Biometrika (2018), **xx**, x, *pp*. 1–11
Printed in Great Britain

Supplementary material for "Integrative linear discriminant analysis with guaranteed error rate improvement"

BY QUEFENG LI

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.

quefeng@email.unc.edu

AND LEXIN LI

Division of Biostatistics, University of California at Berkeley, Berkeley, California 94720, U.S.A.

lexinli@berkeley.edu.

SUMMARY

This supplementary file includes proofs of the proposition and theorems in the main document, the supporting technical lemmas and their proofs, and additional simulation results.

1. PROOF OF PROPOSITION 1

Proof. We partition $\delta = (\delta_{1:d-1}^{\mathrm{T}} \ \delta_d)^{\mathrm{T}}$, where $\delta_{1:d-1}$ and δ_d are the first $d-1$ and the dth coordinates of δ. The covariance matrix Σ is partitioned accordingly. Letting $\alpha = (\sigma_{dd} - \sigma_{1:d-1,d}^{\mathrm{T}} \Sigma_{1:d-1,1:d-1}^{-1} \sigma_{1:d-1,d})^{-1}$, we have

$$
\begin{aligned}
&\delta^{\mathrm{T}}\Sigma^{-1}\delta \\
&= \left(\delta_{1:d-1}^{\mathrm{T}} \ \delta_d\right) \begin{pmatrix} \Sigma_{1:d-1,1:d-1} & \sigma_{1:d-1,d} \\ \sigma_{1:d-1,d}^{\mathrm{T}} & \sigma_{dd} \end{pmatrix}^{-1} \begin{pmatrix} \delta_{1:d-1} \\ \delta_d \end{pmatrix} \\
&= \left(\delta_{1:d-1}^{\mathrm{T}} \ \delta_d\right) \begin{Bmatrix} (\Sigma_{1:d-1,1:d-1} - \sigma_{dd}^{-1}\sigma_{1:d-1,d}\sigma_{1:d-1,d}^{\mathrm{T}})^{-1} & -\alpha\Sigma_{1:d-1,1:d-1}^{-1}\sigma_{1:d-1,d} \\ -\alpha\sigma_{1:d-1,d}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1} & \alpha \end{Bmatrix} \begin{pmatrix} \delta_{1:d-1} \\ \delta_d \end{pmatrix}.
\end{aligned} \tag{S1}
$$

By the Sherman–Morrison–Woodbury formula,

$$
(\Sigma_{1:d-1,1:d-1} - \sigma_{dd}^{-1}\sigma_{1:d-1,d}\sigma_{1:d-1,d}^{\mathrm{T}})^{-1} = \Sigma_{1:d-1,1:d-1}^{-1} + \alpha\Sigma_{1:d-1,1:d-1}^{-1}\sigma_{1:d-1,d}\sigma_{1:d-1,d}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1}.
$$

Then we have

$$
\begin{aligned}
(S1) &= \delta_{1:d-1}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1}\delta_{1:d-1} + \alpha\delta_{1:d-1}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1}\sigma_{1:d-1,d}\sigma_{1:d-1,d}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1}\delta_{1:d-1} \\
&\quad - 2\alpha\delta_{1:d-1}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1}\sigma_{1:d-1,d}\delta_d + \alpha\delta_d^2 \\
&= \delta_{1:d-1}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1}\delta_{1:d-1} + \alpha(\delta_d - \sigma_{1:d-1,d}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1}\delta_{1:d-1})^2 \\
&\geq \delta_{1:d-1}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1}\delta_{1:d-1},
\end{aligned}
$$

where the equality holds if and only if $\sigma_{1:d-1,d}^{\mathrm{T}}\Sigma_{1:d-1,1:d-1}^{-1}\delta_{1:d-1} = \delta_d$, which happens only in a set with Lebesgue measure zero. The last inequality is due to the fact that $\alpha > 0$, since Σ is positive definite. $\square$

2. PROOF OF THEOREM 1

Proof. We use similar technique as in Cai & Liu (2012). First, we bound $|(\widehat{\mu} - \mu_1)^{\mathrm{T}}\widehat{\beta}/(\widehat{\beta}\Sigma\widehat{\beta})^{1/2} - {\Delta_d}^{1/2}/2|$ and $|(\mu_0 - \widehat{\mu})^{\mathrm{T}}\widehat{\beta}/(\widehat{\beta}\Sigma\widehat{\beta})^{1/2} - {\Delta_d}^{1/2}/2|$. Then, we utilize a result of the tail probability of normal distribution to complete the proof.

Letting $\Omega = \Sigma^{-1}$, we have

$$(\mu_1 - \widehat{\mu})^{\mathrm{T}}\widehat{\beta} + \frac{\Delta_d}{2} = (\mu - \widehat{\mu})^{\mathrm{T}}\widehat{\beta} + (\mu_1 - \mu)^{\mathrm{T}}\widehat{\beta} + \frac{\Delta_d}{2} = (\mu - \widehat{\mu})^{\mathrm{T}}\widehat{\beta} + \frac{1}{2}\delta^{\mathrm{T}}(\Omega\delta - \widehat{\beta}).$$

Therefore,

$$\left|(\mu_1 - \widehat{\mu})^{\mathrm{T}}\widehat{\beta} + \frac{\Delta_d}{2}\right| \le |(\mu - \widehat{\mu})^{\mathrm{T}}\widehat{\beta}| + \frac{1}{2}|\delta^{\mathrm{T}}(\Omega\delta - \widehat{\beta})|. \tag{S2}$$

By the normality assumption and the standard concentration result (Bühlmann & Van De Geer, 2011), it holds, with probability at least $1 - C_1 d^{-C_2}$ that,

$$\|\widehat{\mu} - \mu\|_\infty \lesssim \{(\log d)/n\}^{1/2}. \tag{S3}$$

It then follows from Lemma 1 that $\|\widehat{\mu} - \mu\|_{\infty,2} \lesssim \{M(\log d)/n\}^{1/2}$.

Next, we show that, with a high probability, $\|\widehat{\beta} - \beta^*\|_{1,G} = o(1)$. By Lemma 1 and (S3), we have

$$|(\mu - \widehat{\mu})^{\mathrm{T}}\widehat{\beta}| \lesssim \|\beta^*\|_{1,G}\{M(\log d)/n\}^{1/2}. \tag{S4}$$

To prove $\|\widehat{\beta} - \beta^*\|_{1,G} = o_P(1)$, we use the results from Negahban et al. (2012). First, using a similar argument as in (S3), we have with probability at least $1 - C_1 d^{-C_2}$ that $\|\widehat{\delta} - \delta\|_{\infty,2} \lesssim \{M(\log d)/n\}^{1/2}$. Together with Lemma 2, it implies that, $\|\widehat{\Sigma}\beta^* - \widehat{\delta}\|_{\infty,2} \lesssim \{M\Delta_d(\log d)/n\}^{1/2}$. Then, by choosing $\lambda_n = C_0\{M\Delta_d(\log d)/n\}^{1/2}$ for some large constant C_0, Corollary 1 of Negahban et al. (2012) implies that, with probability at least $1 - C_1 d^{-C_2}$, $\|\widehat{\beta} - \beta^*\|_{1,G} \lesssim \|\beta^*\|_{1,G}\{M\Delta_d(\log d)/n\}^{1/2} = o(1)$, where the last equality is ensured by Condition 3.

On the other hand, by Lemma 3, $\|\widehat{\delta} - \widehat{\Sigma}\widehat{\beta}\|_{\infty,2} \le \lambda_n M^{1/2} \asymp \nu_n$. It further implies

$$\|\delta - \widehat{\Sigma}\widehat{\beta}\|_{\infty,2} \lesssim \nu_n. \tag{S5}$$

Then,

$$\begin{aligned}
|\delta^{\mathrm{T}}(\Omega\delta - \widehat{\beta})| &\le |\delta^{\mathrm{T}}\Omega(\delta - \widehat{\Sigma}\widehat{\beta})| + |\delta^{\mathrm{T}}\Omega\widehat{\Sigma}\widehat{\beta} - \delta^{\mathrm{T}}\widehat{\beta}| \\
&\le \|\beta^*\|_{1,G}\|\widehat{\Sigma}\widehat{\beta} - \delta\|_{\infty,2} + \|\widehat{\Sigma}\Omega\delta - \delta\|_{\infty,2}\|\widehat{\beta}\|_{1,G} \\
&\lesssim \|\beta^*\|_{1,G}\|\widehat{\Sigma}\widehat{\beta} - \delta\|_{\infty,2} + \|\widehat{\Sigma}\Omega\delta - \delta\|_{\infty,2}\|\beta^*\|_{1,G} \\
&\lesssim \nu_n\|\beta^*\|_{1,G},
\end{aligned} \tag{S6}$$

where the last inequality follows from (S5) and Lemma 2. Then by (S2), (S4) and (S6), we have

$$\left|(\mu_1 - \widehat{\mu})^{\mathrm{T}}\widehat{\beta} + \frac{\Delta_d}{2}\right| \lesssim \nu_n\|\beta^*\|_{1,G}. \tag{S7}$$

Next we bound $\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta} - \widehat{\delta}^{\mathrm{T}}\Omega\delta$ by

$$|\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta} - \delta^{\mathrm{T}}\Omega\delta| \leq |\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta} - \widehat{\beta}^{\mathrm{T}}\delta| + |\widehat{\beta}^{\mathrm{T}}\delta - \delta^{\mathrm{T}}\Omega\delta|. \tag{S8}$$

For the first term, we have

$$\begin{aligned}|\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta} - \widehat{\beta}^{\mathrm{T}}\delta| &\leq \|\widehat{\beta}\|_{1,G}\|\Sigma\widehat{\beta} - \delta\|_{\infty,2} \lesssim \|\beta^*\|_{1,G}\{\|(\widehat{\Sigma} - \Sigma)\widehat{\beta}\|_{\infty,2} + \|\widehat{\Sigma}\widehat{\beta} - \delta\|_{\infty,2}\} \\ &\lesssim \|\beta^*\|_{1,G}(\|\widehat{\Sigma} - \Sigma\|_{\infty,2}\|\widehat{\beta}\|_{1,G} + \nu_n) \lesssim \|\beta^*\|_{1,G}(M\|\widehat{\Sigma} - \Sigma\|_{\max}\|\beta^*\|_{1,G} + \nu_n) \\ &\lesssim \varphi_n\|\beta^*\|^2_{1,G} + \nu_n\|\beta^*\|_{1,G},\end{aligned}$$

where the third inequality follows from Lemma 1 and (S5), the fourth inequality follows from Lemma 1 and $\varphi_n = M\{(\log d)/n\}^{1/2}$. Together with (S6) and (S8), it implies that

$$|\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta} - \delta^{\mathrm{T}}\Omega\delta| \lesssim \varphi_n\|\beta^*\|^2_{1,G} + \nu_n\|\beta^*\|_{1,G}.$$

Then we have

$$\begin{aligned}\left|(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{-1/2} - (\delta^{\mathrm{T}}\Sigma\delta)^{-1/2}\right| &\leq \frac{|\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta} - \delta^{\mathrm{T}}\Sigma\delta|}{(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{1/2}(\delta^{\mathrm{T}}\Sigma\delta)^{1/2}\{(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{1/2} + (\delta^{\mathrm{T}}\Sigma\delta)^{1/2}\}} \\ &\lesssim \Delta_d^{-3/2}(\varphi_n\|\beta^*\|^2_{1,G} + \nu_n\|\beta^*\|_{1,G}).\end{aligned} \tag{S9}$$

Denote $r_{1n} = (\widehat{\mu} - \mu_1)^{\mathrm{T}}\widehat{\beta}/(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{1/2}$. We have

$$|r_{1n} - {\Delta_d}^{1/2}/2| \leq |r_{1n} - (\Delta_d/2)(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{-1/2}| + |(\Delta_d/2)(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{-1/2} - {\Delta_d}^{1/2}/2|. \tag{S10}$$

For the first term on the right-hand side of (S10), it follows from (S7) that

$$|r_{1n} - (\Delta_d/2)(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{-1/2}| \leq |\{(\widehat{\mu} - \mu_1)^{\mathrm{T}}\widehat{\beta} - \Delta_d/2\}(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{-1/2}| \lesssim \nu_n\|\beta^*\|_{1,G}(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{-1/2}.$$

Since $\delta^{\mathrm{T}}\Omega\delta \geq c_0$,

$$\left|\frac{\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta}}{\delta^{\mathrm{T}}\Omega\delta} - 1\right| \lesssim |\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta} - \delta^{\mathrm{T}}\Omega\delta| = o(1).$$

Then $|r_{1n} - (\Delta_d/2)(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{-1/2}| \lesssim \nu_n\Delta_d^{-1/2}\|\beta^*\|_{1,G}$. For the second term on the right-hand side of (S10), it follows from (S9) that

$$\begin{aligned}|(\Delta_d/2)(\widehat{\beta}\Sigma\widehat{\beta})^{-1/2} - {\Delta_d}^{1/2}/2| &= (\Delta_d/2)|(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{-1/2} - \Delta_d^{-1/2}| \\ &\lesssim \Delta_d^{-1/2}(\varphi_n\|\beta^*\|^2_{1,G} + \nu_n\|\beta^*\|_{1,G}).\end{aligned}$$

Therefore,

$$\begin{aligned}|r_{1n} - {\Delta_d}^{1/2}/2| &\lesssim \nu_n\Delta_d^{-1/2}\|\beta^*\|_{1,G} + \Delta_d^{-1/2}(\varphi_n\|\beta^*\|^2_{1,G} + \nu_n\|\beta^*\|_{1,G}) \\ &\lesssim \Delta_d^{-1/2}(\varphi_n\|\beta^*\|^2_{1,G} + \nu_n\|\beta^*\|_{1,G}).\end{aligned}$$

Then we have

$$\left|\frac{r_{1n}}{{\Delta_d}^{1/2}/2} - 1\right| \lesssim \Delta_d^{-1}(\varphi_n\|\beta^*\|^2_{1,G} + \nu_n\|\beta^*\|_{1,G}) = o(1).$$

Letting $\pi_n = \nu_n\|\beta^*\|_{1,G} + \varphi_n\|\beta^*\|^2_{1,G}$, we have

$$\left|\frac{{\Delta_d}^{1/2}/2}{r_{1n}} - 1\right| \lesssim \Delta_d^{-1}\pi_n.$$

Using the fact that $\Phi(-x) \asymp x^{-1}\exp(-x^2/2)$, we have

$$\frac{\Phi(-r_{1n})}{\Phi(-\sqrt{\Delta_d/2})} \asymp \frac{\sqrt{\Delta_d/2}}{r_{1n}} \exp\left(-\frac{r_{1n}^2}{2} + \frac{\Delta_d}{8}\right) = \{1 + O(\Delta_d^{-1}\pi_n)\} \exp\left(-\frac{r_{1n}^2}{2} + \frac{\Delta_d}{8}\right). \tag{S11}$$

When Δ_d is bounded, (S11) and Condition 2 imply that

$$\frac{\Phi(-r_{1n})}{\Phi(-\sqrt{\Delta_d/2})} - 1 = O(\pi_n). \tag{S12}$$

When $\Delta_d \to \infty$, by the mean value theorem, $\exp(-r_{1n}^2/2 + \Delta_d/8) = 1 + O(-r_{1n}^2/2 + \Delta_d/8) = 1 + O\{\Delta_d(1 - 4r_{1n}^2\Delta_d^{-1})\} = 1 + O(\Delta_d)$. Therefore, in both cases (S12) holds. Similarly, we can show that

$$\frac{\Phi(-r_{2n})}{\Phi(-\sqrt{\Delta_d/2})} - 1 = O(\pi_n),$$

where $r_{2n} = (\mu_0 - \widehat{\mu})^{\mathrm{T}}\widehat{\beta}/(\widehat{\beta}^{\mathrm{T}}\Sigma\widehat{\beta})^{1/2}$. These two results imply that $R_n/R_d^* - 1 = O(\pi_n)$. This completes the proof. □

3. PROOF OF OF THEOREM 2

Proof. Recall that we allow p and d to grow with n. We first prove that $\limsup_{p\to\infty} R_2^*/R_1^* < 1$. This together with $R_{1n}/R_1^* \to 1$ and $R_{2n}/R_2^* \to 1$ in probability imply that

$$\limsup_{n\to\infty} \frac{R_{2n}}{R_{1n}} = \limsup_{n\to\infty} \frac{R_{2n}}{R_2^*} \times \frac{R_2^*}{R_1^*} \times \frac{R_1^*}{R_{1n}} < 1.$$

We use a standard result regarding the normal distribution, see e.g., equation (22) of Shao et al. (2011),

$$\frac{x}{1+x^2} e^{-x^2/2} \le \Phi(-x) \le \frac{1}{x} e^{-x^2/2}.$$

Then we have

$$\frac{R_2^*}{R_1^*} = \frac{\Phi(-\sqrt{\Delta_d/2})}{\Phi(-\sqrt{\Delta_p/2})} \le \frac{4+\Delta_p}{\sqrt{\Delta_d}\sqrt{\Delta_p}} \exp\left\{-\frac{1}{4}(\Delta_d - \Delta_p)\right\}. \tag{S13}$$

When $\Delta_p \to \infty$, by Condition 4,

$$\frac{4+\Delta_p}{\sqrt{\Delta_d}\sqrt{\Delta_p}} \le \frac{4+\Delta_p}{\Delta_p} \to 1, \;\; \exp\left\{-\frac{1}{4}(\Delta_d - \Delta_p)\right\} \le \exp\left(-\frac{1}{4}c_1\right) < 1.$$

Therefore, $\limsup_{p\to\infty} R_2^*/R_1^* < 1$. When $\Delta_p \le C$ for some $C > 0$ but $\Delta_d \to \infty$, it is clear from (S13) that $\limsup_{p\to\infty} R_2^*/R_1^* < 1$. When $\Delta_d \le C$, by the mean value theorem,

$$\begin{aligned}\Phi(-\sqrt{\Delta_d/2}) &= \Phi(-\sqrt{\Delta_p/2}) - \frac{1}{4\sqrt{\xi}}\phi(-\sqrt{\xi/2})(\Delta_d - \Delta_p)\\ &\le \Phi(-\sqrt{\Delta_p/2}) - \frac{1}{4\sqrt{C}}\phi(-\sqrt{C/2})(\Delta_d - \Delta_p),\end{aligned}$$

where $\Delta_p \le \xi \le \Delta_d$, and $\phi(x)$ is the standard normal density function. Therefore,

$$\frac{R_2^*}{R_1^*} \le 1 - \frac{\phi(-\sqrt{C/2})(\Delta_d - \Delta_p)}{4\sqrt{C}\Phi(-\sqrt{\Delta_p/2})} \le 1 - \frac{c_1\phi(-\sqrt{C/2})}{4\sqrt{C}\Phi(-c_0^{-1/2}/2)} < 1,$$

based on the fact that $\Delta_p \le \Delta_d \le C$ and Condition 4. This completes the proof. □

4. PROOF OF THEOREM 3

Proof. By the convex optimization theory, any vector $\beta \in \mathbb{R}^d$ satisfying the following Karush–Kuhn–Tucker conditions (Boyd & Vandenberghe, 2004) are the solution to problem (3)

$$(\widehat{\Sigma}\beta)_{j_m} - \widehat{\delta}_{j_m} + (1-\alpha)\lambda_n \mathrm{sgn}(\beta_{j_m}) + \alpha\lambda_n \frac{\beta_{j_m}}{\|\beta_{S_j}\|_2} = 0,\ j_m \in \mathcal{A}, \tag{S14}$$

$$|(\widehat{\Sigma}\beta)_{j_m} - \widehat{\delta}_{j_m}| < (1-\alpha)\lambda_n,\ j_m \in \mathcal{B}, \tag{S15}$$

$$|(\widehat{\Sigma}\beta)_{j_m} - \widehat{\delta}_{j_m}| < \lambda_n M^{-1/2},\ j_m \in \mathcal{C}, \tag{S16}$$

$$\lambda_{\min}(\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}) > 0. \tag{S17}$$

We prove the theorem through the following three steps. First, we show that there exists a solution $\widehat{\beta}_{\mathcal{A}} \in \mathbb{R}^s$ to equation (S14) within the neighborhood $\mathcal{N} = \{\beta : \|\beta - \beta^*_{\mathcal{A}}\|_\infty \le C\lambda_n\}$. Second, we show that $\widehat{\beta} = (\widehat{\beta}_{\mathcal{A}}, 0)^{\mathrm{T}}$ satisfies (S15) and (S16). Third, we check (S17). The inequality in (S16) further implies the Karush–Kuhn–Tucker condition $\|(\widehat{\Sigma}\beta)_{S_j} - \widehat{\delta}_{S_j}\|_2 < \lambda_n$, which is needed for the ℓ_2-component of the composite penalty we use.

First, we have

$$(\widehat{\Sigma}\beta)_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}} = \widehat{\Sigma}_{\mathcal{A}\mathcal{A}}(\beta_{\mathcal{A}} - \beta^*_{\mathcal{A}}) + \widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}}.$$

By (S19), we have with probability at least $1 - C_1 d^{-C_2}$ that

$$\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}}\|_\infty \le \|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \delta_{\mathcal{A}}\|_\infty + \|\delta_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}}\|_\infty \le C\{\Delta_d(\log d)/n\}^{1/2}. \tag{S18}$$

Define vectors $\tau \in \mathbb{R}^d$ and $\eta \in \mathbb{R}^d$ such that $\tau_{j_m} = \mathrm{sgn}(\beta_{j_m})$ and $\eta_{j_m} = \beta_{j_m}/\|\beta_{S_j}\|_2$ for $j_m \in \mathcal{A}$ and $\tau_{j_m} = \eta_{j_m} = 0$ for $j_m \in \mathcal{A}^c$. Let $f(\beta_{\mathcal{A}}) = \widehat{\Sigma}_{\mathcal{A}\mathcal{A}}(\beta_{\mathcal{A}} - \beta^*_{\mathcal{A}}) + \widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}} + (1-\alpha)\lambda_n\tau_{\mathcal{A}} + \alpha\lambda_n\eta_{\mathcal{A}}$ and $g(\beta_{\mathcal{A}}) = \widehat{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}} f(\beta_{\mathcal{A}}) = \beta_{\mathcal{A}} - \beta^*_{\mathcal{A}} - \widehat{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\{\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}} + (1-\alpha)\lambda_n\tau_{\mathcal{A}} + \alpha\lambda_n\eta_{\mathcal{A}}\}$. By Lemma 4, $\|\widehat{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty$ is bounded with probability at least $1 - C_1 d^{-C_2}$. Hence, by (S18) and the choice of λ_n, we have

$$\begin{aligned}
&\|\widehat{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\{\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}} + (1-\alpha)\lambda_n\tau_{\mathcal{A}} + \alpha\lambda_n\eta_{\mathcal{A}}\}\|_\infty \\
&\le \|\widehat{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty\{\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}}\|_\infty + (1-\alpha)\lambda_n + \alpha\lambda_n\} \\
&\le 2c_0\left[C\{\Delta_d(\log d)/n\}^{1/2} + \lambda_n\right] \\
&\lesssim \lambda_n.
\end{aligned}$$

Then, for a sufficiently large n, if $(\beta_{\mathcal{A}} - \beta^*_{\mathcal{A}})_{j_m} = C\lambda_n$, for some large constant $C > 0$,

$$\{g(\beta_{\mathcal{A}})\}_{j_m} \ge C\lambda_n - [\widehat{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\{\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}} + (1-\alpha)\lambda_n\tau_{\mathcal{A}} + \alpha\lambda_n\eta_{\mathcal{A}}\}]_{j_m} \ge 0.$$

If $(\beta_{\mathcal{A}} - \beta^*_{\mathcal{A}})_{j_m} = -C\lambda_n$,

$$\{g(\beta_{\mathcal{A}})\}_{j_m} \le -C\lambda_n + [\widehat{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\{\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}} + (1-\alpha)\lambda_n\tau_{\mathcal{A}} + \alpha\lambda_n\eta_{\mathcal{A}}\}]_{j_m} \le 0.$$

By the continuity of $g(\beta_{\mathcal{A}})$ and the Miranda's existence theorem (Vrahatis, 1989), $g(\beta_{\mathcal{A}}) = 0$ has a solution $\widehat{\beta}_{\mathcal{A}}$ in $\mathcal{N}$. Obviously, $f(\widehat{\beta}_{\mathcal{A}}) = 0$. Hence, $\widehat{\beta}_{\mathcal{A}}$ also solves (S14).

Second, we have

$$
\begin{aligned}
(\widehat{\Sigma}\beta)_{\mathcal{B}} - \widehat{\delta}_{\mathcal{B}} &= \widehat{\Sigma}_{\mathcal{B}\mathcal{A}}(\beta_{\mathcal{A}} - \beta^*_{\mathcal{A}}) + (\widehat{\Sigma}\beta^* - \widehat{\delta})_{\mathcal{B}} \\
&= \widehat{\Sigma}_{\mathcal{B}\mathcal{A}}\widehat{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\{\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}} + (1-\alpha)\lambda_n\tau_{\mathcal{A}} + \alpha\lambda_n\eta_{\mathcal{A}}\} + (\widehat{\Sigma}\beta^* - \widehat{\delta})_{\mathcal{B}}.
\end{aligned}
$$

Similarly as in (S18), $\|\widehat{\Sigma}\beta^* - \widehat{\delta}\|_\infty \le C\{\Delta_d(\log d)/n\}^{1/2}$ with probability at least $1 - C_1 d^{-C_2}$. By Lemma 4, $\|\widehat{\Sigma}_{\mathcal{B}\mathcal{A}}\widehat{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty \le (1-\alpha)(1-\epsilon/2) < 1-\alpha$ with probability at least $1 - C_1 d^{-C_2}$. Hence, with probability at least $1 - C_1 d^{-C_2}$,

$$
\begin{aligned}
\|(\widehat{\Sigma}\beta)_{\mathcal{B}} - \widehat{\delta}_{\mathcal{B}}\|_\infty &\le (1-\alpha)(1-\epsilon/2)(\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}\beta^*_{\mathcal{A}} - \widehat{\delta}_{\mathcal{A}}\|_\infty + \lambda_n) + \|(\widehat{\Sigma}\beta^* - \widehat{\delta})_{\mathcal{B}}\|_\infty \\
&\le (1-\alpha)(1-\epsilon/2)\left[C\{\Delta_d(\log d)/n\}^{1/2} + \lambda_n\right] + C\{\Delta_d(\log d)/n\}^{1/2} \\
&\le (1-\alpha)(1-\epsilon/2)\lambda_n + (2-\epsilon/2)C\{\Delta_d(\log d)/n\}^{1/2} \\
&< (1-\alpha)\lambda_n,
\end{aligned}
$$

since $\{\Delta_d(\log d)/n\}^{1/2} = o(\lambda_n)$ by the stated choice of λ_n. By an analogous proof, we can show that $\widehat{\beta}$ satisfies (S16).

Finally, (S17) follows from Lemma 4. This completes the proof. □

5. ADDITIONAL LEMMAS AND PROOFS

LEMMA 1. *For a matrix $A \in \mathbb{R}^{d\times d}$, vectors $a \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$, the following statements hold: $|a^{\mathrm{T}}x| \le \|a\|_{\infty,2}\|x\|_{1,G}$; $\|A\|_{\infty,2} \le M\|A\|_{\max}$; $\|x\|_{\infty,2} \le \sqrt{M}\|x\|_\infty$; and $\|Ax\|_{\infty,2} \le \|A\|_{\infty,2}\|x\|_{1,G}$.*

Proof. For the first statement, we have

$$
\begin{aligned}
|a^{\mathrm{T}}x| &= \left|\sum_{j=1}^{p}\left\{(1-\alpha)\sum_{m=1}^{M} a_{j_m}x_{j_m} + \alpha\sum_{m=1}^{M} a_{j_m}x_{j_m}\right\}\right| \\
&\le \sum_{j=1}^{p}\left\{(1-\alpha)\sum_{m=1}^{M}|a_{j_m}x_{j_m}| + \alpha\sum_{m=1}^{M}|a_{j_m}x_{j_m}|\right\} \\
&\le \sum_{j=1}^{p}\left\{(1-\alpha)\left(\max_{1\le m\le M}|a_{j_m}|\right)\|x_{S_j}\|_1 + \alpha\|a_{S_j}\|_2\|x_{S_j}\|_2\right\} \\
&\le \left(\max_{j,m}|a_{j_m}|\right)\sum_{j=1}^{p}(1-\alpha)\|x_{S_j}\|_1 + \left(\max_{1\le j\le p}\|a_{S_j}\|_2\right)\sum_{j=1}^{p}\alpha\|x_{S_j}\|_2 \\
&\le \left(\max_{1\le j\le p}\|a_{S_j}\|_2\right)\sum_{j=1}^{p}\left\{(1-\alpha)\|x_{S_j}\|_1 + \alpha\|x_{S_j}\|_2\right\} \\
&= \|a\|_{\infty,2}\|x\|_{1,G}.
\end{aligned}
$$

The second and third statements follow from some simple algebra.

For the last statement, let $\widetilde{a}_{j_m}$ denote the j_mth row of A. By (1), we have

$$
|\widetilde{a}^{\mathrm{T}}_{j_m}x| \le \|\widetilde{a}_{j_m}\|_{\infty,2}\|x\|_{1,G}.
$$

Then,

$$
\begin{aligned}
\|Ax\|_{\infty,2}^2 &= \max_{1\le j\le p}\sum_{m=1}^{M}(Ax)_{j_m}^2 \le \max_{1\le j\le p}\sum_{m=1}^{M}\|\widetilde{a}_{j_m}\|_{\infty,2}^2\|x\|_{1,G}^2 \le \left(\max_{1\le j\le p}\sum_{m=1}^{M}\|\widetilde{a}_{j_m}\|_{\infty,2}^2\right)\|x\|_{1,G}^2 \\
&= \|A\|_{\infty,2}^2\|x\|_{1,2}^2.
\end{aligned}
$$

LEMMA 2. *Under Condition* 1*, there exist positive constants* C*,* C_1 *and* C_2 *such that it holds with probability at least* $1 - C_1 d^{-C_2}$ *that*

$$
\|\widehat{\Sigma}\Omega\delta - \delta\|_{\infty,2} \le C\{M\Delta_d(\log d)/n\}^{1/2}.
$$

Proof. We use a similar argument as in Cai & Liu (2012). Denote the vectors $U_0 = (X \mid Y = 0) - \mu_0$, and $U_1 = (X \mid Y = 1) - \mu_1$. We have

$$
\begin{aligned}
\widehat{\Sigma} &= \frac{1}{n}\left(\sum_{Y_i=0} U_{i0}U_{i0}^{\mathrm{T}} + \sum_{Y_i=1} U_{i1}U_{i1}^{\mathrm{T}}\right) - \frac{n_0}{n}\bar{U}_0\bar{U}_0^{\mathrm{T}} - \frac{n_1}{n}\bar{U}_1\bar{U}_1^{\mathrm{T}} \\
&= \widetilde{\Sigma} - \frac{n_0}{n}\bar{U}_0\bar{U}_0^{\mathrm{T}} - \frac{n_1}{n}\bar{U}_1\bar{U}_1^{\mathrm{T}}.
\end{aligned}
$$

It suffices to prove the result with $\widehat{\Sigma}$ replaced by $\widetilde{\Sigma}$. To simplify the presentation, denote $Z_i = U_{i0}$ $(1 \le i \le n_0)$ and $Z_i = U_{i1}$ $(n_0 + 1 \le i \le n)$. Then,

$$
\widetilde{\Sigma}\Omega\delta - (\mu_0 - \mu_1) = (\widetilde{\Sigma} - \Sigma)\Omega\delta = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^{\mathrm{T}}\Omega\delta - E(Z_i Z_i^{\mathrm{T}}\Omega\delta).
$$

Denote $\xi_{ij} = Z_{ij}Z_i^{\mathrm{T}}\Omega\delta - E(Z_{ij}Z_i^{\mathrm{T}}\Omega\delta)$. With e_j being a vector with 1 for the jth coordinate and 0 elsewhere, we have

$$
\begin{aligned}
\mathrm{var}(\xi_{ij}) &= \mathrm{var}(e_j^{\mathrm{T}} Z Z^{\mathrm{T}}\Omega\delta) = \mathrm{var}(Z^{\mathrm{T}}\Omega\delta e_j^{\mathrm{T}} Z) = \frac{1}{2}\mathrm{tr}\{(\Omega\delta e_j^{\mathrm{T}} + e_j^{\mathrm{T}}\delta^{\mathrm{T}}\Omega)\Sigma(\Omega\delta e_j^{\mathrm{T}} + e_j^{\mathrm{T}}\delta^{\mathrm{T}}\Omega)\Sigma\} \\
&= \delta_j^2 + \sigma_{jj}\delta^{\mathrm{T}}\Omega\delta \lesssim \delta^{\mathrm{T}}\Omega\delta.
\end{aligned}
$$

Since $\{\xi_{ij}\}_{i=1}^{n}$ are independent sub-exponential random variables with mean 0, we have

$$
\mathrm{pr}\left\{\max_{1\le j\le d}\frac{1}{n}\left|\sum_{i=1}^{n}\xi_{ij}\right| \ge C\left(\frac{\Delta_d\log d}{n}\right)^{1/2}\right\} \le C_1 d^{-C_2}.
$$

Then, by Lemma 1 and the union bound, we have

$$
\begin{aligned}
&\mathrm{pr}\left\{\|\widetilde{\Sigma}\Omega\delta - (\mu_0 - \mu_1)\|_{\infty,2} \ge C\left(\frac{M\Delta_d\log d}{n}\right)^{1/2}\right\} \\
&\le \mathrm{pr}\left\{\|\widetilde{\Sigma}\Omega\delta - (\mu_0 - \mu_1)\|_{\infty} \ge C\left(\frac{\Delta_d\log d}{n}\right)^{1/2}\right\} \\
&\le C_1 d^{-C_2}.
\end{aligned}
\tag{S19}
$$

LEMMA 3. *Let* $\widehat{\beta}$ *be the solution of problem* (3)*. Then it holds that*

$$
\|\widehat{\Sigma}\widehat{\beta} - \widehat{\delta}\|_{\infty,2} \le \lambda_n M^{1/2}.
$$

Proof. Let $F(\beta) = \beta^{\mathrm{T}}\widehat{\Sigma}\beta/2 - \widehat{\delta}^{\mathrm{T}}\beta + \lambda_n \sum_{j=1}^{p} \|\beta_{S_j}\|_G$. Define

$$H(\beta,\gamma,\eta) = \frac{1}{2}\beta^{\mathrm{T}}\widehat{\Sigma}\beta - \widehat{\delta}^{\mathrm{T}}\beta + (1-\alpha)\lambda_n \sum_{j=1}^{p} \gamma_{S_j}^{\mathrm{T}}\beta_{S_j} + \alpha\lambda_n \sum_{j=1}^{p} \eta_{S_j}^{\mathrm{T}}\beta_{S_j}.$$

Then, we have

$$F(\beta) = \max_{\substack{\|\gamma\|_\infty \le 1 \\ \|\eta\|_{\infty,2}\le 1}} H(\beta,\gamma,\eta).$$

By the strong duality, $\widehat{\beta}$ also solves

$$\min_{\beta} F(\beta) = \min_{\beta} \max_{\substack{\|\gamma\|_\infty \le 1 \\ \|\eta\|_{\infty,2}\le 1}} H(\beta,\gamma,\eta) = \max_{\substack{\|\gamma\|_\infty \le 1 \\ \|\eta\|_{\infty,2}\le 1}} \min_{\beta} H(\beta,\gamma,\eta).$$

By the Karush–Kuhn–Tucker condition, we have $\widehat{\Sigma}\widehat{\beta} - \widehat{\delta} + (1-\alpha)\lambda_n\gamma + \alpha\lambda_n\eta = 0$. Since $\|\gamma\|_\infty \le 1$ and $\|\eta\|_{\infty,2} \le 1$, we have

$$\|\widehat{\Sigma}\widehat{\beta} - \widehat{\delta}\|_{\infty,2} \le (1-\alpha)\lambda_n\|\gamma\|_{\infty,2} + \alpha\lambda_n\|\eta\|_{\infty,2} \le (1-\alpha)\lambda_n\sqrt{M} + \alpha\lambda_n \le \lambda_n\sqrt{M}.$$

LEMMA 4. *Under Conditions* 1 *and* 5–8, *if* $s^2\{(\log d)/n\}^{1/2} = o(1)$, *there exist positive constants* C, C_1 *and* C_2 *such that, with probability at least* $1 - C_1 d^{-C_2}$, *we have* $\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \le 2c_0$; $\|\widehat{\Sigma}_{\mathcal{B}\mathcal{A}}\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \le (1-\alpha)(1-\epsilon/2)$; $\|\widehat{\Sigma}_{\mathcal{C}\mathcal{A}}\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \le (1-\epsilon/2)M^{-1/2}$; *and* $\lambda_{\min}(\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}) \ge c_0^{-1}/2$.

Proof. For the first statement, by the standard concentration inequality result, e.g., Equation (10) of Bickel & Levina (2008), there exist positive constants C, C_1 and C_2 such that, for any $1 \le i, j \le d$,

$$\mathrm{pr}\left[|\widehat{\sigma}_{ij} - \sigma_{ij}| > C\{(\log d)/n\}^{1/2}\right] \le C_1 d^{-(C_2+2)}.$$

By the union bound, we have

$$\begin{aligned}
&\mathrm{pr}\left[\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}} - \Sigma_{\mathcal{A}\mathcal{A}}\|_\infty > Cs\{(\log d)/n\}^{1/2}\right] = \mathrm{pr}\left[\max_{i\in\mathcal{A}} \sum_{j\in\mathcal{A}} |\widehat{\sigma}_{ij} - \sigma_{ij}| > Cs\{(\log d)/n\}^{1/2}\right] \\
&\le s\,\mathrm{pr}\left[\sum_{j\in\mathcal{A}} |\widehat{\sigma}_{ij} - \sigma_{ij}| > Cs\{(\log d)/n\}^{1/2}\right] \le s^2\mathrm{pr}\left[|\widehat{\sigma}_{ij} - \sigma_{ij}| > C\{(\log d)/n\}^{1/2}\right] \\
&\le C_1 s^2 d^{-(C_2+2)} \le C_1 d^{-C_2}.
\end{aligned} \tag{S20}$$

Then, with probability at least $1 - C_1 d^{-C_2}$, we have

$$\begin{aligned}
\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty &\le \|\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty + \|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}} - \Sigma_{\mathcal{A}\mathcal{A}}\|_\infty\|\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \\
&\le c_0 + c_0\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty Cs\{(\log d)/n\}^{1/2}.
\end{aligned}$$

Therefore, when n is sufficiently large,

$$\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \le \frac{c_0}{1 - Cc_0 s\{(\log d)/n\}^{1/2}} \le 2c_0.$$

For the second statement, we have

$$\widehat{\Sigma}_{\mathcal{B}\mathcal{A}}\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} = \widehat{\Sigma}_{\mathcal{B}\mathcal{A}}(\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} - \Sigma_{\mathcal{A}\mathcal{A}}^{-1}) + (\widehat{\Sigma}_{\mathcal{B}\mathcal{A}} - \Sigma_{\mathcal{B}\mathcal{A}})\Sigma_{\mathcal{A}\mathcal{A}}^{-1} + \Sigma_{\mathcal{B}\mathcal{A}}\Sigma_{\mathcal{A}\mathcal{A}}^{-1}.$$

Then,

$$\|\widehat{\Sigma}_{\mathcal{B}\mathcal{A}}(\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} - \Sigma_{\mathcal{A}\mathcal{A}}^{-1})\|_\infty \le (\|\Sigma_{\mathcal{B}\mathcal{A}}\|_\infty + \|\widehat{\Sigma}_{\mathcal{B}\mathcal{A}} - \Sigma_{\mathcal{B}\mathcal{A}}\|_\infty)\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty\|\widehat{\Sigma}_{\mathcal{A}\mathcal{A}} - \Sigma_{\mathcal{A}\mathcal{A}}\|_\infty\|\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty. \quad \text{(S21)}$$

By definition, $\|\Sigma_{\mathcal{B}\mathcal{A}}\|_\infty = \max_{i\in\mathcal{B}} \sum_{j\in\mathcal{A}} |\sigma_{ij}| \lesssim s$. Similarly as (S20), we have

$$\text{pr}\left[\|\widehat{\Sigma}_{\mathcal{B}\mathcal{A}} - \Sigma_{\mathcal{B}\mathcal{A}}\|_\infty > Cs\{(\log d)/n\}^{1/2}\right] \le C_1 d^{-C_2}.$$

By (S21) and Condition 6, with probability at least $1 - C_1 d^{-C_2}$, we have

$$\|\widehat{\Sigma}_{\mathcal{B}\mathcal{A}}(\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} - \Sigma_{\mathcal{A}\mathcal{A}}^{-1})\|_\infty \lesssim s^2\{(\log d)/n\}^{1/2}.$$

By a similar argument, $\|(\widehat{\Sigma}_{\mathcal{B}\mathcal{A}} - \Sigma_{\mathcal{B}\mathcal{A}})\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \lesssim s^2\{(\log d)/n\}^{1/2}$. When the sample size n is large enough, both upper bounds become arbitrarily small. Hence, $\|\widehat{\Sigma}_{\mathcal{B}\mathcal{A}}(\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1} - \Sigma_{\mathcal{A}\mathcal{A}}^{-1}) + (\widehat{\Sigma}_{\mathcal{B}\mathcal{A}} - \Sigma_{\mathcal{B}\mathcal{A}})\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\|_2 \le (1-\alpha)\epsilon/2$. Then the result follows from Condition 7.

For the third statement, it can be proved using similar arguments.

For the last statement, Condition 5 implies that $\lambda_{\min}(\Sigma_{\mathcal{A}\mathcal{A}}) \ge c_0^{-1}$. Then by a similar proof, we can show that $\lambda_{\min}(\widehat{\Sigma}_{\mathcal{A}\mathcal{A}}) \ge c_0^{-1}/2$ with probability at least $1 - C_1 d^{-C_2}$. □

6. Additional simulations

We consider two additional simulation examples to inspect the robustness of our method and its adaptivity to block missing values. The settings are similar as in Example A, except that we change the distribution of X or introduce missing values. In Example D, the data follow a heavy-tailed distribution. That is, $X \mid Y = 0 \sim t_3(0, \Sigma)$, and $X \mid Y = 1 \sim \mu_1 + t_3(0, \Sigma)$, where $t_3(0, \Sigma)$ is the multivariate t-distribution with 3 degrees of freedom and the scale parameter Σ. For this example, we add the robust integrative linear discriminant analysis into the comparison. In Example E, the data follow a normal distribution. However, one data type has probability 0·25 to be entirely missing. The missing of different types are assumed independent. For this example, we compare two ways to utilize the data as discussed in Section 6 of the main document, i.e., the effective way and the complete case analysis. Table S1 reports the average criteria, and standard errors in parentheses, all in percentages, over 100 data replications. In Example D, the robust integrative linear discriminant analysis further improves the performance of the non-robust counterpart. In Example E, the effective integrative linear discriminative analysis handles the missing data better than only using the complete data.

In addition, we conduct a simulation study with an increasing M. The setting is the same as in Example A, with $n = 50$, $p = 100$, except that we choose $\beta_{jm}^* = 0{\cdot}5$ $(j = 1, \ldots, 5; m = 1, \ldots, M)$ and the rest equal to zero. We use $M = 2, 4$, and 6. Table S2 reports the average classification error of our method and the corresponding Bayes error over 100 replications. It is observed that, as M increases, both errors decrease, meanwhile the difference between the two errors increases. This observation agrees with Theorem 1, since as M increases, the convergence rate of the integrative classifier relative to the Bayes error can become slower. This is essentially due to the fact that more unknown parameters need to be estimated, which in turn induces a larger estimation error. However, if the additional discriminative information brought by the extra variables exceeds the estimation error they bring, the error rate R_n is guaranteed to decrease, as we show in Theorem 2.

Table S1: Classification and variable selection accuracy (%)

	Example D							
	$n = 50, p = 100, \pi = 1$				$n = 50, p = 100, \pi = 0{\cdot}5$			
	iLDA-r	iLDA	m-vote	sLDA	iLDA-r	iLDA	m-vote	sLDA
error rate	8(4)	12(9)	15(5)	24(7)	12(8)	14(14)	26(10)	32(10)
sensitivity	92(11)	68(25)	54(23)	54(14)	79(22)	59(33)	18(23)	42(20)
specificity	74(22)	78(23)	99(1)	95(2)	66(33)	64(35)	99(1)	96(3)
	$n = 100, p = 200, \pi = 1$				$n = 100, p = 200, \pi = 0{\cdot}5$			
	iLDA-r	iLDA	m-vote	sLDA	iLDA-r	iLDA	m-vote	sLDA
error rate	14(5)	15(7)	18(4)	27(5)	19(8)	20(12)	31(9)	36(8)
sensitivity	97(6)	80(15)	72(21)	65(15)	88(15)	68(30)	24(25)	53(24)
specificity	92(11)	94(15)	100(0)	99(1)	90(16)	90(20)	100(0)	99(1)

	Example E							
	$n = 50, p = 100, \pi = 1$				$n = 50, p = 100, \pi = 0{\cdot}5$			
	iLDA-e	iLDA-c	m-vote	sLDA	iLDA-e	iLDA-c	m-vote	sLDA
error rate	22(9)	25(7)	30(6)	32(6)	36(10)	38(7)	43(7)	42(7)
sensitivity	66(20)	55(25)	62(21)	57(11)	56(35)	51(27)	21(23)	45(21)
specificity	91(27)	72(36)	100(0)	98(2)	81(37)	69(33)	100(0)	98(1)
	$n = 100, p = 200, \pi = 1$				$n = 100, p = 200, \pi = 0{\cdot}5$			
	iLDA-e	iLDA-c	m-vote	sLDA	iLDA-e	iLDA-c	m-vote	sLDA
error rate	21(8)	24(6)	30(5)	33(6)	37(10)	39(9)	43(8)	43(7)
sensitivity	61(21)	47(25)	57(25)	55(14)	53(31)	52(26)	21(22)	44(19)
specificity	98(14)	92(13)	100(0)	98(1)	85(35)	79(16)	100(0)	98(1)

iLDA, the integrative linear discriminant analysis classifier with the composite penalty; iLDA-r, the robust integrative linear discriminant analysis classifier; iLDA-e, the integrative linear discriminant analysis classifier that effectively using all the observations; iLDA-c, the integrative linear discriminant analysis classifier using the complete data only; sLDA, the linear discriminant analysis classifier applied to each individual type separately; m-vote, a majority vote based on the class assigned by sLDA.

REFERENCES

BICKEL, P. J. & LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–2604.

BOYD, S. & VANDENBERGHE, L. (2004). *Convex Optimization.* Cambridge University Press, Cambridge.

BÜHLMANN, P. & VAN DE GEER, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications.* Springer, New York.

CAI, T. & LIU, W. (2012). A direct estimation approach to sparse linear discriminant analysis. *J. Am. Statist. Assoc.* **106**, 1566–1577.

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. & YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27**, 538–557.

Table S2: Classification error (%) of the Bayes rule and our method as M increases

	$M = 2$	$M = 4$	$M = 6$
Bayes error	16	6	2
iLDA	21	13	11

iLDA, the integrative linear discriminant analysis classifier with the composite penalty.

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

Vrahatis, M. N. (1989). A short proof and a generalization of Miranda's existence theorem. *Proc. Am. Math. Soc.* **107**, 701–703.