$\mathcal{DBSDA}$: Lowering the Bound of Misclassification Rate for Sparse Linear Discriminant Analysis via Model Debiasing

Haoyi Xiong, *Member, IEEE*, Wei Cheng, *Member, IEEE*, Jiang Bian, *Student Member, IEEE*, Wenqing Hu, Zeyi Sun, and Zhishan Guo, *Member, IEEE*

***Abstract*— Linear discriminant analysis (LDA) is a well-known technique for linear classification, feature extraction, and dimension reduction. To improve the accuracy of LDA under the high dimension low sample size (HDLSS) settings, shrunken estimators, such as Graphical Lasso, can be used to strike a balance between biases and variances. Although the estimator with induced sparsity obtains a faster convergence rate, however, the introduced bias may also degrade the performance. In this paper, we theoretically analyze how the sparsity and the convergence rate of the precision matrix (also known as inverse covariance matrix) estimator would affect the classification accuracy by proposing an analytic model on the upper bound of an LDA misclassification rate. Guided by the model, we propose a novel classifier, $\mathcal{DBSDA}$, which improves classification accuracy through *debiasing*. Theoretical analysis shows that $\mathcal{DBSDA}$ possesses a reduced upper bound of misclassification rate and better asymptotic properties than sparse LDA (SDA). We conduct experiments on both synthetic datasets and real application datasets to confirm the correctness of our theoretical analysis and demonstrate the superiority of $\mathcal{DBSDA}$ over LDA, SDA, and other downstream competitors under HDLSS settings.**

***Index Terms*— Classification, debiasing, linear discriminant analysis, sparsity.**

I. Introduction

LINEAR discriminant analysis (LDA) [1] is a well-known technique for feature extraction and dimension reduction [2]. It has been widely used in many applications, such as face recognition [3], image retrieval, and so on. An intrinsic limitation of classical LDA is that its objective function relies on the *well-estimated* and *nonsingular* covariance matrices. For many applications, such as the microarray data analysis, all scatter matrices can be *singular* or *ill-posed* since the data are often with high dimension but low sample size (HDLSS) [4].

The classical LDA classifier relies on two key parameters–the mean vector of each type and the precision matrix. Under the HDLSS settings, the sample precision matrix (also known as the inverse of sample covariance matrix) used in LDA is usually ill-estimated and quite different from the inverse of population/true covariance matrix [5]. For example, the largest eigenvalue of the sample covariance matrix is not a consistent estimate of the largest eigenvalue of the population covariance matrix, and the eigenvectors of the sample covariance matrix can be nearly orthogonal to the truth when the number of dimensions is greater than the number of samples [6], [7]. Such inconsistency between the true and the estimated precision matrices degrades the accuracy of LDA classifiers under the HDLSS settings [8], [9].

A plethora of excellent work has been conducted to address such HDLSS data classification problem. For example, Krzanowski *et al.* [10] suggested to use pseudoinverse to approximate the inverse covariance matrix, when the sample covariance matrix is singular. However, the precision of pseudoinverse LDA is usually low and not well guaranteed [11]. Other techniques include the two-stage algorithm principle component analysis + LDA [11], [12], LDA based on Kernels [13]–[16] and/or other nonparametric statistics [17]–[20]. To overcome the singularity of the sample covariance matrices, instead of estimating the inverse covariance matrix and mean vectors separately, Clemmensen *et al.* [21], Cai and Liu [22], and Mai *et al.* [23] proposed to estimate the projection vector for discrimination directly. More popularly, regularized LDA approaches [3], [10], [24] are proposed to solve the problem. These methods can improve the performance of LDA either empirically or theoretically, while few of them can directly address the ill-estimated inverse covariance matrix estimation issue.

One representative regularized LDA approach is to replace the precision matrix used in LDA with a shrunken estimator [3], [10], [24], such as Graphical Lasso [25], so as to achieve a "superior prediction." Intuitively, through replacing the precision matrix used in LDA with a sparse regularized estimation, the ill-posed problem caused by HDLSS settings can be well addressed. The sparse estimators usually converge to the inverse of true/population covariance matrix faster than

Manuscript received August 2, 2017; revised March 1, 2018 and June 1, 2018; accepted June 1, 2018. Date of publication July 24, 2018; date of current version February 19, 2019. *(Corresponding author: Haoyi Xiong.)*

H. Xiong is with the Big Data Laboratory, Baidu Research, Beijing 100193, China, and also with the National Engineering Laboratory of Deep Learning Technology and Application, Beijing 100193, China (e-mail: xhyccc@gmail.com).

W. Cheng is with the NEC Labs of America, Princeton, NJ 08540 USA.

J. Bian is with the Big Data Laboratory, Baidu Research, Beijing 100193, China, and also with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO 65409 USA.

W. Hu and Z. Sun are with the Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65409 USA.

Z. Guo is with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO 65409 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2846783

the sample estimators [5]. With the asymptotic properties, the sparse LDA (SDA) should be close to the optimal LDA. However, the way that the sparsity and the convergence rate of the precision matrix estimator would affect the classification accuracy is not well studied in the literature.

Furthermore, with induced sparsity, the inverse covariance estimator becomes biased [26]. The performance of SDA is frequently bottlenecked due to the bias of the sparse estimators. Recently, researchers tried to debias the Lasso estimator [26]–[28], through adjusting the ℓ_1-penalty for the regularized estimation, so as to achieve a better regression performance. Inspired by this line of research, we propose to improve SDA through debiasing (i.e., desparsifing) in this paper.

Our Contributions: With respect to the aforementioned issues, in this paper, we made following contributions.

1) We propose a novel analytic model for the LDA misclassification rate, based on the convergence rate of inverse) covariance matrix estimator and the sparsity/density of the estimates. This model can derive the upper bounds of LDA misclassification rate on both the Gaussian and non-Gaussian datasets.
2) Guided by the proposed analytic model, we first analyze the most commonly seen SDA via Graphical Lasso [24], and study the upper bounds of the SDA misclassification rate. Inspired by debiased Lasso [28], we then develop a novel classifier $\mathcal{DBSDA}$—debiased sparse discriminant analysis. $\mathcal{DBSDA}$ leverages yet another debiased estimator for a linear classification problem, to reduce the upper bounds of misclassification rate, through balancing the biases and variances in a regularized model.
3) Our theoretical analysis based on the proposed analytic model shows, in terms of asymptotic properties of projection vector (also known as the vector β), $\mathcal{DBSDA}$ converges faster than SDA; in terms of misclassification rate, $\mathcal{DBSDA}$ enjoys a reduced upper bound of misclassification rate than SDA. We also conduct extensive experiments to demonstrate the advantage of the proposed algorithms comparing to other competitors. The results validate the correctness of our theoretical analysis.

The paper is organized as follows. In Section II, we review the LDA models and summarize the existing LDA misclassification rate analysis model. In Section III, we propose an analytic model characterizing the error bound of misclassification rate based on the sparsity and convergence rate of an inverse covariance matrix estimator. In Section IV, guided by the proposed analytic model, we introduce a baseline algorithm SDA, then propose our algorithm, $\mathcal{DBSDA}$, and further compare their performances. In Section V, we validate the proposed algorithms, synthesized datasets, benchmark datasets, and real-world applications. Finally, we review the related work, discuss the limitation, and conclude the paper in Sections VI and VII, respectively.

II. Preliminaries

In this section, we first briefly introduce the binary classifier using traditional LDA. Then, we present the state-of-the-art analytic models on the LDA misclassification rate which assume that the data for classification follow certain Gaussian distributions. The main notations are listed in Table I.

TABLE I

Summary of Notations

Symbol	Definition
(x, ℓ)	the labeled data pairs
m	the number of training samples
x	the data vector for classification
l	the label of the data
$\bar{f}(\cdot)$	the classification rule of FDA model
π_l, π_+, π_-	the frequency of samples
μ^*, μ^*_+, μ^*_-	the population means
Σ^*, Θ^*	the population covariance matrix and its inverse
$\bar{\mu}, \bar{\mu}_+, \bar{\mu}_-$	the sample estimation of means
$\bar{\Sigma}$, $\bar{\Theta}$	the sample covariance matrix and its inverse
β	(optimal) linear discriminant projection vector
$\bar{\beta}$	the estimated linear discriminant projection vector
$\Phi(\cdot)$	the CDF of standard normal distribution
$\bar{\Delta}_\mu$	$\bar{\Delta}_\mu = \bar{\mu}_+ - \bar{\mu}_-$
$\|\| \cdot \|\|_F$	the Frobenius Matrix Norm
$\|\| \cdot \|\|_2$	the Spectral Matrix Norm
M	the full rank square matrix
$\mathbb{E}(\cdot)$	the expectation of target function
$D_{KL}(\cdot)$	the Kullback–Leibler divergence
$\widehat{f}(\cdot)$	the classification result of SDA model
$\widehat{\Theta}$	the Graphical Lasso (GLasso) estimator
C_1, C_2	the positive constant
p	the dimensionality of data sample
d	the maximal degree of the graph
$\widehat{\beta}^G$	the projection vectors based on GLasso
$\widehat{f}^D(\cdot)$	the classification rule of $\mathcal{DBSDA}$ model
$\mathbf{X}$	$m \times p$ matrix where the i^{th} column is x_i
$\mathbf{L}$	$1 \times m$ matrix where the i^{th} row is ℓ_i
$\mathbf{U}$	$m \times p$ matrix with each column as $\frac{\bar{\mu}_+ + \bar{\mu}_-}{2}$
$\mathbf{C^g}$	p-dimensional vector with each row as c^g
$\widehat{\beta}^D$	the de-biased projection vector
$\| \cdot \|_\infty$	the ℓ_∞-vector-norm
$\mathcal{O}_p$	Big-O-Notation with Probability
$\mathbf{I}$	$p \times p$ identity matrix

A. LDA for Binary Classification

To use Fisher's LDA (FDA), given the independent identically distributed (i.i.d.) labeled data pairs $(x_1, l_1) \ldots (x_m, l_m)$, we first estimate the sample covariance matrix $\bar{\Sigma}$ using the pooled sample covariance matrix estimator with respect to the two classes [1], then estimate the sample precision matrix as $\bar{\Theta} = \bar{\Sigma}^{-1}$. Furthernore, $\bar{\mu}_+$ and $\bar{\mu}_-$ are estimated as the mean vectors of the positive samples and the negative samples in the m training samples, respectively.

Given all estimated parameters $\bar{\Sigma}$ (and $\bar{\Theta} = \bar{\Sigma}^{-1}$), $\bar{\mu}_+$ and $\bar{\mu}_-$, the FDA model classifies a new data vector x as the result of

$$\bar{f}(x) = \underset{\ell \in \{-,+\}}{\operatorname{argmax}} \delta(x, \bar{\Theta}, \bar{\mu}_\ell, \pi_\ell)$$

where

$$\delta(x, \bar{\Theta}, \bar{\mu}_\ell, \pi_\ell) = x^T \bar{\Theta} \bar{\mu}_\ell - \frac{1}{2} \bar{\mu}_\ell^T \bar{\Theta} \bar{\mu}_\ell + \log \pi_\ell \qquad (1)$$

where π_+ and π_- refer to the (foreknown) frequencies of positive samples and negative samples in the whole population, respectively.

B. LDA Misclassification Rate for Gaussian Data With Uncertain Covariance Estimates

In this section, we summarize the studies [8], [9], [29], [30] in theoretical misclassification rate of LDA classifiers for classifying multivariate Gaussian data.

We first assume that the data for binary classification follow two (unknown) Gaussian distributions with the same covariance matrix Σ^* (i.e., the inverse covariance matrix $\Theta^* = \Sigma^{*-1}$) but two different means μ_+^* and μ_-^*, i.e., $\mathcal{N}(\mu_+^*, \Sigma^*)$ for positive samples and $\mathcal{N}(\mu_-^*, \Sigma^*)$ for negative samples, respectively. Given the LDA classifier $\bar{f}(x)$ based on the sample estimated mean vectors $\bar{\mu}_+$, $\bar{\mu}_-$, and a specific covariance matrix $\bar{\Sigma}$ (and $\bar{\Theta} = \bar{\Sigma}^{-1}$), Zollanvari *et al.* [8] and Zollanvari and Dougherty [9] modeled the expected misclassification rate of an LDA (i.e., probability of $\ell \neq \bar{f}(x)$) on the data of $\mathcal{N}(\mu_+^*, \Sigma^*)$, $\mathcal{N}(\mu_-^*, \Sigma^*)$ as a function $\varepsilon(\bar{\mu}_+, \bar{\mu}_-, \bar{\Sigma}, \mu_+^*, \mu_-^*, \Sigma^*)$, that is,

$$\begin{aligned}
&\varepsilon\left(\bar{\mu}_+, \bar{\mu}_-, \bar{\Sigma}, \mu_+^*, \mu_-^*, \Sigma^*\right)\\
&= \pi_+ \cdot \Phi\left(-\frac{\left(\mu_+^* - \frac{(\bar{\mu}_+ + \bar{\mu}_-)}{2}\right)^T \bar{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)}{\sqrt{(\bar{\mu}_+ - \bar{\mu}_-)^T \bar{\Theta}\Sigma^*\bar{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)}}\right)\\
&\quad + \pi_- \cdot \Phi\left(\frac{\left(\mu_-^* - \frac{(\bar{\mu}_+ + \bar{\mu}_-)}{2}\right)^T \bar{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)}{\sqrt{(\bar{\mu}_+ - \bar{\mu}_-)^T \bar{\Theta}\Sigma^*\bar{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)}}\right)\\
&= \pi_+ \cdot \Phi\left(\frac{(2(\bar{\mu}_+ - \mu_+^*) - (\bar{\mu}_+ - \bar{\mu}_-))^T \bar{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)}{2\sqrt{(\bar{\mu}_+ - \bar{\mu}_-)^T \bar{\Theta}\Sigma^*\bar{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)}}\right)\\
&\quad + \pi_- \cdot \Phi\left(\frac{(2(\mu_-^* - \bar{\mu}_-) - (\bar{\mu}_+ - \bar{\mu}_-))^T \bar{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)}{2\sqrt{(\bar{\mu}_+ - \bar{\mu}_-)^T \bar{\Theta}\Sigma^*\bar{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)}}\right)
\end{aligned} \tag{2}$$

where $\bar{\Theta} = \bar{\Sigma}^{-1}$ and $\Phi(\cdot)$ refers to the CDF function of a standard normal distribution. It is obvious that the expected misclassification rate is sensitive with the parameters $\mu_+^*, \mu_-^*, \Sigma^*, \bar{\mu}_+, \bar{\mu}_-$, and $\bar{\Sigma}$, while the true parameters μ_+^*, μ_-^*, and Σ^* are usually unknown.

Assumption 1: Given m samples $x_1, x_2, \ldots, x_m$ drawn from $\mathcal{N}(\mu_+^*, \Sigma^*)$ and $\mathcal{N}(\mu_-^*, \Sigma^*)$ with constant priors, $\bar{\mu}_+$ and $\bar{\mu}_-$ are estimated using sample estimators. According to [31], the sample mean $\bar{\mu}_+, \bar{\mu}_-$, and $\bar{\mu}$ converge to the population mean μ_+^*, μ_-^*, and μ^* at the ℓ_2-norm convergence rate $\mathcal{O}(p/m)^{1/2}$ in high probability.

Assumption 2: We assume that for each sample $|x_i|_2 \leq \mathcal{L}$. Thus, there has $|\bar{\mu}_+ - \bar{\mu}_-|_2 \leq 2\mathcal{L}$. For sample covariance matrix $\bar{\Sigma} = m^{-1}\sum_{i=1}^m x_i x_i^T$, there has $\|\bar{\Sigma}\|_2 = \lambda_{\max}(\bar{\Sigma}) \leq \text{trace}(\bar{\Sigma}) \leq \mathcal{L}^2$, as $\bar{\Sigma}$ is a positive semidefinite matrix with all eigenvalues nonnegative.

Assumption 3: As was assumed in [32] and [33], we assume that there exists a positive constant $\mathcal{K}$ that can bound the eigenvalues of Σ^* as $1/\mathcal{K} \leq \lambda_{\min}(\Sigma^*) \leq \lambda_{\max}(\Sigma^*) \leq K$. Since $\Theta^* = \Sigma^{*-1}$, then there also exists $1/\mathcal{K} \leq \lambda_{\min}(\Theta^*) \leq \lambda_{\max}(\Theta^*) \leq \mathcal{K}$. In this way, there exist $\|\Sigma^*\|_2 \leq \mathcal{K}$ and $\|\Theta^*\|_2 \leq \mathcal{K}$.

Theorem 1: We first denote $\bar{\Delta}_\mu$ as $\bar{\Delta}_\mu = \bar{\mu}_+ - \bar{\mu}_-$. Then, based on [8] and [9], the upper bound of the expected misclassification rate of $\bar{f}(x)$ can be reduced to

$$P_{\sim\mathcal{N}}[\ell \cdot \bar{f}(x) < 0] \leq \Phi\left(\frac{C\sqrt{p/m} \cdot |\bar{\Theta}\bar{\Delta}_\mu|_2 - \bar{\Delta}_\mu^T \bar{\Theta}\bar{\Delta}_\mu}{2\sqrt{\bar{\Delta}_\mu^T \bar{\Theta}\Sigma^*\bar{\Theta}\bar{\Delta}_\mu}}\right) \tag{3}$$

where C refers to a positive constant.

Intuitively, when the sample estimation of a covariance matrix $\bar{\Sigma}$ (also $\bar{\Theta}$) is close to the population ones Σ^* (and Θ^*), the expected error rate can be minimized.

III. Main Theory: Upper Bounds of LDA Misclassification Rates

In this section, we analyze the performance of a classical LDA classifier. We derive the upper bounds of the LDA misclassification rate on both the Gaussian and non-Gaussian data. Our result shows, for both the Gaussian and non-Gaussian datasets, the upper bounds of LDA misclassification rates are sensitive to the sparsity of inverse covariance matrix estimator and the convergence rate of the estimator.

A. LDA Misclassification Rate for Gaussian Data

Let denote $\Delta_\mu = |\bar{\mu}_+ - \bar{\mu}_-|_2/2$ referring the gap between means, in the rest of this paper.

Theorem 2: Suppose the data for the binary classification (training and testing) follows the Gaussian distributions $\mathcal{N}(\mu_+^*, \Sigma^*)$ and $\mathcal{N}(\mu_-^*, \Sigma^*)$ (with $\Theta^* = \Sigma^{*-1}$). Given an inverse covariance matrix estimator $\bar{\Theta}$ and mean estimators $\bar{\mu}_+$ and $\bar{\mu}_-$ over m samples drawn from the distribution with constant priors, the misclassification rate of $\bar{f}(x)$ will be upper bounded by

$$P_{\sim\mathcal{N}}[\ell \cdot \bar{f}(x) < 0] \leq \Phi\left(\frac{C}{2}\sqrt{\mathcal{K}\cdot\frac{p}{m}} - \frac{|\bar{\Delta}_\mu|_2}{2\sqrt{\mathcal{K}}\mathcal{L}^2}\|\bar{\Theta}\|_2^{-1}\right). \tag{4}$$

Theorem 1 suggests that the performance of LDA can be improved with *lower misclassification upper bound*, when using a *sparse inverse covariance matrix estimator* with *faster convergence rate* in *spectral norm.*

Proof: To prove Theorem 1, given the symmetric positive definite matrices Σ^* and $\Theta^* = \Sigma^{*-1}$, there must exist the Cholesky decomposition matrix M having $\Sigma^* = M^T M$ and $\Sigma^{*-1} = \Theta^* = M^{-1}(M^{-1})^T$

$$P[\ell \cdot \bar{f}_{\sim\mathcal{N}}(x) \leq 0] \leq \Phi\left(\frac{C\sqrt{p/m} \cdot |\bar{\Theta}\bar{\Delta}_\mu|_2 - \bar{\Delta}_\mu^T \bar{\Theta}\bar{\Delta}_\mu}{2|M\bar{\Theta}\bar{\Delta}_\mu|_2}\right).$$

Since $\Phi(\cdot)$ is monotonically increasing, we have

$$P[\ell \cdot \bar{f}_{\sim\mathcal{N}}(x) \leq 0] \leq \Phi\left(\frac{C}{2}\sqrt{\frac{p}{m}} \cdot \|M^{-1}\|_2 - \frac{\bar{\Delta}_\mu^T \bar{\Theta}\bar{\Delta}_\mu}{2|M\bar{\Theta}\bar{\Delta}_\mu|_2}\right).$$

Since there exists (1) $\|M^{-1}\|_2 = (\lambda_{\max}((M^{-1})^T M^{-1}))^{1/2} = (\lambda_{\max}(\Theta^*))^{1/2} \le (\mathcal{K})^{1/2}$. (2) $\bar{\Delta}_\mu^T \bar{\Theta} \bar{\Delta}_\mu \ge \lambda_{\min}(\bar{\Theta}) |\bar{\Delta}_\mu|_2^2 = 1/\lambda_{\max}(\bar{\Sigma}) \cdot |\bar{\Delta}_\mu|_2^2 \ge (1/\mathcal{L}^2)|\bar{\Delta}_\mu|_2^2$, and (3) $\|M\|_2 = (\lambda_{\max}(M^T M))^{1/2} = (\lambda_{\max}(\Sigma^*))^{1/2} \le (\mathcal{K})^{1/2}$, then we have

$$P[\ell \cdot \bar{f}_{\sim\mathcal{N}}(x) \le 0] \le \Phi\left(\frac{C}{2}\sqrt{\mathcal{K} \cdot \frac{p}{m}} - \frac{|\bar{\Delta}_\mu|_2}{2\sqrt{\mathcal{K}}\mathcal{L}^2}\|\bar{\Theta}\|_2^{-1}\right).$$

□

B. LDA Misclassification Rate for Non-Gaussian Data

In this section, we generalize the LDA misclassification rate from Gaussian data to non-Gaussian data.

Theorem 3: Suppose the labeled data pairs (x, ℓ) follows a joint probability distribution with density function $P(x, \ell)$. The population covariance matrix Σ^* is defined as

$$\Sigma^* = \mathbb{E}_{x \overset{i.i.d.}{\sim} P(X)}[(\mathbf{X} - \bar{\mu})(\mathbf{X} - \bar{\mu})^{\mathrm{T}}] \tag{5}$$

where $\bar{\mu}$ is the sample mean estimated from the training samples drawn i.i.d. from an unknown distribution with density function $P(x) = \sum_{\ell \in \{-1,+1\}} P(x, \ell)$. Based on our assumption on mean vectors, the Gaussian distribution $\mathcal{N}(\bar{\mu}, \Sigma^*)$ is the nearest Gaussian distribution to the data, with minimized Kullback–Leibler divergence [34]. Then, there exists an upper bound of $\bar{f}(x)$'s misclassification rate on such data

$$\begin{aligned} &P[\ell \cdot \bar{f}(x) < 0] \\ &\le\ P_{\sim\mathcal{N}}[\ell \cdot \bar{f}(x) < 0] + \sum_{\ell \in \{-1,+1\}} \pi_\ell \sqrt{\frac{D_{\mathrm{KL}}(P_\ell \| \mathcal{N}(\bar{\mu}_\ell, \Sigma^*))}{2}} \end{aligned} \tag{6}$$

where P_ℓ refers to the distribution of data x with a specific label $\ell \in \{+1, -1\}$ and the probability density function is $P(x|\ell) = (P(x, \ell)/\pi_\ell)$, and $D_{\mathrm{KL}}(P_\ell \| \mathcal{N}(\bar{\mu}_\ell, \Sigma^*))$ refers to the Kullback–Leibler divergence between the distribution of the data and Gaussian distribution $\mathcal{N}(\bar{\mu}_\ell, \Sigma^*)$.

Proof: Given an LDA classifier $\bar{f}(x) : X \to \{+1, -1\}$, we define the misclassification function $\mathbf{1}[\ell \ne \bar{f}(x)] = 1$ when $\ell \ne \bar{f}(x)$ and $\mathbf{1}[\ell = \bar{f}(x)] = 0$ when $\ell = \bar{f}(x)$. Then, the misclassification rate of $\bar{f}(x)$, for any data distribution with density functions $P(x, \ell)$ and $\ell \in \{+1, -1\}$, can be written as

$$P[l \cdot \bar{f}(x) < 0] = \sum_{\ell \in \{+1,-1\}} \int_x \mathbf{1}[\ell \ne \bar{f}(x)] \cdot P(x, \ell) d_x.$$

Consider the density functions $P_{\sim\mathcal{N}}(x|\ell)$ for Gaussian distribution $\mathcal{N}(\bar{\mu}_\ell, \Sigma^*)$ and $\ell \in \{+1, -1\}$

$$\begin{aligned} &P[l \cdot \bar{f}(x) < 0] \\ &= \sum_{\ell \in \{+1,-1\}} \int_x \mathbf{1}[\ell \ne \bar{f}(x)] \cdot \pi_\ell \cdot P_{\sim\mathcal{N}}(x|\ell) d_x \\ &\quad + \sum_{\ell \in \{+1,-1\}} \int_x \mathbf{1}[\ell \ne \bar{f}(x)] \cdot \pi_\ell \cdot (P(x|\ell) - P_{\sim\mathcal{N}}(x|\ell)) d_x. \end{aligned}$$

Consider the misclassification rate on Gaussian distribution $P_{\sim\mathcal{N}}[\ell \cdot \bar{f}(x) < 0]$. Thus

$$\begin{aligned} P[l \cdot \bar{f}(x) < 0] &\le P_{\sim\mathcal{N}}[\ell \cdot \bar{f}(x) < 0] \\ &\quad + \sum_{\ell \in \{+1,-1\}} \pi_\ell \int_x |P(x|\ell) - P_{\sim\mathcal{N}}(x|\ell)| d_x. \end{aligned}$$

Consider Pinsker's inequality

$$\begin{aligned} &P[l \cdot \bar{f}(x) < 0] \\ &\quad \le P_{\sim\mathcal{N}}[\ell \cdot \bar{f}(x) < 0] + \sum_{\ell \in \{-1,+1\}} \pi_\ell \sqrt{\frac{D_{\mathrm{KL}}(P_\ell \| \mathcal{N}_{(\Sigma^*, \bar{\mu}_\ell)})}{2}} \end{aligned}$$

where $D_{\mathrm{KL}}(P_\ell \| \mathcal{N}_{(\Sigma^*, \bar{\mu}_\ell)})$ refers to the Kullback–Leibler divergence (KLD) between the Gaussian distribution $\mathcal{N}(\bar{\mu}_\ell, \Sigma^*)$ to the arbitrary distribution P_ℓ with the density function $P(x|\ell) = (P(x, \ell)/\pi_\ell)$. While the KLD of a given dataset to its nearest Gaussian distributions are frequently fixed, the misclassification rate of $\bar{f}(x)$ on arbitrary data distribution is sensitive with the Gaussian bound $P_{\sim\mathcal{N}}[\ell \cdot \bar{f}(x) < 0]$. □

Remark 1: Theorem 2 suggests that we can consider any distribution as the combination of its nearest Gaussian distribution [i.e., $\mathcal{N}(\bar{\mu}, \Sigma^*)$] and other non-Gaussian components [35]. Given an LDA classifier $\bar{f}(x)$, the misclassification rate is affected by two factors: 1) the misclassification rate of $\bar{f}(x)$ on the nearest Gaussian distribution of the data, i.e., $P_{\sim\mathcal{N}}[\ell \cdot \bar{f}(x) < 0]$ and 2) the divergence between the data distribution and the Gaussian distribution. Since the divergence of the given data to its nearest Gaussian distribution $D_{\mathrm{KL}}(P_\ell \| \mathcal{N}(\bar{\mu}_\ell, \Sigma^*))$ is fixed for a non-Gaussian dataset, we only need to consider the first term of the right-hand side in the inequality. Considering both the theorems, we can thus further conclude that for any datasets, the performance of $\bar{f}(x)$ is sensitive to the *convergence rate* $\|\bar{\Theta} - \Theta^*\|_2$. Such factor is sometimes known or bounded when an estimator is given, thus are the focuses of this paper.

IV. Proposed Algorithm

In this section, we first introduce a baseline algorithm that lowers the aforementioned error bound using the *sparse but biased* precision matrix estimator with fast convergence rate. Then, we propose our algorithm $\mathcal{DBSDA}$ that further improves the baseline algorithms through the model debiasing. Finally, we discuss the theoretical advantages of the proposed algorithm.

A. Sparse Discriminant Analysis via Graphical Lasso

This baseline algorithm, referred to by SDA via Graphical Lasso, was derived from the *Scout family* of LDA introduced in [24]. Compared to the classical Fisher's LDA presented in Section II-A, this baseline algorithm leverages Graphical Lasso [24] estimator to replace the precision matrix estimated using sample covariance matrix. The proposed algorithm is implemented using the discriminant function defined in (1), as

$$\widehat{f}(x) = \underset{\ell \in \{-,+\}}{\operatorname{argmax}}\ \delta(x, \widehat{\Theta}, \bar{\mu}_\ell, \pi_\ell) \tag{7}$$

where $\widehat{\Theta}$ refers to the Graphical Lasso estimator based on the sample covariance matrix $\bar{\Sigma}$

$$\widehat{\Theta} = \underset{\Theta > 0}{\operatorname{argmin}} \left(\mathrm{tr}(\bar{\Sigma}\Theta) - \log\det(\Theta) + \lambda \sum_{j \ne k} |\Theta_{jk}| \right). \tag{8}$$

According to Theorem 1 and the convergence rate of Graphical Lasso [36], the misclassification rate of $\widehat{f}(x)$ is addressed in Theorem 3.

Theorem 4: Suppose the sample covariance matrix is estimated using m samples drawn i.i.d. from a Gaussian distribution $\mathcal{N}(\mu, \Sigma^*)$, the upper bound of the misclassification rate of $\widehat{f}(x)$ on Gaussian distributions—$\mathcal{N}(\mu_+, \Sigma^*)$ and $\mathcal{N}(\mu_-, \Sigma^*)$—should be

$$\begin{aligned} &P_{\sim\mathcal{N}}[\ell \cdot \widehat{f}(x) < 0] \\ &\quad = \Phi\left(\mathcal{O}\left(\left(\frac{p}{m}\right)^{\frac{1}{2}} - \left(\frac{m^{\frac{1}{2}}}{m^{\frac{1}{2}} + ((p+d)\log\ p)^{\frac{1}{2}}}\right)\right)\right) \end{aligned} \tag{9}$$

with high probability, where the rate $((p+d)\log p/m)^{1/2}$ is derived from the Frobenius-norm convergence rate of Graphical Lasso [36], and $d = \max_{1\le i\le p}|\{j : \Sigma^{*^{-1}}_{i,j} \neq 0\}|$ refers to the maximal degree of the graph (i.e., population inverse covariance matrix).

Proof: According to Zollanvari's model, the misclassification rate of $\widehat{f}(x)$ on the Gaussian data should be

$$P_{\sim\mathcal{N}}[\ell \cdot \widehat{f}(x) < 0] = \varepsilon\big(\bar{\mu}_+, \bar{\mu}_-, \widehat{\Sigma}, \mu^*_+, \mu^*_-, \Sigma^*\big) \tag{10}$$

where $\widehat{\Sigma}^{-1} = \widehat{\Theta}$ is the Graphical Lasso estimator.

According to [36], while the spectral-norm convergence rate is not yet known, the Frobenius-norm convergence rate of Graphical Lasso is known as

$$\begin{aligned} \|\widehat{\Theta}\|_2 &\le \|\Theta^*\|_2 + \|\widehat{\Theta} - \Theta^*\|_2 \le \sqrt{\mathcal{K}} + \|\widehat{\Theta} - \Theta^*\|_F \\ &= \mathcal{O}_p\left(\frac{\sqrt{m} + \sqrt{(p+d)\log p}}{\sqrt{m}}\right) \end{aligned} \tag{11}$$

where $d = \max_{1\le i\le p} |\{j : \Theta^* \neq 0\}|$ refers to the maximal degree of the true graph $\Theta^* = \Sigma^{*-1}$. Then, according to the definition of $\mathcal{O}(\cdot)$, we can conclude

$$\begin{aligned} &P_{\sim\mathcal{N}}[\ell \cdot \widehat{f}(x) < 0] \\ &\quad = \Phi\left(\mathcal{O}\left(\left(\frac{p}{m}\right)^{\frac{1}{2}} - \frac{m^{\frac{1}{2}}}{m^{\frac{1}{2}} + ((p+d)\log p)^{\frac{1}{2}}}\right)\right). \end{aligned} \tag{12}$$

□

B. $\mathcal{DBSDA}$: Debiased Sparse Discriminant Analysis

Intuitively, a sparse estimator, such as SDA aforementioned, can be further improved through debiasing. For example, Lasso can be improved, with even faster convergence rate (lower error bound), by debiased Lasso [28].

In this section, we introduce our proposed algorithm $\mathcal{DBSDA}$(via Graphical Lasso). We first present the linear classifier form of the SDA. Then, we propose a debiased estimator of the linear projection vector (i.e., β), which is derived from the well-known debiased Lasso. Later, we address the overall algorithm design.

The SDA algorithm introduced in (7) can be rewritten in a linear classifier form as

$$\begin{aligned} \widehat{f}(x) &= \mathrm{sign}(\delta(x, \widehat{\Theta}, \bar{\mu}_+, \pi_+) - \delta(x, \widehat{\Theta}, \bar{\mu}_-, \pi_-)) \\ &= \mathrm{sign}(x^T \widehat{\beta}^G + c^g) \end{aligned} \tag{13}$$

where sign$(\cdot)$ function returns $+1$ if the input is nonnegative, and -1 when the input is negative. The vector $\widehat{\beta}^G = \widehat{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)$ and the scalar $c^g = -(1/2) \cdot (\bar{\mu}_+ + \bar{\mu}_-)^T \beta^G + \log(\pi_+/\pi_-)$. Obviously, β^G is the vector of projection coefficients for linear classification.

Inspired by the debiased Lasso [26]–[28], we propose to improve the performance of SDA through debiasing β^G. Given m labeled training data $(x_1, \ell_1), (x_2, \ell_2), \ldots (x_m.\ell_m)$ with balanced labels, the Graphical Lasso estimator $\hat{\Theta}$ on the data and the SDA model (i.e., $\hat{\beta}^G$ and c^g), we propose a novel debiased estimator of $\hat{\beta}^D$ that takes the form as

$$\widehat{\beta}^D \leftarrow \widehat{\beta}^G + \frac{1}{m} \cdot \widehat{\Theta}(\mathbf{X} - \mathbf{U})^T (\mathbf{L} - \mathbf{X}^T \widehat{\beta}^G - \mathbf{C^g}) \tag{14}$$

where we denote $\mathbf{X}$ as an $m \times p$ matrix, where $1 \le i \le m$ and the i^{th} column is x_i; $\mathbf{L}$ as a $1 \times m$ matrix (i.e., vector) whose ith row is ℓ_i; $\mathbf{U}$ is an $m \times p$ matrix with each column as $(\bar{\mu}_+ + \bar{\mu}_-/2)$; and $\mathbf{C^g}$ is a p-dimensional vector with each row as c^g. Note that the debiased estimator addressed in (14) is quite different from debiased Lasso [28], with respect to the structure of LDA as a classifier. We analyze the performance of debiased estimator in Section IV-C.

Given the debiased estimator $\widehat{\beta}^D$, our proposed algorithm $\mathcal{DBSDA}$ is designed as

$$\widehat{f}^D(x) = \mathrm{sign}\left(\left(x^T - \frac{\bar{\mu}_+ + \bar{\mu}_-}{2}\right)^T \widehat{\beta}^D + \log(\pi_+/\pi_-)\right). \tag{15}$$

In Section IV-C, we present the analytical results of $\mathcal{DBSDA}$.

C. Theoretical Properties of $\mathcal{DBSDA}$

In this section, we first present Theorem 4 that provides a upper bound of the misclassification rate of $\mathcal{DBSDA}$. Then, we present Theorem 5 addressing asymptotic properties of $\widehat{\beta}^D$. Finally, we remark the theoretical performance comparison between $\mathcal{DBSDA}$ and SDA.

1) $\mathcal{DBSDA}$—Misclassification Rate Analysis: Further, we aim at analyzing the effect of convergence rates to the upper bound of $\mathcal{DBSDA}$ misclassification rate. Then, we have the following theorem.

Theorem 5: Under the same conditions, the upper bound of $\widehat{f}^D(x)$ misclassification rate on Gaussian distributions $\mathcal{N}(\mu^*_+, \Sigma^*)$ and $\mathcal{N}(\mu^*_-, \Sigma^*)$ (with equal priors) should be

$$\begin{aligned} &P_{\sim\mathcal{N}}[\ell \cdot \widehat{f}^D(x) < 0] \\ &\qquad = \Phi\left(\mathcal{O}\left(\left(\frac{p}{m}\right)^{\frac{1}{2}} - \frac{m^{\frac{1}{2}}}{m^{\frac{1}{2}} + (p\log\ p)^{\frac{1}{2}}}\right)\right) \end{aligned} \tag{16}$$

with high probability.

Lemma 1: Consider the definition of the debiased LDA estimator $\widehat{\beta}^D$ introduced in (14), we have

$$\begin{aligned} \widehat{\beta}^D = \widehat{\beta}^G &+ \frac{1}{m} \cdot \widehat{\Theta}\mathbf{X}^T L - \frac{1}{m} \cdot \widehat{\Theta}\mathbf{U}^T L \\ &- \frac{1}{m} \cdot \widehat{\Theta}(\mathbf{X} - \mathbf{U})^T (\mathbf{X} - \mathbf{U})\widehat{\beta}^G. \end{aligned} \tag{17}$$

As was defined $\widehat{\beta}^G = \widehat{\Theta}(\bar{\mu}_+ - \bar{\mu}_-) = (1/m) \cdot \widehat{\Theta}\mathbf{X}^T L$. With the assumption of equal priors ($\pi_+ = \pi_- = 0.5$), thus $\mathbf{L}$ is a label vector that half of its elements are $+1$ while the rest are all -1. As $\mathbf{U}$ is matrix where each column is a constant vector $(\bar{\mu}_+ + \bar{\mu}_-)/2$, thus $(1/m) \cdot \widehat{\Theta}\mathbf{U}^T L = \mathbf{0}$. As each column of $\mathbf{X}$ refers to a sample drawn from the original data distribution, thus $(1/m)(\mathbf{X} - \mathbf{U})^T(\mathbf{X} - \mathbf{U}) = \bar{\Sigma}$ is the sample covariance matrix estimator. With all above in mind, we have

$$\widehat{\beta}^D = \widehat{\beta}^G + (\mathbf{I} - \widehat{\Theta}\bar{\Sigma})\widehat{\beta}^G = (2\widehat{\Theta} - \widehat{\Theta}\bar{\Sigma}\widehat{\Theta})\bar{\Delta}_\mu \qquad (18)$$

where $\mathbf{I}$ refers to a $p \times p$ identity matrix. Note that $(\mathbf{I} - \widehat{\Theta}\bar{\Sigma})\widehat{\beta}^G$ can be considered the desparsification term that debiases $\widehat{\beta}^G$ through adjusting the Karush–Kuhn–Tucker (K.K.T) condition given the Graphical Lasso estimator.

Proof: As was mentioned in Lemma1, $\mathcal{DBSDA}$ indeed can be considered as an LDA classifier that leverages Θ^D as its precision matrix estimator and $\widehat{\Theta}^D = (2 \cdot \widehat{\Theta} - \widehat{\Theta}\bar{\Sigma}\widehat{\Theta})$. Considering the known Frobenius-norm convergence rate

$$\begin{aligned} \|\widehat{\Theta}^D\|_2 &\le \|\Theta^*\|_2 + \|\widehat{\Theta}^D - \Theta^*\|_2 \le \sqrt{\mathcal{K}} + \|\widehat{\Theta} - \Theta^*\|_F \\ &= \mathcal{O}_p\sqrt{p \log p/m}. \end{aligned} \qquad (19)$$

According to the definition of $\mathcal{O}(\cdot)$, we can obtain the result

$$\begin{aligned} &P_{\sim\mathcal{N}}[\ell \cdot \widehat{f}^D(x) < 0] \\ &\quad = \Phi\left(\mathcal{O}\left(\left(\frac{p}{m}\right)^{\frac{1}{2}} - \frac{m^{\frac{1}{2}}}{m^{\frac{1}{2}} + (p \log\ p)^{\frac{1}{2}}}\right)\right) \end{aligned} \qquad (20)$$

with high probability. □

2) $\mathcal{DBSDA}$—Asymptotic Analysis: In order to analyze the performance of $\mathcal{DBSDA}$, we first define the linear projection vector of the optimal LDA as $\beta^* = \Sigma^{*-1}(\mu^*_+ - \mu^*_-)$, then we intend to understand how close $\widehat{\beta}^G$ and $\widehat{\beta}^D$ approximate to the optimal estimation β^*. Here, we continue assuming the population mean vectors μ^*, μ^*_+, and μ^*_- are estimated as sample mean vectors $\bar{\mu}$, $\bar{\mu}_+$, and $\bar{\mu}_-$, and have the following theorem:

Theorem 6: Given the m samples for training $(x_1, \ell_1), \ldots, (x_m, \ell_m)$ drawn i.i.d. from $\mathcal{N}(\mu^*_+, \Sigma^*)$ and $\mathcal{N}(\mu^*_-, \Sigma^*)$ with the equal priors, the ℓ_∞-vector-norm convergence rate of $\widehat{\beta}^D$ and $\widehat{\beta}^G$ approximating to the optimal estimation β^* are

$$\begin{aligned} |\widehat{\beta}^D - \beta^*|_2 &= \mathcal{O}_p(\sqrt{p \log p/m}) \\ |\widehat{\beta}^G - \beta^*|_2 &= \mathcal{O}_p(\sqrt{(p + d) \log p/m}) \end{aligned} \qquad (21)$$

under the assumption that $|\mu^*_+ - \mu^*_-|_2$ is estimated as $|\bar{\mu}_+ - \bar{\mu}_-|_2$ and $|\mu^*_+ - \mu^*_-|_2$ is assumed a constant.

Proof: Here, we first prove the upper bound of $|\widehat{\beta}^G - \beta^*|_\infty$. As was defined $\widehat{\beta}^G = \widehat{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)$, then we have

$$\begin{aligned} |\widehat{\beta}^G - \beta^*|_2 &= \left|\widehat{\Theta}(\bar{\mu}_+ - \bar{\mu}_-) - \Theta^*\left(\mu^*_+ - \mu^*_-\right)\right|_2 \\ &\le \left|(\widehat{\Theta} - \Theta^*)\bar{\Delta}_\mu|_2 + |\Theta^*\left(\mu^*_+ - \bar{\mu}_+\right)\right|_2 \\ &\quad + \left|\Theta^*\left(\mu^*_- - \bar{\mu}_-\right)\right|_2 \\ &\le \|\widehat{\Theta} - \Theta^*\|_2 \cdot |\bar{\Delta}_\mu|_2 \\ &\quad + \|\Theta^*\|_2 \cdot \left(\left|\mu^*_+ - \bar{\mu}_+\right|_2 + \left|\mu^*_- - \bar{\mu}_-\right|_2\right). \end{aligned} \qquad (22)$$

Since 1) both $\bar{\mu}_+$ and $\bar{\mu}_-$ converge at the rate $\mathcal{O}_p(\sqrt{p/m})$ in ℓ_2-norm; 2) the spectral-norm convergence rate of $\widehat{\Theta}$ [32] is $\|\widehat{\Theta} - \Theta^*\|_2 \le \|\widehat{\Theta} - \Theta^*\|_F = \mathcal{O}_p(((p + d) \cdot \log p/m)^{1/2})$; and 3) $\bar{\Delta}_\mu \le 2\mathcal{L}$ is bounded by a constant, we thus can conclude that, with high probability

$$|\widehat{\beta}^G - \beta^*|_2 = \mathcal{O}_p(\sqrt{(p + d) \cdot \log p/m}). \qquad (23)$$

□

Proof: Given Lemma 1, we prove the upper bound of $|\widehat{\beta}^D - \beta^*|_2$ as

$$\begin{aligned} |\widehat{\beta}^D - \beta^*|_2 &= \left|(2\widehat{\Theta} - \widehat{\Theta}\bar{\Sigma}\widehat{\Theta})(\bar{\mu}_+ - \bar{\mu}_-) - \Theta^*\left(\mu^*_+ - \mu^*_-\right)\right|_2 \\ &\le |(2\widehat{\Theta} - \widehat{\Theta}\bar{\Sigma}\widehat{\Theta} - \Theta^*)(\bar{\mu}_+ - \bar{\mu}_-)|_2 \\ &\quad + \left|\Theta^*\left(\mu^*_+ - \bar{\mu}_+\right)\right|_2 + \left|\Theta^*\left(\mu^*_- - \bar{\mu}_-\right)\right|_2 \\ &\le \|2\widehat{\Theta} - \widehat{\Theta}\bar{\Sigma}\widehat{\Theta} - \Theta^*\|_2 \cdot |\bar{\mu}_+ - \bar{\mu}_-|_2 \\ &\quad + \|\Theta^*\|_2 \cdot \left(\left|\bar{\mu}_+ - \mu^*_+\right|_2 + \left|\bar{\mu}_- - \mu^*_-\right|_2\right). \end{aligned} \qquad (24)$$

According to [33], the spectral-norm convergence rate of the desparisified estimator $\widehat{\Theta}^D = (2 \cdot \widehat{\Theta} - \widehat{\Theta}\bar{\Sigma}\widehat{\Theta})$ under mild conditions should be $\|\widehat{\Theta}^D - \Theta^*\|_2 \le \sqrt{p} \cdot \|\widehat{\Theta}^D - \Theta^*\|_\infty = \mathcal{O}_p(p \cdot \log p/m)^{1/2}$. In this way, considering Assumption 1, we conclude the convergence rate as

$$|\widehat{\beta}^D - \beta^*|_2 = \mathcal{O}_p(\sqrt{p \log p/m}). \qquad (25)$$

□

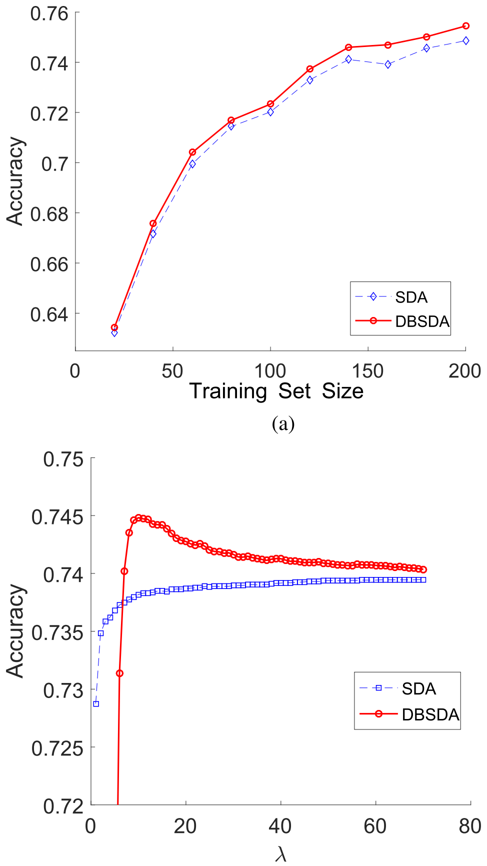
Note that, to highlight the effect of precision matrix to the accuracy of classification, throughout the paper, we make no assumptions on the mean vectors. We consider the sample estimation $\bar{\mu}$, $\bar{\mu}_+$, and $\bar{\mu}_-$ as the mean vectors μ^*, μ^*_+, and μ^*_- in the Gaussian distribution. It is quite often in multivariate statistics to follow such settings [31].

Remark 2: Compared to SDA's $\widehat{\beta}^G$, our method $\mathcal{DBSDA}$ recovers the linear projection vector $\widehat{\beta}^D$ with a faster convergence rate, i.e., $(\log p/m)^{1/2} < ((p + d) \cdot \log p/m)^{1/2}$ in a mild condition. Thus, we can conclude that $\mathcal{DBSDA}$ outperforms SDA with reduced upper bounds of ℓ_∞ vector norm estimation errors. We name $\widehat{\beta}^D$ as the debiased estimator of $\widehat{\beta}^G$ due to following reasons: with assumptions addressed in Theorem 4, we can rewrite $\widehat{\beta}^G = \widehat{\Theta}(\bar{\mu}_+ - \bar{\mu}_-)$ and $\widehat{\beta}^D = (2 \cdot \widehat{\Theta} - \widehat{\Theta}\bar{\Sigma}\widehat{\Theta})(\bar{\mu}_+ - \bar{\mu}_-)$, while $(2 \cdot \widehat{\Theta} - \widehat{\Theta}\bar{\Sigma}\widehat{\Theta})$ is the debiased estimator of $\widehat{\Theta}$, according to [33].

Remark 3: In terms of misclassification rate comparison, $\mathcal{O}_p((p \log p/m)^{1/2}) < \mathcal{O}_p(((p + d) \log p/m)^{1/2})$ while the $B = \|(I - \widehat{\Theta}\bar{\Sigma})\widehat{\Theta}\|_2$ is not fully known. However, $\widehat{\Theta}\bar{\Sigma}$ should be very close to the identity matrix I, considering the K.K.T condition in (8). In this way, $\widehat{f}^D(x)$ can outperform $\widehat{f}^G(x)$ with lower misclassification rate.

V. Experiments

In this section, we first validate different properties of $\mathcal{DBSDA}$ on the synthesized data, from which we can gain insight into the superiority of $\mathcal{DBSDA}$. Then, we experimentally evaluate the performance of $\mathcal{DBSDA}$ on benchmark datasets, in terms of the accuracy for binary classification. Finally, we demonstrate the performance of $\mathcal{DBSDA}$ on real-world HDLSS dataset for the application of diseases early detection.



(a)

(b)

Fig. 1. Classification accuracy of $\mathcal{DBSDA}$ versus SDA on pseudorandom synthesized data. (a) $\mathcal{DBSDA}$ versus SDA. (b) λ Tunning.

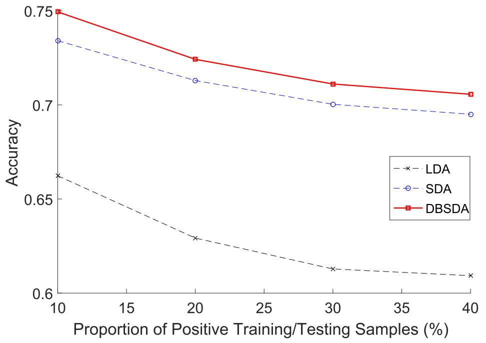


Fig. 2. Classification accuracy of $\mathcal{DBSDA}$ versus SDA on unbalanced datasets ($m = 160$).

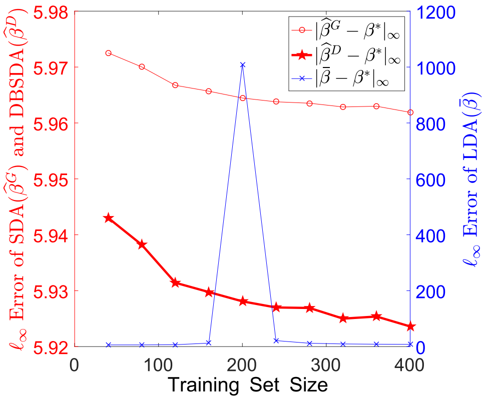


Fig. 3. Asymptotic properties of $\mathcal{DBSDA}$ versus SDA on pseudorandom synthesized data.

A. Synthesized Data Evaluation

To validate our algorithms, we evaluate our algorithms on a synthesized dataset, which are obtained through a pseudorandom simulation. The synthetic data are generated by two predefined Gaussian distributions $\mathcal{N}(\mu^*_+, \Sigma^*)$ and $\mathcal{N}(\mu^*_-, \Sigma^*)$ with equal priors. The settings of μ^*_+, μ^*_-, and Σ^* are as follows: Σ^* is a $p \times p$ symmetric and positive definite matrix, where each element $\Sigma^*_{i,j} = 0.8^{|i-j|}$, $1 \leq i \leq p$ and $1 \leq j \leq p$. μ^*_+ and μ^*_- are both p-dimensional vectors, where $\mu^*_+ = \langle 1, 1, \ldots, 1, 0, 0, \ldots, 0\rangle^T$ (the first 10 elements are all 1's, while the rest $p-10$ elements are 0's) and $\mu^*_- = \mathbf{0}$. (Settings of the two Gaussian distributions first appear in [37].) In our experiment, we set $p = 200$. To simulate the HDLSS settings, we train SDA and $\mathcal{DBSDA}$, with 20–200 samples randomly drawn from the distributions with equal priors, and test the two algorithms using 500 samples. For each setting, we repeat the experiments for 100 times and take the averaged results, under the aforementioned cross-validation procedure.

In this experiment, we compare $\mathcal{DBSDA}$, SDA, and LDA (with pseudoinverse). The results of LDA are not included here, as it performs extremely worse than both SDA and $\mathcal{DBSDA}$ under the HDLSS settings. Fig. 1(a) presents the comparison between $\mathcal{DBSDA}$ and SDA, in terms of accuracy, where each algorithm is fine-tuned with the best parameter λ. A detailed example of parameter tuning is reported in Fig. 1(b), where we run both the algorithms, with training set size as 160, when varying λ from 1 to 70. From Fig. 1(a), it is obvious that $\mathcal{DBSDA}$ outperforms SDA marginally. The λ tuning comparison addressed in Fig. 1(b) shows that, given a small λ, both SDA and $\mathcal{DBSDA}$ cannot perform well, as the sparse approximation of $\widehat{\beta}^G$ and $\widehat{\beta}^D$ cannot be well recovered in such case [24], [33]. When $\lambda \geq 6$, $\mathcal{DBSDA}$ starts outperforming SDA, while the advantage of $\mathcal{DBSDA}$ to SDA decreases when increasing λ. However, even with an extremely large λ, $\mathcal{DBSDA}$ still outperforms SDA. In Fig. 2, we present the evaluation results based on unbalanced datasets, where the accuracy of algorithms using $m = 160$ training samples drawn with varying priors is illustrated.

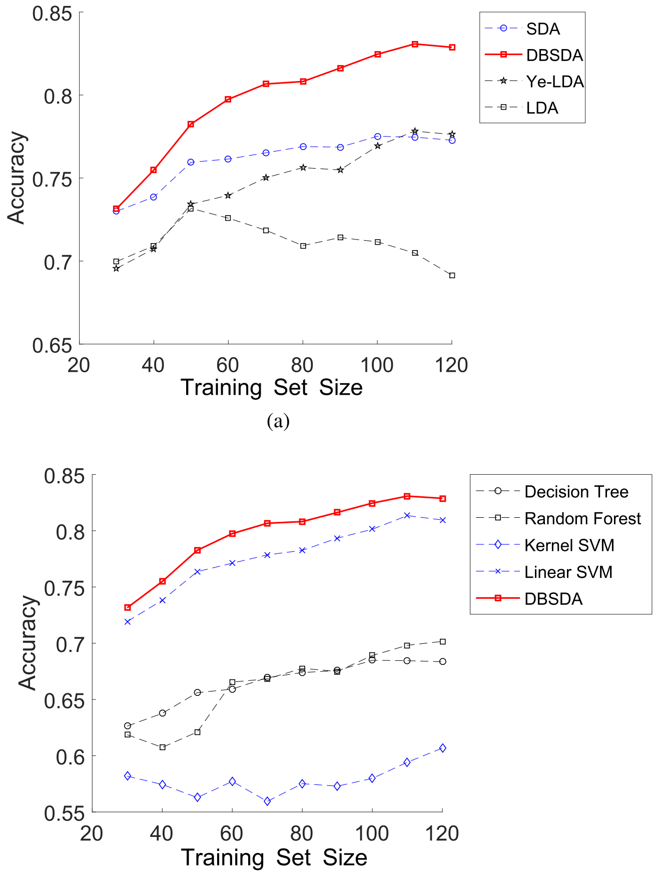
To further verify our algorithms, we propose the optimal LDA classifier $\beta^* = \Theta^*(\mu^*_+ - \mu^*_-)$, which is all based on the population parameters. We compare $\widehat{\beta}^D$, $\widehat{\beta}^G$, and $\bar{\beta}$ estimated by $\mathcal{DBSDA}$, SDA, and LDA (with pseudoinverse) to β^*. Fig. 3 presents the comparison among $|\widehat{\beta}^D - \beta^*|_2$, $|\widehat{\beta}^G - \beta^*|_2$, and $|\bar{\beta} - \beta^*|_2$. It is obvious that $\widehat{\beta}^D$ is more close to β^* than $\widehat{\beta}^G$ and $\bar{\beta}$. This observation further verifies Theorem 4. We also compare the accuracy of β^* to SDA, $\mathcal{DBSDA}$ and LDA. β^* outperform these algorithms and the accuracy of β^* is around 84.4%. It is reasonable to conclude that $\mathcal{DBSDA}$ outperforms SDA, because $\widehat{\beta}^D$ is more close to β^*.

B. Benchmark Evaluation Results

In Fig. 4(a), we compare $\mathcal{DBSDA}$ and other LDA algorithms, including LDA with pseudoinverse, SDA via

TABLE II
EARLY DETECTION OF DISEASES ACCURACY COMPARISON BETWEEN $\mathcal{DBSDA}$ AND BASELINES

	Training Set Size						
Algorithm	100	200	300	400	500	600	700
$\mathcal{DBSDA}$	**0.659±0.022**	**0.677±0.028**	**0.691±0.024**	**0.692±0.023**	**0.690±0.021**	**0.696±0.024**	**0.701±0.023**
LDA	0.543±0.034	0.586±0.033	0.616±0.022	0.642±0.029	0.642±0.022	0.657±0.025	0.658±0.026
Ye-LDA	0.627±0.050	0.620±0.077	0.652±0.063	0.620±0.067	0.655±0.062	0.637±0.064	0.670±0.045
Decision Tree	0.621±0.046	0.649±0.031	0.652±0.041	0.655±0.030	0.671±0.028	0.665±0.031	0.668±0.040
Linear SVM	0.615±0.026	0.628±0.030	0.647±0.023	0.666±0.029	0.666±0.021	0.670±0.030	0.675±0.029
Kernel SVM	0.635±0.032	0.669±0.027	0.674±0.039	0.678±0.021	0.668±0.038	0.688±0.024	0.682±0.029
AdaBoost	0.631±0.035	0.630±0.039	0.620±0.028	0.622±0.027	0.621±0.022	0.617±0.025	0.626±0.070
SDA	0.658±0.023	0.676±0.024	0.682±0.028	0.686±0.022	0.683±0.021	0.692±0.025	0.695±0.018
Random Forest	0.590±0.035	0.602±0.035	0.653±0.031	0.602±0.040	0.674±0.024	0.666±0.026	0.658±0.032



(a)

(b)

Fig. 4. Performance comparison on benchmark datasets ($p = 300$ and $p \gg m$). (a) $\mathcal{DBSDA}$ versus LDA baselines. (b) $\mathcal{DBSDA}$ versus downstream classifiers.

Graphical Lasso, and Ye-LDA derived from [38], on the Web datasets [39]. To simulate the HDLSS settings ($p \gg m$), we vary the training sample sizes from 30 to 120 while using 400 samples for testing. The numbers of dimensions p is 300. For each algorithm, the reported result is averaged over 100 randomly selected subsets of the training/testing data with equal priors. SDA and $\mathcal{DBSDA}$ are fine-tuned with the best λ. The experimental settings show that $\mathcal{DBSDA}$ consistently outperforms other competitors in different settings. The nonmonotonic trend of LDA with the increasing training set size is partially due to the poor/uncontrollable performance of pseudoinverse used in LDA. Note that the whole Web dataset consists of three subsets (Web-1, Web-2, and Web-3). Due to space limitation, we only report the results on Web-1 in Fig. 4(a). Similar results can be observed on Web-2 and Web-3.

In addition to LDA classifiers, we also compared $\mathcal{DBSDA}$ with other downstream algorithms including *Decision Tree*, *Random Forest*, *Linear Support Vector Machine (SVM)*, and *Kernel SVM with Gaussian Kernel*. The comparison results are listed in Fig. 4(b). All algorithms are fine-tuned with the best parameters under our experiment settings (under cross validation).

C. Early Detection of Diseases on EHR Datasets

To demonstrate the effectiveness of $\mathcal{DBSDA}$ in handling the real problems, we evaluate $\mathcal{DBSDA}$ on the real-world Electronic Health Records (EHRs) data for early detection of diseases [40]. In this application, each patient's EHR data are represented by a $p = 295$ dimensional vector, referring to the outpatient record on the physical disorders diagnosed. Patients are labeled with either "positive" or "negative," indicating whether he/she was diagnosed with depression and anxiety disorders. Through supervised learning on the datasets, the trained binary classifier is expected to predict whether a (new) patient is at risk or would develop to the depression and anxiety disorders from their historical outpatient records (physical disorder records) [41].

We evaluate $\mathcal{DBSDA}$ and other competitors, including linear SVM, nonlinear SVM with Gaussian kernel, decision tree, AdaBoost, random forest, and other LDA baselines, with varying training dataset size m from 100 to 700. Table II presents the comparison results. To simplify the comparison, we only present the results of the algorithm with fine-tuned parameter, which is selected through 10-fold cross validation. It is obvious that $\mathcal{DBSDA}$ and SDA outperform other baseline algorithms significantly, while $\mathcal{DBSDA}$ performs better than SDA. The advantage of $\mathcal{DBSDA}$ over other algorithms, such as SVM, is extremely obvious when the size of a training dataset m is small. With the increasing sample size, though

TABLE III

EARLY DETECTION OF DISEASES F1-SCORE COMPARISON BETWEEN $\mathcal{DBSDA}$ AND OTHER BASELINES

	Training Set						
Algorithm	100	200	300	400	500	600	700
$\mathcal{DBSDA}$	0.690±0.028	0.708±0.027	**0.722±0.024**	**0.729±0.018**	**0.727±0.0118**	**0.736±0.018**	**0.734±0.022**
LDA	0.539±0.048	0.580±0.044	0.611±0.030	0.646±0.027	0.644±0.025	0.662±0.028	0.663±0.032
Ye-LDA	0.644±0.100	0.657±0.124	0.688±0.071	0.678±0.057	0.698±0.035	0.698±0.035	0.712±0.027
Decision Tree	0.626±0.120	0.671±0.074	0.675±0.088	0.703±0.032	0.695±0.034	0.676±0.078	0.690±0.097
Linear SVM	0.616±0.031	0.627±0.041	0.651±0.026	0.675±0.031	0.675±0.026	0.680±0.035	0.690±0.031
Kernel SVM	**0.701±0.063**	**0.723±0.022**	0.702±0.115	0.726±0.016	0.681±0.115	0.734±0.019	0.715±0.071
AdaBoost	0.560±0.081	0.533±0.107	0.498±0.065	0.503±0.078	0.500±0.080	0.482±0.066	0.503±0.070
SDA	0.696±0.021	0.716±0.021	0.719±0.024	0.725±0.018	0.721±0.015	0.733±0.021	0.734±0.016
Random Forest	0.419±0.126	0.509±0.102	0.613±0.067	0.509±0.110	0.661±0.036	0.640±0.058	0.603±0.063

the margins of $\mathcal{DBSDA}$ over the rest of algorithms decrease, $\mathcal{DBSDA}$ still outperforms other algorithms. We also measured the F1-score of all algorithms, $\mathcal{DBSDA}$ still outperforms other competitors in most cases, which is shown in Table III.

VI. Related Work and Discussion

In this section, we review several most relevant studies of our research. To address the HDLSS issues for LDA, a line of research [21], [22], [42], [43] proposed to directly estimate a sparse projection vector without estimating the inverse covariance matrix (sample covariance matrix is not invertible) and mean vectors separately. On the other hand, Peck and Ness [3], Witten and Tibshirani [24], and Bickel and Levina [44] proposed to first estimate the inverse covariance matrix through shrunken covariance estimators, and then estimate the projection vector with sample mean vectors. Through regularizing the (inverse) covariance matrix estimation, these algorithms are expected to estimate a sparse projection vector with (sub)optimal discrimination power [9].

In our paper, we focus on improving covariance-regularized LDA [24] through debiasing the projection vector estimated using Graphical Lasso [36]. Our work is distinct due to the following reasons: 1) our work is the first to study the problem of debiasing the SDA [26]–[28]; 2) compared to the existing solution to the debiased linear regression models [28], we proposed a novel debiased estimator (using a different formulate in (14)) for the *covariance-regularized SDA* [24], [36]; 3) we analyzed the debiased estimator and obtained the upper bound of its misclassification rate (based on our main theory) as well as its asymptotic properties; and 4) we validate our algorithms through comparing a wide range of baselines on both the synthesized and real-world datasets, where the evaluation result backups our theory (e.g., asymptotic properties proved in Theorem 4 versus the curve shown in Fig 3).

In our future work, we intend to study the performance of $\mathcal{DBSDA}$ for feature extraction [45], [46], and metrics/ representation learning [47]. Though our work study the asymptotic property of the estimator under the *IID sampling* assumption, we plan to study the scheme to further improve LDA with faster rate leveraging other sampling strategies [48]. In addition to debias the covariance-regularized LDA, we plan to study the debiased estimators for other SDA [21], [22], [49], [50] under HDLSS settings. In addition to Fisher's discriminant analysis that relies on the estimation of the inverse covariance matrix, our future work intends to model the performance of a tensor discriminant analysis that preserves higher order discriminant information in tensors [51], as well as the performance of other projection methods [52] that learn linear subspace for optimal classification.

VII. Conclusion

In this paper, we extend the existing theory [9], [30] and propose a novel analytic model characterizing the misclassification upper bound of LDA under uncertainty of inverse covariance matrix estimation. Based on the analytic model, we analyzed the misclassification rate of SDA, and proposed $\mathcal{DBSDA}$—a novel $\mathcal{D}$e-$\mathcal{B}$iased $\mathcal{S}$parse $\mathcal{D}$iscriminant $\mathcal{A}$nalysis classifier that reduces the upper bound of LDA misclassification rate through debiasing the shrunken (sparse) estimator [24]. Our analysis shows that $\mathcal{DBSDA}$ is with a reduced upper bound of misclassification rate and better asymptotic properties, compared to SDA, under HDLSS settings. The experimental results on synthesized and real-world datasets show $\mathcal{DBSDA}$ outperformed all baseline algorithms. Furthermore, the empirical studies on estimator comparison validate our theoretical analysis.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, USA: Wiley, 2001.

[2] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, 2013.

[3] R. Peck and J. Van Ness, "The use of shrinkage estimators in linear discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 5, pp. 530–537, Sep. 1982.

[4] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[5] T. T. Cai, Z. Ren, and H. H. Zhou, "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation," *Electron. J. Statist.*, vol. 10, no. 1, pp. 1–59, 2016.

[6] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math. USSR-Sbornik*, vol. 1, no. 4, p. 457, 1967.

[7] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Statist.*, vol. 29, no. 2, pp. 295–327, 2001.

[8] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, "Analytic study of performance of error estimators for linear discriminant analysis," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4238–4255, Sep. 2011.

[9] A. Zollanvari and E. R. Dougherty, "Random matrix theory in pattern classification: An application to error estimation," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2013, pp. 884–887.

[10] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, "Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data," *Appl. Statist.*, vol. 44, no. 1, pp. 101–115, 1995.

[11] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

[12] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. NIPS*, Cambridge, MA, USA, 2004, pp. 1569–1576.

[13] N. D. Lawrence and B. Schölkopf, "Estimating a Kernel Fisher discriminant in the presence of label noise," in *Proc. ICML*, vol. 1, 2001, pp. 306–313.

[14] Z. Zhang, "Learning metrics via discriminant kernels and multidimensional scaling: Toward expected Euclidean representation," in *Proc. ICML*, vol. 2, 2003, pp. 872–879.

[15] S.-J. Kim, A. Magnani, and S. Boyd, "Optimal kernel selection in kernel Fisher discriminant analysis," in *Proc. ICML*, 2006, pp. 465–472.

[16] Z. Zhang, G. Dai, C. Xu, and M. I. Jordan, "Regularized discriminant analysis, ridge regression and beyond," *J. Mach. Learn. Res.*, vol. 11, pp. 2199–2228, Aug. 2010.

[17] S. Kaski and J. Peltonen, "Informative discriminant analysis," in *Proc. ICML*, 2003, pp. 329–336.

[18] C. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and K-means clustering," in *Proc. ICML*, 2007, pp. 521–528.

[19] R. He, B.-G. Hu, and X.-T. Yuan, "Robust discriminant analysis based on nonparametric maximum entropy," in *Proc. Asian Conf. Mach. Learn.*. Berlin, Germany: Springer, 2009, pp. 120–134.

[20] M. Chen, W. Carson, M. Rodrigues, L. Carin, and R. Calderbank, "Communications inspired linear discriminant analysis," in *Proc. ICML*, 2012, pp. 1–8.

[21] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[22] T. Cai and W. Liu, "A direct estimation approach to sparse linear discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 106, no. 496, pp. 1566–1577, 2011.

[23] Q. Mai, H. Zou, and M. Yuan, "A direct approach to sparse discriminant analysis in ultra-high dimensions," *Biometrika*, vol. 99, no. 1, pp. 29–42, 2012.

[24] D. M. Witten and R. Tibshirani, "Covariance-regularized regression and classification for high dimensional problems," *J. Roy. Statist. Soc. B*, vol. 71, no. 3, pp. 615–636, Jun. 2009.

[25] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.

[26] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 76, no. 1, pp. 217–242, 2014.

[27] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal confidence regions and tests for high-dimensional models," *Ann. Statist.*, vol. 42, no. 3, pp. 1166–1202, 2014.

[28] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2869–2909, 2014.

[29] Š. Raudys and D. M. Young, "Results in statistical discriminant analysis: A review of the former Soviet Union literature," *J. Multivariate Anal.*, vol. 89, no. 1, pp. 1–35, Apr. 2004.

[30] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, Feb. 1968.

[31] C. Field, "Small sample asymptotic expansions for multivariate M-estimates," *Ann. Statist.*, vol. 10, no. 3, pp. 672–689, Sep. 1982.

[32] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electron. J. Statist.*, vol. 2, pp. 494–515, Jun. 2008.

[33] J. Janková and Sara van de Geer, "Confidence intervals for high-dimensional inverse covariance estimation," *Electron. J. Statist.*, vol. 9, no. 1, pp. 1205–1229, 2015.

[34] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[35] G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller, "In search of non-Gaussian components of a high-dimensional distribution," *J. Mach. Learn. Res.*, vol. 7, pp. 247–282, Feb. 2006.

[36] D. M. Witten, J. H. Friedman, and N. Simon, "New insights and faster computations for the graphical lasso," *J. Comput. Graph. Statist.*, vol. 20, no. 4, pp. 892–900, 2011.

[37] L. Tian, B. Jayaraman, Q. Gu, and D. Evans, "Aggregating private sparse learning models using multi-party computation," in *Proc. NIPS Workshop Private Multi-Party Mach. Learn.*, Barcelona, Spain, 2016.

[38] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 982–994, Aug. 2004.

[39] J. C. Platt, "12 fast training of support vector machines using sequential minimal optimization," *Adv. Kernel Methods*, vol. 1, pp. 185–208, 1999.

[40] J. C. Turner and A. Keller, "College Health Surveillance Network: Epidemiology and health care utilization of college students at US 4-year universities," *J. Amer. College Health*, vol. 63, no. 8, pp. 530–538, Jun. 2015.

[41] J. Zhang, H. Xiong, Y. Huang, H. Wu, K. Leach, and L. E. Barnes, "M-SEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data," in *Proc. BigData*, 2015, pp. 2569–2577.

[42] Z. Qiao, L. Zhou, and J. Z. Huang, "Effective linear discriminant analysis for high dimensional, low sample size data," in *Proc. World Congr. Eng.*, vol. 2, 2008, pp. 2–4.

[43] J. Shao, Y. Wang, X. Deng, and S. Wang, "Sparse linear discriminant analysis by thresholding for high dimensional data," *Ann. Statist.*, vol. 39, no. 2, pp. 1241–1265, 2011.

[44] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Statist.*, vol. 36, no. 1, pp. 199–227, 2008.

[45] Y. Hou, I. Song, H.-K. Min, and C. H. Park, "Complexity-reduced scheme for feature extraction with linear discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 1003–1009, Jun. 2012.

[46] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, "Effective discriminative feature selection with nontrivial solution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 796–808, Apr. 2016.

[47] A. Iosifidis, A. Tefas, and I. Pitas, "On the optimal class representation in linear discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1491–1497, Sep. 2013.

[48] B. Zou, L. Li, Z. Xu, T. Luo, and Y. Y. Tang, "Generalization performance of Fisher linear discriminant based on Markov sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 2, pp. 288–300, 2013.

[49] J. Zhao, L. Shi, and J. Zhu, "Two-stage regularized linear discriminant analysis for 2-D data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1669–1681, Aug. 2015.

[50] X. Zhang, D. Chu, and R. C. E. Tan, "Sparse uncorrelated linear discriminant analysis for undersampled problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 7, pp. 1469–1485, Jul. 2016.

[51] Z. Lai, Y. Xu, J. Yang, J. Tang, and D. Zhang, "Sparse tensor discriminant analysis," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3904–3915, Oct. 2013.

[52] Z. Lai, W. K. Wong, Y. Xu, J. Yang, and D. Zhang, "Approximate orthogonal sparse embedding for dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 723–735, Apr. 2016.

Haoyi Xiong (S'12–M'15) received the B.Eng. degree in electrical engineering and automation from the Huazhong University of Science and Technology, Wuhan, China, in 2009, the M.Sc. degree in information technology from The Hong Kong University of Science and Technology, Hong Kong, in 2010, and the Ph.D. degree in computer science from Telecom SudParis, Evry, France, jointly with Pierre and Marie Curie University (Paris VI), Paris, France, in 2015.

From 2015 to 2016, he was a Post-Doctoral Research Associate with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA. He is currently a Senior Research Scientist with the Beijing Big Data Laboratory, Baidu Research, Beijing. He is also affiliated with the National Engineering Laboratory of Deep Learning Application and Technology, Beijing. Prior to joining Baidu, he was an Assistant Professor (tenure-track) with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA. He has published extensively in a series of top conferences and journals, including the ACM International Joint Conference on Pervasive and Ubiquitous Computing (ACM UbiComp), the AAAI Conference on Artificial Intelligence, the International Joint Conference on Artificial Intelligence, the IEEE International Conference on Data Mining, the IEEE International Conference on Pervasive Computing and Communications, and IEEE/ACM TRANSACTIONS. His current research interests include applied machine learning and ubiquitous computing.

Dr. Xiong was a recipient of the Best Paper Award from the 9th IEEE International Conference on Ubiquitous Intelligence and Computing (IEEE UIC), Fukuoka, Japan, in 2012, the 2015 Outstanding Ph.D. Thesis Runner-Up Award from the SAMOVAR Lab (UMR 5157) at the French National Science Research Center (CNRS), Evry, France, and the Excellent Service Award from IEEE UIC'17, San Francisco, CA, USA. He was also a co-receipt of several miscellaneous awards. He served as referees for a number of excellent conference and journals, including ACM UbiComp, the *Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *ACM Transactions on Intelligent Systems and Technology*, *ACM Transactions on Multimedia*, and the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, the IEEE TRANSACTIONS ON BIG DATA, the IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS: SYSTEMS, the *IEEE Communications Magazine*. His research career has been generously supported and financed by the EU FP7 Program, the Paris-Saclay Pole Systematic Program, the UVA HobbyâĽ™s Fellowship of Computational Sciences, UMSystem Funds, and Baidu Research.

Wei Cheng (S'10–M'15) received the B.S. degree from the School of Software, Nanjing University, Nanjing, China, in 2006, the M.Sc. degree from the School of Software, Tsinghua University, Beijing, China, in 2010, and the Ph.D. degree from the Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, in 2015.

He is currently a Research Staff Member with the Data Science Department, NEC Laboratories America, Inc., Princeton, NJ, USA. His current research interests include data science, machine learning, web applications, and bioinformatics.

Jiang Bian (S'16) received the B.Eng. degree in logistics engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014, and the M.Sc. degree in industrial engineering from the University of Florida, Gainesville, FL, USA in 2016. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA.

He is currently a Research Intern with the Beijing Big Data Laboratory, Baidu Research, Beijing, China. He has published extensively in a series of top conferences, including the AAAI Conference on Artificial Intelligence (AAAI), the International Joint Conference on Artificial Intelligence, and the IEEE International Conference on Data Mining. His current research interests include ubiquitous computing and distributed learning algorithms.

Mr. Bian served as a volunteer in AAAI-18 and awarded travel grant by the committee.

Wenqing Hu received the B.S. degree from the School of Mathematical Science, Peking University, Beijing, China, in 2008, and the Ph.D. degree in mathematics from the Department of Mathematics, University of Maryland at College Park, College Park, MD, USA, in 2016.

He was a Post-Doctoral Associate with the School of Mathematics, University of Minnesota Twin Cities, Minneapolis, MN, USA, from 2013 to 2016. He is currently an Assistant Professor of mathematics with the Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO, USA. His current research interests include probability theory, stochastic analysis, differential equations, dynamical systems, mathematical physics, and statistical methodology.

Zeyi Sun received the B.Eng. degree in material science and engineering from Tongji University, Shanghai, China, in 2002, the M.Phil. degree in manufacturing from the University of Michigan, Ann Arbor, MI, USA, in 2010, and the Ph.D. degree in industrial engineering and operation research from the University of Illinois at Chicago, Chicago, IL, USA, in 2015.

He is currently an Assistant Professor with the Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO, USA. His current research interests include energy efficiency management of manufacturing systems, electricity demand response of manufacturing systems, system modeling of cellulosic biofuel manufacturing, energy modeling, and control in additive manufacturing and intelligent maintenance of manufacturing systems.

Zhishan Guo (S'10–M'16) received the B.Eng. degree in computer science and technology from Tsinghua University, Beijing, China, in 2009, the M.Phil. degree in mechanical and automation engineering from The Chinese University of Hong Kong, Hong Kong, in 2011, and the Ph.D. degree in computer science from the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, in 2016.

He is currently an Assistant Professor with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA. His current research interests include real-time and cyber-physical systems, neural networks, and computational intelligence.