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 \( \delta = (\delta_{1:d-1}^T \ \delta_d)^T \), where \( \delta_{1:d-1} \) and \( \delta_d \) are the first \( d - 1 \) and the \( d \)th coordinates of \( \delta \). The covariance matrix \( \Sigma \) is partitioned accordingly. Letting \( \alpha = (\sigma_{dd} - \Sigma_{1:d-1,d}^{-1} \Sigma_{1:d-1,1:d-1} \sigma_{1:d-1,d})^{-1} \), we have

\[
\delta^T \Sigma^{-1} \delta = (\delta_{1:d-1}^T \ \delta_d) \begin{pmatrix} \Sigma_{1:d-1,d-1} & \Sigma_{1:d-1,d} \\ \Sigma_{1:d-1,d}^T & \sigma_{dd} \end{pmatrix}^{-1} \begin{pmatrix} \delta_{1:d-1} \\ \delta_d \end{pmatrix} = (\delta_{1:d-1}^T \ \delta_d) \begin{pmatrix} \Sigma_{1:d-1,d-1} - \sigma_{dd}^{-1} \Sigma_{1:d-1,d} \sigma_{1:d-1,d}^T \\ -\alpha \sigma_{dd}^{-1} \Sigma_{1:d-1,d} \sigma_{1:d-1,d}^T \alpha \end{pmatrix}^{-1} \begin{pmatrix} \delta_{1:d-1} \\ \delta_d \end{pmatrix}.
\]

By the Sherman–Morrison–Woodbury formula,

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

Then we have

\[
(S1) = \delta_{1:d-1}^T \Sigma_{1:d-1,d-1}^{-1} \delta_{1:d-1} + \alpha \delta_{1:d-1}^T \Sigma_{1:d-1,d-1}^{-1} \Sigma_{1:d-1,d} \sigma_{1:d-1,d}^T \Sigma_{1:d-1,d}^{-1} \delta_{1:d-1} - 2\alpha \delta_{1:d-1}^T \Sigma_{1:d-1,d-1}^{-1} \sigma_{1:d-1,d} \delta_d + \alpha \delta_d^2
\]

\[
\geq \delta_{1:d-1}^T \Sigma_{1:d-1,d-1}^{-1} \delta_{1:d-1} + \alpha \delta_d^2 - \alpha \delta_{1:d-1}^T \Sigma_{1:d-1,d-1}^{-1} \Sigma_{1:d-1,d} \delta_{1:d-1} \delta_{1:d-1}^T \Sigma_{1:d-1,d}^{-1} \delta_{1:d-1}.
\]
where the equality holds if and only if $\sigma_{1:d-1,d}^{-1}\Sigma_{1:d-1,1:d-1}\delta_{1:d-1} = \delta_d$, which happens only in a set with Lebesgue measure zero. The last inequality is due to the fact that $\alpha > 0$, since $\Sigma$ is positive definite.

\[ \Box \]

2. PROOF OF THEOREM 1

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

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

\[
(\mu_1 - \hat{\mu})^T\hat{\beta} + \frac{\Delta_d}{2} = (\mu - \hat{\mu})^T\hat{\beta} + (\mu_1 - \mu)^T\hat{\beta} + \frac{\Delta_d}{2} = (\mu - \hat{\mu})^T\hat{\beta} + \frac{1}{2}\delta^T(\Omega\delta - \hat{\beta}).
\]

Therefore,

\[
\left| (\mu_1 - \hat{\mu})^T\hat{\beta} + \frac{\Delta_d}{2} \right| \leq |(\mu - \hat{\mu})^T\hat{\beta}| + \frac{1}{2}|\delta^T(\Omega\delta - \hat{\beta})|.
\]

(S2)

By the normality assumption and the standard concentration result (Bühlmann & Van De Geer, 2011), it holds, with probability at least $1 - C_1d^{-C_2}$ that,

\[
\|\hat{\mu} - \mu\|_\infty \lesssim \{(\log d)/n\}^{1/2}.
\]

(S3)

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

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

\[
|\hat{\beta} - \beta^*| \lesssim \|\beta^*\|_{1,G}\{M(\log d)/n\}^{1/2}.
\]

(S4)

To prove $\|\hat{\beta} - \beta^*\|_{1,G} = o_P(1)$, we use the results from Negahban et al. (2012). First, using a similar argument as in (S3), we have with probability at least $1 - C_1d^{-C_2}$ that $\|\hat{\beta} - \beta\|_{\infty,2} \lesssim \{M(\log d)/n\}^{1/2}$. Together with Lemma 2, it implies that, $\|\hat{\Sigma}\beta^* - \hat{\beta}\|_{\infty,2} \lesssim \{M\Delta_d(\log d)/n\}^{1/2}$. 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 $1 - C_1d^{-C_2}$, $\|\hat{\beta} - \beta^*\|_{1,G} \lesssim \|\beta^*\|_{1,G}\{M\Delta_d(\log d)/n\}^{1/2} = o(1)$, where the last equality is ensured by Condition 3.

On the other hand, by Lemma 3, $\|\hat{\delta} - \hat{\Sigma}\beta\|_{\infty,2} \lesssim \lambda_nM^{1/2} \simeq \nu_n$. It further implies

\[
\|\delta - \hat{\Sigma}\beta\|_{\infty,2} \lesssim \nu_n.
\]

(S5)

Then,

\[
|\delta^T(\Omega\delta - \hat{\beta})| \leq |\delta^T\Omega(\delta - \hat{\Sigma}\beta)| + |\delta^T\Omega\hat{\Sigma}\beta - \delta^T| \\
\leq \|\beta^*\|_{1,G}\|\hat{\Sigma}\beta - \delta\|_{\infty,2} + \|\hat{\Sigma}\Omega\delta - \delta\|_{\infty,2}\|\beta^*\|_{1,G} \\
\lesssim \|\beta^*\|_{1,G}\|\hat{\Sigma}\beta - \delta\|_{\infty,2} + \|\hat{\Sigma}\Omega\delta - \delta\|_{\infty,2}\|\beta^*\|_{1,G} \\
\lesssim \nu_n\|\beta^*\|_{1,G}.
\]

(S6)

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

\[
\left| (\mu_1 - \hat{\mu})^T\hat{\beta} + \frac{\Delta_d}{2} \right| \lesssim \nu_n\|\beta^*\|_{1,G}.
\]

(S7)
Next we bound $\hat{\beta}^T \Sigma \hat{\beta} - \delta^T \Omega \delta$ by

$$\|\hat{\beta}^T \Sigma \hat{\beta} - \delta^T \Omega \delta\| \leq \|\hat{\beta}^T \Sigma \hat{\beta} - \hat{\beta}^T \delta\| + \|\hat{\beta}^T \delta - \delta^T \Omega \delta\|. \quad (S8)$$

For the first term, we have

$$\|\hat{\beta}^T \Sigma \hat{\beta} - \delta^T \Omega \delta\| \leq \|\beta\|_{1,G} \|\Sigma \hat{\beta} - \delta\|_{\infty,2} \leq \|\beta\|_{1,G} \{(\|\Sigma - \Sigma\|_{\infty,2} + \|\Sigma \hat{\beta} - \delta\|_{\infty,2}\}
\leq \|\beta\|_{1,G} (\|\Sigma - \Sigma\|_{\infty,2} \|\hat{\beta}\|_{1,G} + \nu_n) \approx \|\beta\|_{1,G} (M \|\Sigma - \Sigma\|_{\max} \|\beta\|_{1,G} + \nu_n)
\leq \varphi_n \|\beta\|_{1,G}^2 + \nu_n \|\beta\|_{1,G},$$

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

$$\|\hat{\beta}^T \Sigma \hat{\beta} - \delta^T \Omega \delta\| \leq \varphi_n \|\beta\|_{1,G}^2 + \nu_n \|\beta\|_{1,G}.$$

Then we have

$$\left|\left(\hat{\beta}^T \Sigma \hat{\beta} - \delta^T \Omega \delta\right)^{-1/2} - \left(\delta^T \Omega \delta\right)^{-1/2}\right| \leq \frac{\|\hat{\beta}^T \Sigma \hat{\beta} - \delta^T \Omega \delta\|}{\left(\hat{\beta}^T \Sigma \hat{\beta} - \delta^T \Omega \delta\right)^{1/2} \left(\delta^T \Omega \delta\right)^{1/2} \left(\|\delta^T \Omega \delta\|\right)^{1/2} \left(\|\hat{\beta}^T \Sigma \hat{\beta}\|\right)^{1/2} \left(\|\beta\|_{1,G}\right)^{1/2}}$$

$$\leq \Delta_d^{-3/2} (\varphi_n \|\beta\|_{1,G}^2 + \nu_n \|\beta\|_{1,G}). \quad (S9)$$

Denote $r_1 = (\hat{\mu} - \mu_1)^T \hat{\beta} / (\hat{\beta}^T \Sigma \hat{\beta})^{1/2}$. We have

$$|r_1 - \Delta_d^{1/2}/2| \leq |r_1 - (\Delta_d/2)(\hat{\beta}^T \Sigma \hat{\beta})^{-1/2}| + |(\Delta_d/2)(\hat{\beta}^T \Sigma \hat{\beta})^{-1/2} - \Delta_d^{1/2}/2|. \quad (S10)$$

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

$$|r_1 - (\Delta_d/2)(\hat{\beta}^T \Sigma \hat{\beta})^{-1/2}| \leq |\{(\hat{\mu} - \mu_1)^T \hat{\beta} - \Delta_d/2\}(\hat{\beta}^T \Sigma \hat{\beta})^{-1/2}| \leq \nu_n \|\beta\|_{1,G}(\hat{\beta}^T \Sigma \hat{\beta})^{-1/2}.$$

Since $\delta^T \Omega \delta \geq c_0$,

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

Then $|r_1 - (\Delta_d/2)(\hat{\beta}^T \Sigma \hat{\beta})^{-1/2}| \leq \nu_n \Delta_d^{-1/2} \|\beta\|_{1,G}$. For the second term on the right-hand side of (S10), it follows from (S9) that

$$|(\Delta_d/2)(\hat{\beta}^T \Sigma \hat{\beta})^{-1/2} - \Delta_d^{1/2}/2| = (\Delta_d/2)(\hat{\beta}^T \Sigma \hat{\beta})^{-1/2} - \Delta_d^{-1/2}| \leq \Delta_d^{-1/2} (\varphi_n \|\beta\|_{1,G}^2 + \nu_n \|\beta\|_{1,G}).$$

Therefore,

$$|r_1 - \Delta_d^{1/2}/2| \leq \nu_n \Delta_d^{-1/2} \|\beta\|_{1,G} + \Delta_d^{-1/2} (\varphi_n \|\beta\|_{1,G}^2 + \nu_n \|\beta\|_{1,G})$$

$$\leq \Delta_d^{-1/2} (\varphi_n \|\beta\|_{1,G}^2 + \nu_n \|\beta\|_{1,G}).$$

Then we have

$$\left|\frac{r_1}{\Delta_d^{1/2}/2} - 1\right| \leq (\varphi_n \|\beta\|_{1,G}^2 + \nu_n \|\beta\|_{1,G}) = o(1).$$

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

$$\left|\frac{\Delta_d^{1/2}/2}{r_1} - 1\right| \leq \Delta_d^{-1} \pi_n.$$
Using the fact that $\Phi(-x) \approx x^{-1} \exp(-x^2/2)$, we have

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

(S11)

When $\Delta_d$ is bounded, (S11) and Condition 2 imply that

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

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

3. PROOF 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

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

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

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

Then we have

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

(S13)

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

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

Therefore, $\limsup_{p \to \infty} R_2^*/R_1^* < 1$. When $\Delta_p \leq C$ 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 $\Delta_d \leq C$, by the mean value theorem,

$$\Phi(-\sqrt{\Delta_d/2}) = \Phi(-\sqrt{\Delta_p/2}) - \frac{1}{4\sqrt{\xi}} \phi(-\sqrt{\xi/2})(\Delta_d - \Delta_p)$$

$$\leq \Phi(-\sqrt{\Delta_p/2}) - \frac{1}{4\sqrt{C}} \phi(-\sqrt{C/2})(\Delta_d - \Delta_p),$$

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

$$\frac{R_2^*}{R_1^*} \leq 1 - \frac{\phi(-\sqrt{C/2})(\Delta_d - \Delta_p)}{4\sqrt{C}(\Delta_d - \Delta_p)} \leq 1 - \frac{c_1 \phi(-\sqrt{C/2})}{4\sqrt{C} \Phi(-c_1^{-1/2}/2)} < 1,$$
Integrative Linear Discriminant Analysis

based on the fact that \( \Delta_p \leq \Delta_d \leq C \) and Condition 4. This completes the proof. \( \square \)

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)

\[
(\bar{\Sigma} \beta)_{j_m} - \hat{\beta}_{j_m} + (1 - \alpha)\lambda_n \text{sgn}(\beta_{j_m}) + \alpha \lambda_p \frac{\beta_{j_m}}{\|\beta_{S_j}\|_2} = 0, \quad j_m \in A, \tag{S14}
\]

\[
|\bar{\Sigma} \beta_{j_m} - \hat{\beta}_{j_m}| < (1 - \alpha)\lambda_n, \quad j_m \in B, \tag{S15}
\]

\[
|\bar{\Sigma} \beta_{j_m} - \hat{\beta}_{j_m}| < \lambda_n M^{-1/2}, \quad j_m \in C, \tag{S16}
\]

\[
\lambda_{\min}(\hat{\Sigma}_{AA}) > 0. \tag{S17}
\]

We prove the theorem through the following three steps. First, we show that there exists a solution \( \hat{\beta}_A \in \mathbb{R}^n \) to equation (S14) within the neighborhood \( \mathcal{N} = \{ \beta : \|\beta - \hat{\beta}_A\|_\infty \leq C\lambda_n \} \). Second, we show that \( \hat{\beta} = (\hat{\beta}_A, 0)^T \) satisfies (S15) and (S16). Third, we check (S17). The inequality in (S16) further implies the Karush–Kuhn–Tucker condition \( \|\bar{\Sigma} \beta_{S_j} - \hat{\beta}_{S_j}\|_2 < \lambda_n \), which is needed for the \( \ell_2 \)-component of the composite penalty we use.

First, we have

\[
(\bar{\Sigma} \beta)_A - \delta_A = \hat{\Sigma}_{AA}(\beta_A - \beta_A^*) + \hat{\Sigma}_{AA}\beta_A^* - \delta_A.
\]

By (S19), we have with probability at least \( 1 - C_1d^{-C_2} \) that

\[
\|\hat{\Sigma}_{AA}\beta_A^* - \delta_A\|_\infty \leq \|\hat{\Sigma}_{AA}\beta_A^* - \delta_A\|_\infty + \|\delta_A - \delta_A\|_\infty \leq C\{\Delta_d(d \log d)/n\}^{1/2}. \tag{S18}
\]

Define vectors \( \tau \in \mathbb{R}^d \) and \( \eta \in \mathbb{R}^d \) such that \( \tau_{j_m} = \text{sgn}(\beta_{j_m}) \) and \( \eta_{j_m} = \beta_{j_m}/\|\beta_{S_j}\|_2 \) for \( j_m \in A \) and \( \tau_{j_m} = \eta_{j_m} = 0 \) for \( j_m \in A^c \). Let \( f(\beta_A) = \hat{\Sigma}_{AA}(\beta_A - \beta_A^*) + \hat{\Sigma}_{AA}\beta_A^* - \delta_A + (1 - \alpha)\lambda_n \tau_A + \alpha \lambda_n \eta_A \) and \( g(\beta_A) = \hat{\Sigma}_{AA}^{-1}f(\beta_A) = \beta_A - \beta_A^* - \hat{\Sigma}_{AA}^{-1}\{\hat{\Sigma}_{AA}\beta_A^* - \delta_A + (1 - \alpha)\lambda_n \tau_A + \alpha \lambda_n \eta_A\} \). By Lemma 4, \( \|\hat{\Sigma}_{AA}^{-1}\|_\infty \) is bounded with probability at least \( 1 - C_1d^{-C_2} \). Hence, by (S18) and the choice of \( \lambda_n \), we have

\[
\|\hat{\Sigma}_{AA}^{-1}\{\hat{\Sigma}_{AA}\beta_A^* - \delta_A + (1 - \alpha)\lambda_n \tau_A + \alpha \lambda_n \eta_A\}\|_\infty \\
\leq \|\hat{\Sigma}_{AA}^{-1}\|_\infty \{\|\hat{\Sigma}_{AA}\beta_A^* - \delta_A\|_\infty + (1 - \alpha)\lambda_n + \alpha \lambda_n\} \\
\leq 2C_0 \left[C\{\Delta_d(d \log d)/n\}^{1/2} + \lambda_n\right] \\
\leq \lambda_n.
\]

Then, for a sufficiently large \( n \), if \( (\beta_A - \beta_A^*) j_m = C\lambda_n \), for some large constant \( C > 0 \),

\[
\{g(\beta_A)\}_m \geq C\lambda_n - \{\hat{\Sigma}_{AA}^{-1}\{\hat{\Sigma}_{AA}\beta_A^* - \delta_A + (1 - \alpha)\lambda_n \tau_A + \alpha \lambda_n \eta_A\}\}_m \geq 0.
\]

If \( (\beta_A - \beta_A^*) j_m = -C\lambda_n \),

\[
\{g(\beta_A)\}_m \leq -C\lambda_n + \{\hat{\Sigma}_{AA}^{-1}\{\hat{\Sigma}_{AA}\beta_A^* - \delta_A + (1 - \alpha)\lambda_n \tau_A + \alpha \lambda_n \eta_A\}\}_m \leq 0.
\]

By the continuity of \( g(\beta_A) \) and the Miranda’s existence theorem (Vrahatis, 1989), \( g(\beta_A) = 0 \) has a solution \( \hat{\beta}_A \) in \( \mathcal{N} \). Obviously, \( f(\hat{\beta}_A) = 0 \). Hence, \( \hat{\beta}_A \) also solves (S14).
Second, we have
\[(\hat{\Sigma}\beta)_B - \hat{\delta}_B = \hat{\Sigma}_{BA}(\beta_A - \beta_A^*) + (\hat{\Sigma}\beta^* - \hat{\delta})_B\]
\[= \hat{\Sigma}_{BA}\hat{\Sigma}^{-1}_{AA}(\hat{\Sigma}A\beta_A^* - \hat{\delta}_A + (1 - \alpha)\lambda_n\tau_A + \alpha\lambda_n\eta_A) + (\hat{\Sigma}\beta^* - \hat{\delta})_B.\]

Similarly as in (S18), \(\|\hat{\Sigma}\beta^* - \hat{\delta}\|_\infty \leq C\{\Delta_d(\log d)/n\}^{1/2}\) with probability at least \(1 - C_1d^{-C_2}\).

By Lemma 4, \(\|\hat{\Sigma}_{BA}\hat{\Sigma}^{-1}_{AA}\|_\infty \leq (1 - \alpha)(1 - \epsilon/2) < 1 - \alpha\) with probability at least \(1 - C_1d^{-C_2}\).

Hence, with probability at least \(1 - C_1d^{-C_2}\),
\[\|(\hat{\Sigma}\beta)_B - \hat{\delta}_B\|_\infty \leq (1 - \alpha)(1 - \epsilon/2)(\|\hat{\Sigma}_{BA}\hat{\Sigma}^{-1}_{AA}\beta_A^* - \hat{\delta}_A\|_\infty + \lambda_n) + \|(\hat{\Sigma}\beta^* - \hat{\delta})_B\|_\infty\]
\[\leq (1 - \alpha)(1 - \epsilon/2)\left[C\{\Delta_d(\log d)/n\}^{1/2} + \lambda_n\right] + C\{\Delta_d(\log d)/n\}^{1/2}\]
\[\leq (1 - \alpha)(1 - \epsilon/2)\lambda_n + (2 - \epsilon/2)C\{\Delta_d(\log d)/n\}^{1/2}\]
\[< (1 - \alpha)\lambda_n,\]

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

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

\[\square\]

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: \(|a^Tx| \leq \|a\|_\infty \|x\|_1, G; \|A\|_{\infty, 2} \leq M\|A\|_{\max}; \|x\|_{\infty, 2} \leq \sqrt{M}\|x\|_{\infty}; \) and \(\|Ax\|_{\infty, 2} \leq \|A\|_{\infty, 2}\|x\|_{1, G}^2\).

**Proof.** For the first statement, we have
\[|a^Tx| = \sum_{j=1}^p \left\{ (1 - \alpha) \sum_{m=1}^M a_{jm}x_{jm} + \alpha \sum_{m=1}^M a_{jm}x_{jm} \right\}\]
\[\leq \sum_{j=1}^p \left\{ (1 - \alpha) \max_{1 \leq m \leq M} |a_{jm}| \|x_{S_j}\|_1 + \alpha \|a_{S_j}\|_2 \|x_{S_j}\|_2 \right\}\]
\[\leq \left( \max_{j, m} |a_{jm}| \right) \sum_{j=1}^p \{ (1 - \alpha) \|x_{S_j}\|_1 + \alpha \|a_{S_j}\|_2 \|x_{S_j}\|_2 \}\]
\[\leq \left( \max_{1 \leq j \leq p} \|a_{S_j}\|_2 \right) \sum_{j=1}^p \{ (1 - \alpha) \|x_{S_j}\|_1 + \alpha \|x_{S_j}\|_2 \}\]
\[= \|a\|_{\infty, 2}\|x\|_{1, G}^2.\]

The second and third statements follow from some simple algebra.

For the last statement, let \(\tilde{a}_{jm}\) denote the \(j_m\)th row of \(\tilde{A}\). By (1), we have
\[|\tilde{a}_{jm}^Tx| \leq \|\tilde{a}_{jm}\|_{\infty, 2}\|x\|_{1, G}^2.\]
Then,
\[
\|Ax\|_{\infty,2}^2 = \max_{1 \leq j \leq n} \sum_{m=1}^{M} (Ax)_m^2 \leq \max_{1 \leq j \leq p} \sum_{m=1}^{M} \|\tilde{a}_{jm}\|_{\infty,2}^2 \|x\|_{1,G}^2 \leq \left( \max_{1 \leq j \leq p} \sum_{m=1}^{M} \|\tilde{a}_{jm}\|_{\infty,2}^2 \right) \|x\|_{1,G}^2
\]
\[
= \|A\|_{\infty,2}^2 \|x\|_{1,G}^2.
\]

**Lemma 2.** Under Condition 1, there exist positive constants \(C, C_1\) and \(C_2\) such that it holds with probability at least \(1 - C_1d^{-C_2}\) that
\[
\|\hat{\Sigma} - \Sigma\|_{\infty,2} \leq C \{ M \Delta_d (\log d)/n \}^{1/2}.
\]

**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
\[
\hat{\Sigma} = \frac{1}{n} \left( \sum_{Y_i=0} U_{i0}U_{i0}^T + \sum_{Y_i=1} U_{i1}U_{i1}^T \right) - \frac{n_0}{n} \tilde{U}_0 \tilde{U}_0^T - \frac{n_1}{n} \tilde{U}_1 \tilde{U}_1^T
\]
\[
= \hat{\Sigma} - \frac{n_0}{n} \tilde{U}_0 \tilde{U}_0^T - \frac{n_1}{n} \tilde{U}_1 \tilde{U}_1^T.
\]
It suffices to prove the result with \(\hat{\Sigma}\) replaced by \(\tilde{\Sigma}\). To simplify the presentation, denote \(Z_i = U_{i0} (1 \leq i \leq n_0)\) and \(Z_i = U_{i1} (n_0 + 1 \leq i \leq n)\). Then,
\[
\tilde{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} Z_iZ_i^T \Omega - E(Z_iZ_i^T \Omega).
\]
Denote \(\xi_{ij} = Z_{ij}Z_i^T \Omega - E(Z_{ij}Z_i^T \Omega)\). With \(e_j\) being a vector with 1 for the \(j\)th coordinate and 0 elsewhere, we have
\[
\text{var}(\xi_{ij}) = \text{var}(e_j^T Z Z^T \Omega) = \text{var}(Z^T \Omega e_j^T Z) = \frac{1}{2} \text{tr} \{ (\Omega e_j^T e_j^T \Omega + \sigma_{jj} \delta^T \delta) \Sigma(\Omega e_j^T e_j^T \Omega + \sigma_{jj} \delta^T \delta) \} = \delta_j^2 + \sigma_{jj} \delta^T \delta \lesssim \delta^T \Omega \delta.
\]
Since \(\{\xi_{ij}\}_{i=1}^n\) are independent sub-exponential random variables with mean 0, we have
\[
\text{pr} \left\{ \max_{1 \leq j \leq d} \frac{1}{n} \sum_{i=1}^{n} \xi_{ij} \geq C \left( \frac{\Delta_d \log d}{n} \right)^{1/2} \right\} \leq C_1d^{-C_2}.
\]
Then, by Lemma 1 and the union bound, we have
\[
\text{pr} \left\{ \|\hat{\Sigma} - \mu_0 - \mu_1\|_{\infty,2} \geq C \left( \frac{\Delta_d \log d}{n} \right)^{1/2} \right\}
\]
\[
\leq \text{pr} \left\{ \|\hat{\Sigma} - \mu_0 - \mu_1\|_{\infty} \geq C \left( \frac{\Delta_d \log d}{n} \right)^{1/2} \right\} \leq C_1d^{-C_2}.
\]

**Lemma 3.** Let \(\hat{\beta}\) be the solution of problem (3). Then it holds that
\[
\|\hat{\Sigma} \hat{\beta} - \hat{\delta}\|_{\infty,2} \leq \lambda_n M^{1/2}.
\]
Proof. Let \( F(\beta) = \beta^T \Sigma \beta / 2 - \hat{\delta}^T \beta + \lambda_n \sum_{j=1}^{p} \| \beta S_j \|_G \). Define
\[
H(\beta, \gamma, \eta) = \frac{1}{2} \beta^T \Sigma \beta - \hat{\delta}^T \beta + (1 - \alpha) \lambda_n \sum_{j=1}^{p} \gamma_{S_j}^T \beta S_j + \alpha \lambda_n \sum_{j=1}^{p} \eta_{S_j}^T \beta S_j.
\]
Then, we have
\[
F(\beta) = \max_{\| \gamma \|_2 \leq 1} H(\beta, \gamma, \eta).
\]
By the strong duality, \( \hat{\beta} \) also solves
\[
\min_{\beta} F(\beta) = \min_{\beta} \max_{\| \gamma \| \leq 1, \| \eta \|_2 \leq 1} H(\beta, \gamma, \eta) = \max_{\| \gamma \| \leq 1} \min_{\| \eta \|_2 \leq 1} H(\beta, \gamma, \eta).
\]
By the Karush–Kuhn–Tucker condition, we have \( \hat{\Sigma} \hat{\beta} - \hat{\delta} + (1 - \alpha) \lambda_n \gamma + \alpha \lambda_n \eta = 0 \). Since \( \| \gamma \|_\infty \leq 1 \) and \( \| \eta \|_2 \leq 1 \), we have
\[
\| \hat{\Sigma} \hat{\beta} - \hat{\delta} \|_2 \leq (1 - \alpha) \lambda_n \| \gamma \|_\infty,2 + \alpha \lambda_n \| \eta \|_\infty,2 \leq (1 - \alpha) \lambda_n \sqrt{M} + \alpha \lambda_n \leq \lambda_n \sqrt{M}.
\]

**Lemma 4.** Under Conditions 1 and 5–8, if \( s \{ \log d \} / n \}^{1/2} = o(1) \), there exist positive constants \( C, C_1 \) and \( C_2 \) such that, with probability at least \( 1 - C_1 d^{-C_2} \), we have \( \| \Sigma_{\hat{\beta}}^{-1} \|_\infty \leq 2c_0; \| \Sigma_{\beta A} \Sigma_{\hat{\beta} A}^{-1} \|_\infty \leq (1 - \alpha)(1 - \epsilon / 2); \| \Sigma_{\hat{\beta} A} \Sigma_{\hat{\beta} A}^{-1} \|_\infty \leq (1 - \epsilon / 2) M^{-1/2}; \) and \( \lambda_{\min}(\Sigma_{\beta A}) \geq c_0^{-1/2} \).

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 \( 1 \leq i, j \leq d \),
\[
\Pr [ | \hat{\sigma}_{ij} - \sigma_{ij} | > C \{ \log d \} / n }^{1/2} \] \leq C_1 d^{-C_2 + 2}.
\]
By the union bound, we have
\[
\Pr [ \| \Sigma_{\hat{\beta} A} - \Sigma_{\beta A} \|_\infty > C s \{ \log d \} / n }^{1/2} \] = \Pr \left[ \max_{i \in A} \sum_{j \in A} | \hat{\sigma}_{ij} - \sigma_{ij} | > C s \{ \log d \} / n }^{1/2} \right]
\leq s \Pr [ \| \hat{\sigma}_{ij} - \sigma_{ij} | > C s \{ \log d \} / n }^{1/2} \]
\leq C_1 s^{-2} d^{(C_2 + 2)} \leq C_1 d^{-C_2}.
\]

Then, with probability at least \( 1 - C_1 d^{-C_2} \), we have
\[
\| \Sigma_{\hat{\beta} A}^{-1} \|_\infty \leq \| \Sigma_{\beta A}^{-1} \|_\infty + \| \Sigma_{\beta A}^{-1} \|_\infty \| \Sigma_{\hat{\beta} A} - \Sigma_{\beta A} \|_\infty \| \Sigma_{\beta A}^{-1} \|_\infty \leq c_0 + c_0 \| \Sigma_{\hat{\beta} A} \|_\infty C s \{ \log d \} / n }^{1/2}.
\]
Therefore, when \( n \) is sufficiently large,
\[
\| \Sigma_{\hat{\beta} A}^{-1} \|_\infty \leq \frac{c_0}{1 - C_0 s \{ \log d \} / n }^{1/2} \leq 2c_0.
\]
Then, \(\|\Sigma_{BA}^{-1} - \Sigma_{AA}^{-1}\|_\infty \leq (\|\Sigma_{BA}\|_\infty + \|\Sigma_{BA} - \Sigma_{AA}\|_\infty)\|\Sigma_{BA}^{-1} - \Sigma_{AA}^{-1}\|_\infty\).

By definition, \(\|\Sigma_{BA}\|_\infty = \max_{i\in B} \sum_{j\in A} |\sigma_{ij}| \leq s\). Similarly as (S20), we have
\[
\Pr\left[\|\Sigma_{BA}^{-1} - \Sigma_{AA}^{-1}\|_\infty > Cs\left((\log d)/n\right)^{1/2}\right] \leq C_1 d^{-C_2}.
\]

By (S21) and Condition 6, with probability at least \(1 - C_1 d^{-C_2}\), we have
\[
\|\Sigma_{BA}^{-1} - \Sigma_{AA}^{-1}\|_\infty \leq s^2\left((\log d)/n\right)^{1/2}.
\]

By a similar argument, \(\|\hat{\Sigma}_{BA}^{-1} - \hat{\Sigma}_{AA}^{-1}\|_\infty \leq s^2\left((\log d)/n\right)^{1/2}\). When the sample size \(n\) is large enough, both upper bounds become arbitrarily small. Hence, \(\|\hat{\Sigma}_{BA}^{-1} - \hat{\Sigma}_{AA}^{-1}\|_2 \leq (1 - \alpha)\epsilon/2\). 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 \(\lambda_{\min}(\hat{\Sigma}_{AA}) \geq c_0^{-1}\). Then by a similar proof, we can show that \(\lambda_{\min}(\hat{\Sigma}_{AA}) \geq c_0^{-1}/2\) 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, \(X | Y = 0 \sim t_3(0, \Sigma)\), and \(X | Y = 1 \sim \mu_1 + t_3(0, \Sigma)\), where \(t_3(0, \Sigma)\) is the multivariate t-distribution with 3 degrees of freedom and the scale parameter \(\Sigma\). 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 \(\beta_{jm}^* = 0.5\) \((j = 1, \ldots, 5; m = 1, \ldots, M)\) 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

<table>
<thead>
<tr>
<th></th>
<th>$n = 50, p = 100, \pi = 1$</th>
<th>$n = 50, p = 100, \pi = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>iLDA-r</td>
<td>iLDA</td>
</tr>
<tr>
<td>error rate</td>
<td>8(4)</td>
<td>12(9)</td>
</tr>
<tr>
<td>sensitivity</td>
<td>92(11)</td>
<td>68(25)</td>
</tr>
<tr>
<td>specificity</td>
<td>74(22)</td>
<td>78(23)</td>
</tr>
</tbody>
</table>

$M = 50, p = 200, \pi = 1$  

|                | 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

<table>
<thead>
<tr>
<th></th>
<th>$n = 50, p = 100, \pi = 1$</th>
<th>$n = 50, p = 100, \pi = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>iLDA-e</td>
<td>iLDA-e</td>
</tr>
<tr>
<td>error rate</td>
<td>22(9)</td>
<td>25(7)</td>
</tr>
<tr>
<td>sensitivity</td>
<td>66(20)</td>
<td>55(25)</td>
</tr>
<tr>
<td>specificity</td>
<td>91(27)</td>
<td>72(36)</td>
</tr>
</tbody>
</table>

$M = 50, p = 200, \pi = 1$  

|                | iLDA-e | iLDA-e | m-vote | sLDA | iLDA-e | iLDA-e | 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


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

<table>
<thead>
<tr>
<th></th>
<th>$M = 2$</th>
<th>$M = 4$</th>
<th>$M = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes error</td>
<td>16</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>iLDA</td>
<td>21</td>
<td>13</td>
<td>11</td>
</tr>
</tbody>
</table>

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