log|\mathcal{B}|\right\},\end{aligned}$$

which vanishes under condition (C)(i).

In order to verify (5), it is sufficient to show that

$$\max_{\mathcal{B}^c} \left|\sum_{i=1}^n a_{kk'i}\epsilon_{ij} - \lambda_{1n} e_j^T \hat{\Sigma}(\bar{\tau}_{k'}^{-1}\hat{s}_{k'} - \bar{\tau}_k^{-1}\hat{s}_k)\right| \leq \alpha_n^{\min},$$

with probability tending to one as $n \to \infty$. Using the triangle inequality, we only need to show that

$$\max_{\mathcal{B}^c} \left|\lambda_{1n} e_j^T \hat{\Sigma}(\bar{\tau}_{k'}^{-1}\hat{s}_{k'} - \bar{\tau}_k^{-1}\hat{s}_k)\right| \leq \alpha_n^{\min}/2 \tag{6}$$

and

$$\max_{\mathcal{B}^c} \left|\sum_{i=1}^n a_{kk'i}\epsilon_{ij}\right| \leq \alpha_n^{\min}/2. \tag{7}$$

Because of (3), it is easy to see that the inequality (6) holds under condition (C)(ii). Then, we turn to (7). For $(k, k', j) \in \mathcal{B}^c$, let $\zeta_j^{kk'} \sim N(0, 2b^*_{\max})$ so that

$$\begin{aligned}\mathrm{E}(\zeta_j^{kk'})^2 &= \mathrm{E}(\xi_j^{kk'})^2, \qquad \text{for all } (k, k', j) \in \mathcal{B}^c, \\ \mathrm{E}(\zeta_j^{kk'}\zeta_t^{ll'}) &\geq \mathrm{E}(\xi_j^{kk'}\xi_t^{ll'}), \quad \text{for all } (k, k', j), (l, l', t) \in \mathcal{B}^c \text{ and } j \neq t.\end{aligned}$$

Then, again, by Slepian's inequality and Chernoff's bound for standard Gaussian variables, we have

$$\begin{aligned}\Pr(\max_{\mathcal{B}^c} |\xi_j^{kk'}| \geq \alpha_n^{\min}/2) &\leq \Pr(\max_{\mathcal{B}^c} |\zeta_j^{kk'}| \geq \alpha_n^{\min}/2) \\ &\leq \sum_{\mathcal{B}^c} 2\exp\left\{-\frac{(\alpha_n^{\min})^2}{16b^*_{\max}}\right\} \\ &= 2\exp\left\{-\frac{(\alpha_n^{\min})^2}{16b^*_{\max}} + \log|\mathcal{B}^c|\right\},\end{aligned}$$

which vanishes under condition (C)(ii). Hence, the proof of Theorem 2(ii) is completed. □

Proof of Theorem 3. Given the estimates $\hat{\omega}$, $\hat{\mu}$ and $\hat{\Omega}$, a new observation x^* is assigned to the kth class if

$$x^{*T}\hat{\Omega}(\hat{\mu}_k - \hat{\mu}_l) > \log(\hat{\omega}_l/\hat{\omega}_k) + \{(\tilde{\mu}_k + \tilde{\mu}_l)/2\}^T\hat{\Omega}(\hat{\mu}_k - \hat{\mu}_l) \tag{8}$$

for $l = 1, \ldots, K$ and $l \neq k$, where $\tilde{\mu}_s = \sum_{i=1}^n I(y_i = s)x_i / \sum_{i=1}^n I(y_i = s)$, $s = 1, \ldots, K$.

Given data (y_i, x_i) for $i = 1, \dots, n$, the conditional misclassification rate of the proposed method is given by

$$R_n = \frac{1}{2}\sum_{k=1}^{2} \Phi\left\{\frac{(-1)^k \hat{\delta}^T \hat{\Omega}(\mu_k^* - \tilde{\mu}_k) - \hat{\delta}^T \hat{\Omega}\tilde{\delta}/2}{\sqrt{\hat{\delta}^T \hat{\Omega}\Sigma^* \hat{\Omega}\hat{\delta}}}\right\},$$

where $\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_2$ and $\tilde{\delta} = \tilde{\mu}_1 - \tilde{\mu}_2$.

(i) Since $\|\hat{\Omega} - \Omega^*\|^2 = O_p(\rho_{n2})$ for a sequence $\rho_{n2} \to 0$, we have

$$\begin{aligned} \|\hat{\Sigma} - \Sigma^*\| &= \|\hat{\Sigma}(\hat{\Omega} - \Omega^*)\Sigma^*\| \\ &\le \|\hat{\Sigma}\| \cdot \|\hat{\Omega} - \Omega^*\| \cdot \|\Sigma^*\| \\ &\le \|\hat{\Sigma}\| \cdot O_p(\kappa_2 \rho_{n2}^{1/2}). \end{aligned}$$

Note that $\|\hat{\Sigma}\| \le \{\lambda_{\min}(\hat{\Omega} - \Omega^*) + \lambda_{\min}(\Omega^*)\}^{-1} = O_p(1)$. Hence,

$$\|\hat{\Sigma} - \Sigma^*\|^2 = O_p(\rho_{n2}).$$

Consequently,

$$\hat{\delta}^T \hat{\Omega}\Sigma^* \hat{\Omega}\hat{\delta} = \hat{\delta}^T \hat{\Omega}\hat{\delta}\{1 + O_p(\rho_{n2}^{1/2})\} = \hat{\delta}^T \Omega^* \hat{\delta}\{1 + O_p(\rho_{n2}^{1/2})\}.$$

Without loss of generality, we assume that $\hat{\delta} = (\hat{\delta}_1^T, 0^T)^T$, where $\hat{\delta}_1$ is the $\hat{b}_n$-dimensional vector containing nonzero components of $\hat{\delta}$. Let $\delta_\mu^* = (\delta_1^{*T}, 0^T)^T$, where δ_1^* is the b_n-dimensional vector containing nonzero components of δ_μ^*. Then, from Theorem 2, we have $\hat{b}_n = b_n$ and consequently,

$$\|\hat{\delta} - \delta_\mu^*\|_2^2 = \|\hat{\delta}_1 - \delta_1^*\|_2^2 = O_p(b_n \rho_{n1})$$

with a probability tending to one. It together with condition (A) implies that $(\hat{\delta} - \delta_\mu^*)^T \Omega^* (\hat{\delta} - \delta_\mu^*) = O_p(b_n \rho_{n1})$. Thus, $(\hat{\delta} - \delta_\mu^*)^T \Omega^* \delta_\mu^* \le \Delta_{p_n} O_p(b_n^{1/2} \rho_{n1}^{1/2})$ and

$$\begin{aligned} \hat{\delta}^T \Omega^* \hat{\delta} &= (\hat{\delta} - \delta_\mu^*)^T \Omega^* (\hat{\delta} - \delta_\mu^*) + 2(\hat{\delta} - \delta_\mu^*)^T \Omega^* \delta_\mu^* + \Delta_{p_n}^2 \\ &= \Delta_{p_n}^2 \{1 + O_p(b_n^{1/2} \rho_{n1}^{1/2}/\Delta_{p_n})\}. \end{aligned}$$

Let $\tilde{\mu}_1 - \mu_1^* = (\gamma_1^T, \gamma_2^T)^T$, where γ_1 is a b_n-dimensional vector. Partition Ω^* into

$$\Omega^* = \begin{bmatrix} \Omega_{11}^* & \Omega_{12}^* \\ \Omega_{12}^{*T} & \Omega_{22}^* \end{bmatrix},$$

where Ω_{11}^* is a $b_n \times b_n$ matrix, and partition Σ^*, $\hat{\Omega}$ and $\hat{\Sigma}$ in the same way. Then,

$$\hat{\delta}^T \hat{\Omega}(\tilde{\mu}_1 - \mu_1^*) = \hat{\delta}_1^T \hat{\Omega}_{11}\gamma_1 + \hat{\delta}_1^T \hat{\Omega}_{12}\gamma_2,$$

with a probability tending to one. Further, by Cauchy-Schwarz inequality and the fact $\Omega_{11}^{*-1} \le \Sigma_{11}^*$, we have $(\hat{\delta}_1^T \hat{\Omega}_{11}\gamma_1)^2 \le (\hat{\delta}^T \hat{\Omega}\hat{\delta}) O_p(b_n/n)$ and $(\hat{\delta}_1^T \hat{\Omega}_{12}\gamma_2)^2 \le (\hat{\delta}^T \hat{\Omega}\hat{\delta})[\gamma_2^T \Omega_{12}^{*T} \Sigma_{11}^* \Omega_{12}^* \gamma_2 \{1 + O_p(\rho_{n2}^{1/2})\}]$. Note that all eigenvalues of sub-matrices of Ω^* and Σ^* are bounded under condition (A). Then, we have that

$$\begin{aligned} \mathrm{E}(\gamma_2^T \Omega_{12}^{*T} \Sigma_{11}^* \Omega_{12}^* \gamma_2) &\le \kappa_2 \mathrm{E}(\gamma_2^T \Omega_{12}^{*T} \Omega_{12}^* \gamma_2) \\ &\le \frac{\kappa_2^2}{n} \mathrm{tr}(\Omega_{12}^* \Omega_{12}^{*T}) \\ &\le \kappa_2^2 a_n/n. \end{aligned}$$

Therefore,

$$\frac{\hat\delta^T\hat\Omega(\tilde\mu_1-\mu_1^*)}{\sqrt{\hat\delta^T\hat\Omega\Sigma^*\hat\Omega\hat\delta}}=\frac{O_p(\sqrt{b_n/n})+O_p(\sqrt{a_n/n})}{\sqrt{1+O_p(\rho_{n2}^{1/2})}},$$

which also holds when $\tilde\mu_1-\mu_1^*$ is replaced by $\tilde\mu_2-\mu_2^*$ or $\tilde\delta-\delta_\mu^*$. Furthermore, $\hat\delta^T\hat\Omega\tilde\delta=\hat\delta^T\hat\Omega\hat\delta+\hat\delta^T\hat\Omega(\tilde\delta-\delta_\mu^*)+\hat\delta^T\hat\Omega(\delta_\mu^*-\hat\delta)$ and $\{\hat\delta^T\hat\Omega(\delta_\mu^*-\hat\delta)\}^2\le(\hat\delta^T\Omega^*\hat\delta)O_p(b_n\rho_{n1})$. Therefore,

$$\begin{aligned}\frac{(-1)^k\hat\delta^T\hat\Omega(\mu_k^*-\tilde\mu_k)-\hat\delta^T\hat\Omega\tilde\delta/2}{\sqrt{\hat\delta^T\hat\Omega\Sigma^*\hat\Omega\hat\delta}}&=\frac{O_p(\sqrt{b_n/n})+O_p(\sqrt{a_n/n})+O_P(\sqrt{b_n\rho_{n1}})}{\sqrt{1+O_p(\rho_{n2}^{1/2})}}\\&\quad-\frac{\Delta_{p_n}\sqrt{1+O_p(b_n^{1/2}\rho_{n1}^{1/2}/\Delta_{p_n})}}{2\sqrt{1+O_p(\rho_{n2}^{1/2})}}\\&=-\{1+O_p(c_n)\}\Delta_{p_n}/2,\end{aligned}$$

which implies the result in (i).

(ii) Let ϕ be the density of Φ. Then, by the result in (i),

$$R_n-R_{\text{OPT}}=\phi(\nu_n)O_p(c_n),$$

where ν_n is between $-\Delta_{p_n}/2$ and $-\{1+O_p(c_n)\}\Delta_{p_n}/2$. Since Δ_{p_n} is bounded, $\phi(\nu_n)$ is bounded by a constant and R_{OPT} is bounded away from 0. Hence, the proposed method is asymptotically optimal and $R_n/R_{\text{OPT}}-1=O_p(c_n)$.

(iii) When $\Delta_{p_n}\to\infty$, $R_{\text{OPT}}\to 0$ and by the result in (i), $R_n\xrightarrow{P}0$, we have $R_n-R_{\text{OPT}}\xrightarrow{P}0$.

(iv) If $\Delta_{p_n}\to\infty$ and $c_n\Delta_{p_n}^2\to 0$, then, by Lemma 1 in Shao et al. (2011), we have $R_n/R_{\text{OPT}}\xrightarrow{P}1$. □

2. FIGURES FOR THE KIDNEY TRANSPLANT REJECTION AND TISSUE INJURY

Figure 1 summarizes the classification accuracy using boxplots for the proposed covariance-enhanced discriminant analysis, fusion-regularized linear discriminant analysis (Guo, 2010), doubly l_1-penalized linear discriminant analysis, sparse discriminant analysis (Clemmensen et al., 2011) and l_1-penalized linear discriminant analysis (Witten & Tibshirani, 2011). Figure 2 presents the heatmap of the estimated centroids for the 19 most informative genes selected in the kidney transplant rejection and tissue injury data set.

REFERENCES

BICKEL, P. J. & LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

CLEMMENSEN, L., HASTIE, T. J., WITTEN, D. M. & ERSBØLL, B. (2011). Sparse discriminant analysis. *Technometrics* **53**, 406–413.

GUO, J. (2010). Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis. *Biostatistics* **11**, 599–608.

LEDOUX, M. & TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Berlin Heidelberg, 1st ed.

RINALDO, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist.* **37**, 2922–2952.

SHAO, J., WANG, Y., DENG, X. & WANG, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39**, 1241–1265.

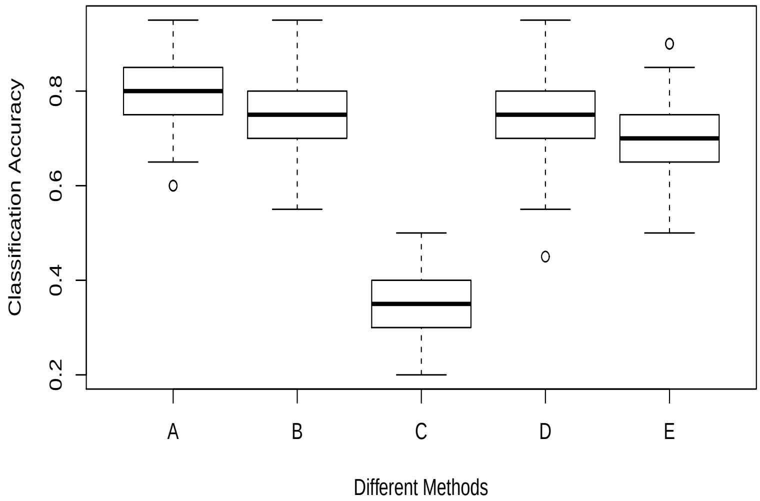

Fig. 1. Classification accuracies of the five methods on the kidney transplant rejection and tissue injury data set. The procedures from A to E are the proposed covariance-enhanced discriminant analysis, fusion-regularized linear discriminant analysis (Guo, 2010), doubly l_1-penalized linear discriminant analysis, the sparse discriminant (Clemmensen et al., 2011) and l_1-penalized linear discriminant analysis (Witten & Tibshirani, 2011).

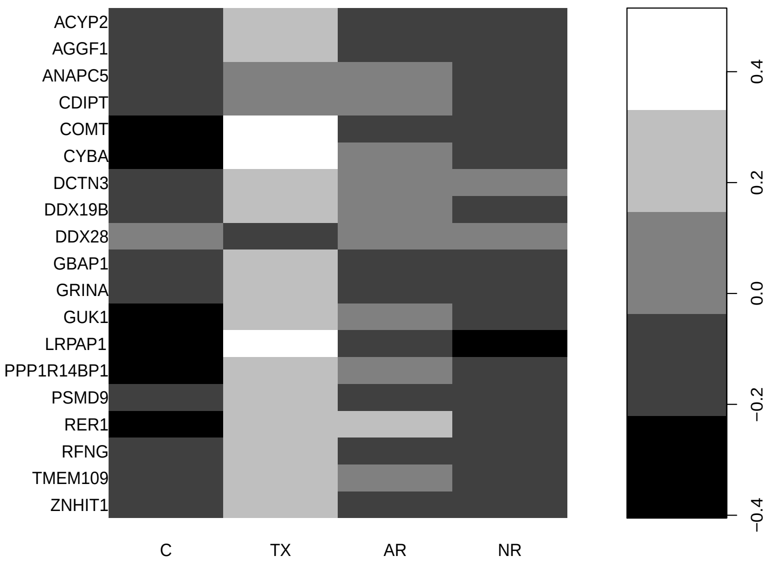

Fig. 2. The heatmap of the estimated centroids for the 19 most informative genes selected in the kidney transplant rejection and tissue injury data set. Rows correspond to genes and columns to classes. The right is the color key.

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

[*Received April* 2012. *Revised September* 2012]