



Improving covariance-regularized discriminant analysis for EHR-based predictive analytics of diseases

Sijia Yang¹ · Haoyi Xiong² · Kaibo Xu³ · Licheng Wang¹ · Jiang Bian⁴ · Zeyi Sun³

Published online: 13 August 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Linear Discriminant Analysis (LDA) is a well-known technique for feature extraction and dimension reduction. The performance of classical LDA however, significantly degrades on the High Dimension Low Sample Size (HDLSS) data for the *ill-posed inverse problem*. Existing approaches for HDLSS data classification typically assume the data in question are with Gaussian distribution and deal the HDLSS classification problem with regularization. However, these assumptions are too strict to hold in many emerging real-life applications, such as enabling personalized predictive analysis using Electronic Health Records (EHRs) data collected from an extremely limited number of patients who have been diagnosed with or without the target disease for prediction. In this paper, we revised the problem of predictive analysis of disease using personal EHR data and LDA classifier. To fill the gap, in this paper, we first studied an analytical model that understands the accuracy of LDA for classifying data with arbitrary distribution. The model gives a theoretical upper bound of LDA error rate that is controlled by two factors: (1) the *statistical convergence rate* of (inverse) covariance matrix estimators and (2) the divergence of the training/testing datasets to fitted distributions. To this end, we could lower the error rate by balancing the two factors for better classification performance. Hereby, we further proposed a novel LDA classifier *De-Sparse* that leverages *De-sparsified Graphical Lasso* to improve the estimation of LDA, which outperforms state-of-the-art LDA approaches developed for HDLSS data. Such advances and effectiveness are further demonstrated by both theoretical analysis and extensive experiments on EHR datasets <https://www.overleaf.com/project/5d2728c718f6ff3b2bcf5991>.

Keywords Linear discriminant analysis · De-sparsified graphical lasso · Electronic health records · High dimension low sample size

1 Introduction

Linear Discriminant Analysis (LDA) [1] is a well-known technique for feature extraction and dimension reduction. It has been widely used in many applications [2, 3] such as face recognition, image retrieval, etc. Typically, LDA finds the projection directions such that for the projected data, the between-class variance has been maximized relative to the within-class variance, thus achieving maximum discrimination. An intrinsic limitation of classical LDA is that its objective function requires the nonsingularity of one

of the scatter matrices. For many applications, such as the microarray data analysis, all scatter matrices can be singular or ill-posed since the data is often with high dimension but low sample size (HDLSS) [4].

Recently, many efforts have been devoted to bear on such HDLSS problems. For example, Krzanowski et al. proposed a pseudo-inverse LDA to approximate the inverse covariance matrix, when the sample covariance matrix is singular. However, the accuracy of pseudo-inverse LDA is usually low and not well guaranteed [5]. Another technique to alleviate this problem is a two-stage algorithm, *i.e.*, PCA+LDA [6, 7]. More popularly, regularized LDA approaches are proposed to solve the problem and improve the performance [8]. For example, researchers proposed a series of algorithms to regularize the covariance matrix estimation [2, 5, 9]. The regularized linear discriminant hyperplane was studied in [4, 10, 11]. All regularized LDA approaches intend to improve LDA through regularizing the estimation of key parameters used in LDA, such

✉ Zeyi Sun
sunzeyi@mininglamp.com

✉ Licheng Wang
wanglc@bupt.edu.cn

as the covariance matrix and/or the linear coefficients for discrimination.

One representative regularized LDA approach is Covariance Regularized Discriminant Analysis (CRDA) proposed in [9] based on the sparse inverse covariance estimation leveraging Graphical Lasso [12]. CRDA was originally proposed to estimate the inverse covariance matrix via a shrunk estimator, so as to achieve “*superior prediction*”. Intuitively, through replacing the sample covariance matrix used in LDA with the regularized estimation, the HDLSS problem can be well addressed since the regularized estimators usually outperform the sample covariance matrix estimator [13]. To better elucidate the performance of LDA classifiers with uncertain covariance matrix estimates for Gaussian data classification, [14] studied a model of error rate by matching the estimated *vs.* true covariance matrices, and the estimated *vs.* true means. While it is reasonable to assume that the estimated mean can easily converge to the population/true mean, the population/true covariance matrix is usually unknown and can be very different with the estimated one [13]. For example, the largest eigenvalue of the sample covariance matrix, which represents the principle component of the data distribution, is not consistent with the population one and the eigenvectors of the sample covariance matrix can be almost orthogonal to the truth under HDLSS [15, 16]. Further, the data for classification is usually Non-Gaussian. Thus, it is highly desirable to develop a new analytical model to characterize the error rate for the data with arbitrary distribution (both Gaussian and Non-Gaussian). Two “known factors” of covariance matrix estimation are useful for developing such analytical models, one is the convergence rate and the other one is the sparsity/density of (inverse) covariance matrix estimators [13]. The sparsity/density is already known once the (inverse) covariance matrix is estimated. The convergence rates reflect the maximal error of estimation, and for some estimators, they are well bounded under certain assumptions, such as spectral-norm convergence rate of Graphical Lasso [17].

Among a wide range of HDLSS data classification tasks, in this work, we focus on the problem of using LDA to classify EHR [19] for personalized predictive analytics of target disease. EHRs play a critical role in modern health information management and service innovations. A patient’s EHR contains his/her histories of medical visit, medication, diagnoses, treatment plans, allergies and so on as shown in Fig. 1. Per each visit a diagnosis record would be updated indicating the disease state, i.e., a set of codes referring to the diseases that diagnosed at a time of visit. One significant feature is the interchangeability of EHR, as a standard protocol for medical/health data generation, storage and communication. The health information is built and managed by authorized

institutions in a unified digital format (e.g., ICD-9/10, CPT-9/10 used in EHR standards) such that researchers and scientists can share and analyze the EHR data to enable innovative health services, such as providing computer-assisted diagnosis and offering medication advice. Among these services, predictive analytics of diseases (or namely early detection of diseases) using patients’ past longitudinal health information of the EHR system, has recently attracted significant attention from research communities.

There has been a series of works [19–24], which attempt to predict future disease of patients, through data mining techniques using EHR data. Prior literature usually first selected important features, such as diagnosis-frequencies [19], pairwise diagnosis transitions [22], and graphs of diagnosis sequences [24], to represent the EHR data of the patients. Then, a wide range of supervised learning algorithms were adopted to build predictive models for early disease detection, on top of well-represented EHR data.

In this paper we first propose a novel analytical model for LDA error rate, based on the statistical convergence of (inverse) covariance matrix estimators and the divergence to the Gaussian distributions. Guided by the proposed analytical model, we propose a novel LDA classifier leveraging the (inverse) covariance matrix estimators with faster convergence rate. We apply our model to a large-scale EHR dataset for the predictive analytics of diseases and demonstrate the advantage of the proposed algorithms over other state-of-the-arts. Specifically, in this paper we made contributions as follows.

1. We studied the problem of high-dimensional linear classification using LDA models and proposed a novel analytical model, derived from the existing LDA models on Gaussian data [14, 25], to understand the LDA error rate for both Gaussian and Non-Gaussian data, with respect to the statistical error of covariance matrices estimation and the divergence between fitted Gaussian distribution and the data.
2. On top of the analytical model, we proposed *De-Sparse*, which extends the well-known baseline approach – *Covariance Regularized Discriminant Analysis (CRDA)* [9, 26], using De-sparsified Graphical Lasso [27]. Theoretical analysis based on the proposed analytical model shows that *De-Sparse* can bound the maximal error rate, under mild conditions. Compared to CRDA, *De-Sparse* could achieve lower error rate, due to the faster convergence rate of De-sparsified Graphical Lasso.
3. To show the practical contribution of the proposed method, we evaluated *De-Sparse* extensively through experiments with large-scale, real-world EHR datasets [28]. In the experiments, we used *De-Sparse* to predict the risk of mental health disorders in college students from ten U.S. universities, using their EHR data from

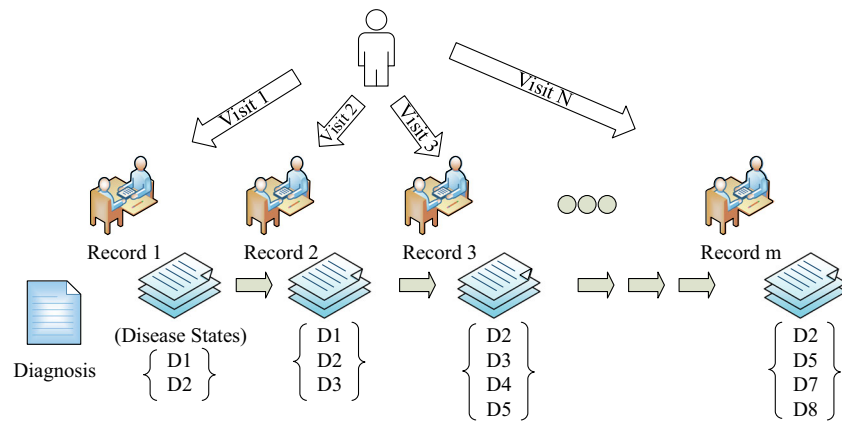


Fig. 1 Medicare visits and Electronic Health Records (EHRs). EHRs of a patient consist of the records of diagnoses and treatments. In this example, totally m medical visits have been placed. For every medical visit, the patient would receive a set of ICD/CPT codes [18] referring

the diseases/treatments that have been diagnosed/carried out. One can enable the early diagnosis/detection of diseases through classifying the EHR data, with big data and machine learning techniques

primary care visits. We compared *De-Sparse* with seven baseline algorithms including other regularized LDA models and downstream classifiers. The evaluation result shows that *De-Sparse* outperforms all baselines, and further validates our theoretical analysis.

The paper is organized as follows. In Section 2, we review the problem of high-dimensional linear classification using LDA models and summarize the existing work on EHR-based predictive analytics of diseases. In Section 3, we first introduce the existing covariance-regularized discriminant analysis (CRDA) based on Graphical Lasso, then present de-sparsified covariance regularized LDA algorithms, based on novel de-sparsified inverse covariance matrix estimators, to classify EHR samples for the predictive analytics. In Section 4, we validate the proposed algorithms with real-world datasets through extensive experiments. Finally, we conclude the paper in Section 5.

2 Preliminaries

2.1 LDA for binary classification

To use Fisher's Linear Discriminant Analysis (FDA), given N labeled data pairs $(x_1, l_1), (x_2, l_2), (x_3, l_3) \dots (x_N, l_N)$ and $\forall x_i, 1 \leq i \leq N$ refers to a d -dimensional vector, we first estimate the sample covariance matrix (an symmetric $d \times d$ matrix) using maximized likelihood estimator:

$$\bar{\Sigma} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{\mu})(x_i - \bar{\mu})^\top, \quad (1)$$

where $\bar{\mu}$ refers to the d -dimensional mean vector of all N training samples $(x_1, l_1), (x_2, l_2) \dots (x_N, l_N)$. Then, $\bar{\mu}_+$ and

$\bar{\mu}_-$ are estimated as the mean vectors of the positive training samples and negative training samples in the N training samples, respectively.

Corollary 1 (Fisher's Discriminant Analysis for Binary Classification [1]) *Given all estimated parameters $\bar{\Sigma}$, $\bar{\mu}_+$, and $\bar{\mu}_-$, the FDA model classifies a new data vector x as the result of (2) as follows.*

$$f_{\bar{\Sigma}}(x) = \text{sign} \left(\log \frac{x^\top \bar{\Sigma}^{-1} \bar{\mu}_+ - \frac{1}{2} \bar{\mu}_+^\top \bar{\Sigma}^{-1} \bar{\mu}_+ + \log \pi_+}{x^\top \bar{\Sigma}^{-1} \bar{\mu}_- - \frac{1}{2} \bar{\mu}_-^\top \bar{\Sigma}^{-1} \bar{\mu}_- + \log \pi_-} \right), \quad (2)$$

where $\text{sign}(\cdot) : \mathbb{R} \rightarrow \{\pm 1\}$ refers to the signal function, π_+ and π_- refer to the (foreknown) frequencies of positive samples and negative samples in the whole population.

To present LDA with other covariance matrix estimator, based on the LDA paradigm listed in (2), we use the notations as follows.

Notations Note that, in the rest of this paper, we denote $f_{\hat{\Sigma}}(x)$ as an LDA classifier with a specific covariance matrix estimator $\hat{\Sigma}$, using the sample estimated mean vectors $\bar{\mu}_+$ and $\bar{\mu}_-$. When $\hat{\Sigma} = \bar{\Sigma}$, then the classifier $f_{\bar{\Sigma}}(x)$ becomes the traditional Fisher's Linear Discriminant Analysis. When $\hat{\Sigma} = \hat{\Theta}^{-1}$ and $\hat{\Theta}$ is the Graphical Lasso estimator [12], then $f_{\hat{\Theta}^{-1}}(x)$ refers to the covariance regularized LDA [9, 26].

Apparently, the performance of LDA depends on (1) whether the realistic training/testing datasets follow Gaussian distributions and (2) how the mean vectors and inverse covariance matrices are estimated from the datasets.

2.2 Electronic health records and predictive analytic of disease

Prior to learning a predictive analytic model for certain diseases, one needs to model the EHR data with a suitable data representation. The most simple yet effective way to represent EHR data is to use *diagnosis-frequency vector* [29–31], which is similar to Term Frequency (TF) or Term Frequency-Inverse Document Frequency (TF-IDF) approach that deals with traditional NLP data [32–34]. Given each patient's EHR data (shown in Fig. 1), this representation method first retrieves the diagnosis codes [35] recorded during each visit. Inspired by some Natural Language Processing (NLP) and text mining practices [36], researchers also proposed using some deep learning based NLP approaches to embed EHR records for data representation learning [37–42]. For example, [38] studied “Patient2Vec” which embeds patients' past EHR records into vectors while preserving structural information for personalized predictive analysis. Bai et al. [40] focuses on the interpolation and interpretability of EHR representation learning, where authors well-balanced the performance of predictive analysis and the understanding to the longitude disease progress of each individual patient, both using the EHR data with the learned representation. Comprehensive surveys could be found in [18, 43, 44].

In our work, we follow the line of research that uses *diagnosis-frequency vector* of each patient for EHR-based predictive analysis [29–31], as the diagnosis-frequency in a certain duration could well characterize the health status of patients while the coefficient of LDA can represent the significance of every diagnosis code. The frequency of each diagnosis appearing in all past visits (of the last two years) is counted here, followed by further transformation on the frequency of each diagnosis into a vector of frequencies. For example, a *diagnosis-frequency vector* can be illustrated as $\langle 1, 0, \dots, 3 \rangle$, where 0 means the second diagnosis does not exist in all past visits. Note that the dimension $d \geq 15,000$ when using original ICD-9 codes, $d = 295$ even when using clustered ICD-9 codes [45], while the number of samples for training N in our experiment is significantly smaller than d .

2.3 Discussion on preliminaries

In our work, we revisited the linear discriminant analysis as a classifier and learner for High-Dimensional and Low Sample Size (HDLSS) settings. Indeed, many efforts have been made in the literature for HDLSS data classification. For example, in addition to LDA-type methods, a number of feature extraction or variable selection methods have been studied. Lin et al. [46] proposed a feature selection algorithm to classify the high-dimensional gene expression data through incorporating the neighborhood entropy-based

uncertainty measures. Over the rough set, the same group of authors [47] adopted a joint feature selection approach that incorporates neighborhood entropy and the fisher scores, for tumor classification. Further, some automatic feature weighting paradigm has been proposed to select features for gene expression data classification [48]. These studies demonstrate that the feature selection algorithms could significantly improve the accuracy of HDLSS data classification, while avoiding the full set of features. The over-reduction problem of LDA has been studied in [49]. In addition to the EHR data, similar regularized projection methods have been used for early diagnosis of diseases for biomedical health data [50, 51].

In terms of methodologies, the most close work to this study is covariance-regularized linear discriminant analysis (CRDA) [9], Graphical Lasso [17], and the de-sparsified Graphical Lasso [27]. CRDA regularizes the estimation of (inverse) covariance matrices inside the estimation of LDA, while improving the performance of LDA for both prediction and inference. Authors in [26] were the first to bring CRDA for EHR classification and early detection of diseases. We included the algorithms in [26] for comparison and found that *De-Sparse* outperformed CRDA with higher accuracy and F1-score. Compared to the Graphical Lasso estimator [17] that has been frequently used to enhance the inverse covariance matrices estimation, our work followed the ideas of de-biased estimator [52] and used de-sparsified Graphical Lasso estimator [27] to improve the LDA for EHR classification. We would provide a comprehensive discussion on the performance comparisons between Graphical Lasso and de-sparsified one from the perspectives of predictive analytics based on EHR data and LDA.

3 De-Sparse: De-sparsified covariance-regularized discriminant analysis

In this section, we first introduce the baseline algorithm based on Covariance-Regularized Discriminant Analysis (CRDA) that is derived from [26]. Then, we present the proposed algorithm *De-Sparse*, an extended Covariance Regularized Discriminant Analysis via De-sparsified Graphical Lasso [27]. Then, using our proposed analytical model of LDA error rate, we compare two methods and demonstrate the advantages of *De-Sparse*.

3.1 CRDA: The baseline of covariance-regularized discriminant analysis via graphical lasso inverse covariance matrix estimator

Compared to the sample LDA introduced in Section 2, CRDA [9, 26] was proposed to use ℓ_1 -penalized inverse

covariance matrix estimator to replace the inverse of sample covariance matrix. Given the labeled data pairs for training $(x_1, l_1), (x_2, l_2) \dots (x_N, l_N)$, the algorithm first estimates the sample covariance matrix $\bar{\Sigma}$ and the sample mean vectors $\bar{\mu}_+, \bar{\mu}_-$ using the algorithms introduced in Section 2.1. With the sample covariance matrix $\bar{\Sigma}$, this method estimates a sparse inverse covariance matrix $\hat{\Theta}$ using the Graphical Lasso estimator [12] as follows.

Corollary 2 (Graphical Lasso Estimator [12]) *Given the sample estimation of the covariance matrix $\bar{\Sigma}$, the Graphical Lasso estimator provides an ℓ_1 -regularized sparse positive-definite approximation to the inverse covariance matrix (denoted as $\hat{\Theta}$) as follows.*

$$\hat{\Theta} = \underset{\Theta > 0}{\operatorname{argmin}} \left(\operatorname{tr}(\bar{\Sigma}\Theta) - \log \det(\Theta) + \lambda \sum_{j \neq k} |\Theta_{jk}| \right), \quad (3)$$

where $\Theta > 0$ refers to the constraint of symmetric positive definiteness (SPD), the term $\operatorname{tr}(\bar{\Sigma}\Theta) - \log \det(\Theta)$ refers to the negative log-likelihood of the optimization objective Θ over the sample estimate $\bar{\Sigma}$, the term $\sum_{j \neq k} |\Theta_{jk}|$ refers to the sum of absolute values of the non-diagonal elements in the matrix Θ (which is the same as the ℓ_1 -norm of Θ without diagonal elements considered), and λ refers to tuning parameter that makes trade-off between the sparsity and the fitness to samples. Please refer to [12] for the implementation of the algorithms.

Corollary 3 (Statistical Convergence of Graphical Lasso [17]) *Suppose the random vector \mathbf{X} is with d -dimensions and zero mean (i.e., $\mathbf{X} \in \mathbb{R}^d$ and $\mathbb{E}(\mathbf{X}) = \mathbf{0}$), where the population estimate of the covariance matrix is $\Sigma = \mathbb{E}(\mathbf{X}\mathbf{X}^\top)$ and the inverse of population covariance matrix is $\Theta = \Sigma^{-1}$. Given N samples $x_1, x_2, x_3, \dots, x_N$ randomly and independently drawn from \mathbf{X} , the sample estimate of the covariance matrix here should be $\bar{\Sigma} = \frac{1}{N} \sum_{i=1}^N x_i x_i^\top$.*

With the increasing number of samples (N) given and growing number of dimensions of the data (d), the graphical lasso estimate $\hat{\Theta}$ based on the sample covariance matrix converges to the population estimate Θ at the rate of Frobenius-norm under mild sparsity conditions, as follows [17].

$$\|\hat{\Theta} - \Theta\|_F = \mathcal{O} \left(\sqrt{\frac{(d+s) \log p}{N}} \right), \quad (4)$$

where $s = \max_{1 \leq i \leq d} \|\Theta_i\|_0$ refers to the maximal degree of the graph in Θ , $\|\cdot\|_0$ refers to the ℓ_0 -norm of the input vector, and Θ_i refers to the i^{th} column vector ($1 \leq i \leq d$) of the matrix Θ_i .

To the end, the classification rule of CRDA is written as follows

$$\text{CRDA}(x) = \operatorname{sign} \left(\log \frac{x^\top \hat{\Theta} \bar{\mu}_+ - \frac{1}{2} \bar{\mu}_+^\top \hat{\Theta} \bar{\mu}_+ + \log \pi_+}{x^\top \hat{\Theta} \bar{\mu}_- - \frac{1}{2} \bar{\mu}_-^\top \hat{\Theta} \bar{\mu}_- + \log \pi_-} \right), \quad (5)$$

which can be viewed as an LDA classifier using $\hat{\Theta}^{-1}$ as the covariance matrix. Apparently, the accuracy of CRDA depends on how the covariance matrices and the mean vectors are estimated. We are going to interpret the performance of CRDA in the Section 3.3.

3.2 De-Sparse: The improved algorithm of covariance-regularized LDA via de-sparsified graphical lasso

As shown in (3), the estimator of sparse inverse covariance matrix induced ℓ_1 -penalization and might hurt the estimation due to the over-penalization or over-sparsification. To address this issue, we proposed a de-sparsified Graphical Lasso estimator [27] to replace the vanilla Graphical Lasso.

Corollary 4 (De-sparsified Graohical Lasso [27]) *Given the Graphical Lasso estimator $\hat{\Theta}$ and the sample estimation $\bar{\Sigma}$, we consider the inverse of Graphical Lasso $\hat{\Theta}^{-1}$ as an approximation to the covariance matrix. In this way, the bias of $\hat{\Theta}^{-1}$, caused by the sparsity regularizer of Graphical Lasso, for covariance estimation could be written as follows.*

$$\hat{Z} = \bar{\Sigma} - \hat{\Theta}^{-1}. \quad (6)$$

Using the Kronecker product, authors in [27] consider the potential bias term of $\hat{\Theta}$ against the inverse of population covariance matrix as follows.

$$\text{Bias}(\bar{\Sigma}, \hat{\Theta}) = \hat{\Theta} \hat{Z} \hat{\Theta} = \hat{\Theta} \bar{\Sigma} \hat{\Theta} - \hat{\Theta}. \quad (7)$$

The de-sparsified Graphical Lasso estimator \hat{T} de-biases the Graphical Lasso estimator $\hat{\Theta}$ through removing the potential bias term caused by the sparsity regularizer, as follows.

$$\hat{T} = \hat{\Theta} - \text{Bias}(\bar{\Sigma}, \hat{\Theta}) = 2\hat{\Theta} - \hat{\Theta} \bar{\Sigma} \hat{\Theta}. \quad (8)$$

On top of the Graphical Lasso, the de-sparsified Graphical Lasso estimator can efficiently approximate an estimation of inverse covariance matrix using the data with faster convergence rate in a mild condition.

Corollary 5 (Statistical Convergence of De-sparsified Graphical Lasso [27]) *Suppose the random vector \mathbf{X} is with d -dimensions and zero mean (i.e., $\mathbf{X} \in \mathbb{R}^d$ and $\mathbb{E}(\mathbf{X}) = \mathbf{0}$), where the population estimate of the covariance matrix is $\Sigma = \mathbb{E}(\mathbf{X}\mathbf{X}^\top)$ and the inverse of population covariance matrix is $\Theta = \Sigma^{-1}$. Given N samples $x_1, x_2, x_3, \dots, x_N$*

randomly and independently drawn from \mathbf{X} , the sample estimate of the covariance matrix here should be $\hat{\Sigma} = \frac{1}{m} \sum_{i=0}^{m-1} x_i x_i^\top$. The Graphical Lasso estimator and the de-sparsified estimator are denoted as $\hat{\Theta}$ and \hat{T} , respectively.

With the increasing number of samples (N) given and growing number of dimensions of the data (d), the De-sparsified Graphical Lasso estimator \hat{T} converges to the population estimate Θ at the rate of ℓ_∞ -norm under mild sparsity conditions, as follows [27].

$$\|\hat{\Theta} - \Theta\|_\infty = \mathcal{O}\left(\sqrt{\frac{\log d}{N}}\right). \quad (9)$$

Note that, above convergence rate of de-sparsified Graphical Lasso was obtained under similar sparsity assumptions as [17], while the ℓ_2 -norm or ℓ_∞ -norm convergence rates of Graphical Lasso are not known yet.

Based on Notations, we denote the De-sparsified Covariance Regularized Discriminant Analysis (namely *De-Sparse*) as $\text{Desparse}(x)$, using the De-sparsified Graphical Lasso \hat{T} and the mean vectors $\bar{\mu}_+$, $\bar{\mu}_-$.

$$\text{Desparse}(x) = \text{sign}\left(\log \frac{x^\top \hat{T} \bar{\mu}_+ - \frac{1}{2} \bar{\mu}_+^\top \hat{T} \bar{\mu}_+ + \log \pi_+}{x^\top \hat{T} \bar{\mu}_- - \frac{1}{2} \bar{\mu}_-^\top \hat{T} \bar{\mu}_- + \log \pi_-}\right). \quad (10)$$

With the de-sparsified inverse covariance matrix estimator \hat{T} enjoying better statistical properties, *De-Sparse* is expected to outperform CRDA with better classification accuracy. Detailed comparison will be discussed in the following sections.

3.3 Performance analysis of LDA, CRDA, and De-Sparse

In this section, we first review the previous studies on the LDA error rate estimation for Gaussian data [14, 25], then we generalize LDA error rate from Gaussian data to non-Gaussian data. Finally, we provide a discussion on the classification accuracy comparison among vanilla LDA, CRDA, and *De-Sparse*.

3.3.1 LDA error rate for Gaussian data via random matrix theory

We first assume the data for binary classification follow two (unknown) Gaussian distributions with the same covariance matrix Σ 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 $f_{\hat{\Sigma}}(x)$ based on the sample estimated mean vectors $\bar{\mu}_-$, $\bar{\mu}_+$ and a specific covariance matrix $\hat{\Sigma}$, the expected error rate of a

linear discriminant analysis (i.e., probability of $l \neq f_{\hat{\Sigma}}(x)$) on the data of $\mathcal{N}(\mu_+, \Sigma)$, $\mathcal{N}(\mu_-, \Sigma)$ is modeled as follows.

Corollary 6 (RMT-based LDA Error Rate Estimation [14])

According to the random matrix theory, [14] models the expectation of classification error rate of LDA (using estimated parameters $\bar{\mu}_+$, $\bar{\mu}_-$, and $\hat{\Sigma}$) on Gaussian distributions $\mathcal{N}(\mu_+, \Sigma)$ and $\mathcal{N}(\mu_-, \Sigma)$ as $\varepsilon(\bar{\mu}_+, \bar{\mu}_-, \hat{\Sigma}, \mu_+, \mu_-, \Sigma)$, as follows.

$$\begin{aligned} & \varepsilon(\bar{\mu}_+, \bar{\mu}_-, \hat{\Sigma}, \mu_+, \mu_-, \Sigma) \\ &= \pi_+ \cdot \Phi\left(-\frac{(\mu_+ - \frac{(\bar{\mu}_+ + \bar{\mu}_-)}{2})^\top \hat{\Sigma}^{-1} (\bar{\mu}_+ - \bar{\mu}_-)}{\sqrt{(\bar{\mu}_+ - \bar{\mu}_-)^\top \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\bar{\mu}_+ - \bar{\mu}_-)}}\right) \\ & \quad + \pi_- \cdot \Phi\left(\frac{(\mu_- - \frac{(\bar{\mu}_+ + \bar{\mu}_-)}{2})^\top \hat{\Sigma}^{-1} (\bar{\mu}_+ - \bar{\mu}_-)}{\sqrt{(\bar{\mu}_+ - \bar{\mu}_-)^\top \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\bar{\mu}_+ - \bar{\mu}_-)}}\right) \end{aligned} \quad (11)$$

where Φ refers to the CDF function of a standard normal distribution.

According to **Corollary 6** and [14], we could conclude that the expected error rate is sensitive with the parameters μ_+ , μ_- , Σ , $\bar{\mu}_+$, $\bar{\mu}_-$ and $\hat{\Sigma}$, while the true parameters μ_+ , μ_- , Σ are assumed unknown. Compared to the (inverse) covariance matrices estimation, the error of sample mean vector estimation is relatively small [53]. Thus, we adopt the settings in studies [2, 5, 25] as follows.

Assumption 1 In this paper, we make no assumptions on the mean vectors μ_+ , μ_- , μ and always use the sample mean $\bar{\mu}_+$, $\bar{\mu}_-$, $\bar{\mu}$ to estimate μ_+ , μ_- , μ . Even under the HDLSS settings, with a certain number of samples, it is reasonable to assume the sample estimation of mean vectors $\bar{\mu}_+$ and $\bar{\mu}_-$ should be close to the population mean vectors, i.e., $|\mu_+ - \bar{\mu}_+| \rightarrow 0$, $|\mu_- - \bar{\mu}_-| \rightarrow 0$, and $|\mu - \bar{\mu}| \rightarrow 0$.

Lemma 1 Thus, based on Theorem 1 and the sample mean relaxation (Assumption 1), the expected error rate of $f_{\hat{\Sigma}}(x)$ can be reduced to

$$\varepsilon(\hat{\Sigma}, \Sigma) = \Phi\left(-\frac{(\bar{\mu}_+ - \bar{\mu}_-)^\top \hat{\Sigma}^{-1} (\bar{\mu}_+ - \bar{\mu}_-)}{2\sqrt{(\bar{\mu}_+ - \bar{\mu}_-)^\top \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\bar{\mu}_+ - \bar{\mu}_-)}}\right). \quad (12)$$

In this way, to improve the LDA classifier with the sample mean vectors, there needs an estimator $\hat{\Sigma}$ to minimize or lower the expected error rate $\varepsilon(\hat{\Sigma}, \Sigma)$.

Lemma 2 Furthermore, when the estimated covariance matrix $\hat{\Sigma}$ is set to the oracle one Σ (the LDA is perfectly

fitted with the data), the expected error rate reaches the optimal error rate

$$\varepsilon(\hat{\Sigma}, \hat{\Sigma}) = \Phi \left(-\frac{\sqrt{(\bar{\mu}_+ - \bar{\mu}_-)^T \hat{\Sigma}^{-1} (\bar{\mu}_+ - \bar{\mu}_-)}}{2} \right). \quad (13)$$

Above result suggests that when the covariance matrix $\hat{\Sigma}$ is perfectly estimated $\hat{\Sigma} \rightarrow \Sigma$ and $\hat{\Sigma}^{-1} \Sigma \rightarrow I$, the LDA classifier would approach to its optimal error rate.

The estimate in (12) reduces the estimation of LDA classification error rate to the divergence between the population covariance matrix Σ and the estimated one $\hat{\Sigma}$. On the other hand, (13) models the generalization error of a model with “perfectly-fitted” covariances [14].

3.3.2 LDA error rate for non-Gaussian data via Kullback-Leibler divergence

We deliver the analysis on LDA classification error rate through incorporating additional assumptions as follows.

Assumption 2 Suppose every positive sample x_+ and negative sample x_- are realized from random variables \mathbf{X}_+ and \mathbf{X}_- respectively. We denote $\mu_+ = \mathbb{E}[\mathbf{X}_+]$ and $\mu_- = \mathbb{E}[\mathbf{X}_-]$ as the expectations of \mathbf{X}_+ and \mathbf{X}_- respectively. In this way we define Σ_+^* and Σ_-^* as the oracle covariance matrices for two classes respectively, such that

$$\begin{aligned} \Sigma_+^* &= \mathbb{E}[(\mathbf{X}_+ - \mu_+)(\mathbf{X}_+ - \mu_+)^T] \text{ and} \\ \Sigma_-^* &= \mathbb{E}[(\mathbf{X}_- - \mu_-)(\mathbf{X}_- - \mu_-)^T]. \end{aligned}$$

We further denote the oracle between-class covariance matrix $\Sigma^* = (\Sigma_+^* + \Sigma_-^*)/2$. To simplify our analysis, we further assume $\Sigma^* \approx \Sigma_+^* \approx \Sigma_-^*$.

Lemma 3 Considering the divergence between the Gaussian distributions $\mathcal{N}(\mu_+, \Sigma_+^*)$ and $\mathcal{N}(\mu_-, \Sigma_-^*)$ to the real data, we could bound the expected error rate of the LDA classifier $f_{\hat{\Sigma}}(x)$ (i.e., the LDA classifier based on parameters $\bar{\mu}_+$, $\bar{\mu}_-$, $\hat{\Sigma}$) as:

$$\text{err}(\bar{\mu}_+, \bar{\mu}_-, \hat{\Sigma}) \lesssim \left(\varepsilon(\hat{\Sigma}, \Sigma^*) + \sum_{l \in \{-1, +1\}} \pi_l \sqrt{\frac{D_{\text{KL}}(\mathbf{X}_l \| \mathcal{N}(\mu_l, \Sigma_l^*))}{2}} \right) \quad (14)$$

where $D_{\text{KL}}(\mathbf{X}_l \| \mathcal{N}(\mu_l, \Sigma_l^*))$ refers to the Kullback–Leibler divergence between the distribution of the data \mathbf{X}_l and Gaussian distribution $\mathcal{N}(\mu_l, \Sigma_l^*)$.

Proof To prove Lemma 3, we first define the error function $\mathbf{1}[l \neq f_{\hat{\Sigma}}(x)] = 1$: when $l \neq f_{\hat{\Sigma}}(x)$ and $\mathbf{1}[l \neq f_{\hat{\Sigma}}(x)] = 0$ when $l = f_{\hat{\Sigma}}(x)$. Then, the error rate of $f_{\hat{\Sigma}}(x)$, for any

data distribution with density functions $P(x, l)$ and $l \in \{+1, -1\}$, can be written as:

$$\text{err}(\bar{\mu}_+, \bar{\mu}_-, \hat{\Sigma}) = \int \mathbf{1}[l \neq f_{\hat{\Sigma}}(x)] \mathbf{d} P(x, l).$$

Consider the error rate on Gaussian distribution $\varepsilon(\hat{\Sigma}, \Sigma^*)$, the conditional probability based on Gaussian distribution $P_{\Sigma_l^*}(x|l)$, and the conditional probability $P(x, l) = P(x|l) \cdot \pi_l$.

$$\lesssim \varepsilon(\hat{\Sigma}, \Sigma^*) + \sum_{l \in \{+1, -1\}} \pi_l \int_x |P(x|l) - P_{\Sigma_l^*}(x|l)| \mathbf{d}_x$$

Consider Pinsker’s inequality to bound the divergence.

$$\leq \varepsilon(\hat{\Sigma}, \Sigma^*) + \sum_{l \in \{-1, +1\}} \pi_l \sqrt{\frac{D_{\text{KL}}(\mathbf{X}_l \| \mathcal{N}(\mu_l, \Sigma_l^*))}{2}} \quad \square$$

Lemma 3 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 [54]. Given an LDA classifier $f_{\hat{\Sigma}}$, the error rate is upper bounded by two factors: (1) the error rate of $f_{\hat{\Sigma}}$ on the nearest Gaussian distribution of the data, i.e., $\varepsilon(\hat{\Sigma}, \Sigma^*)$, and (2) the divergence between the data distribution and the Gaussian distribution. Considering Lemma 1, we can further conclude that two factors: the divergence of the given data to the Gaussian distribution $D_{\text{KL}}(P_l \| \mathcal{N}(\bar{\mu}_l, \Sigma_l^*))$ and the statistical convergence of Σ^{*-1} to Σ^{-1} would affect the performance of LDA on classification accuracy.

3.3.3 Performance Comparisons

We compare the convergence rate of \hat{T} and $\hat{\Theta}$ to the inverse population covariance matrix $\Theta^* = \Sigma^{*-1}$, so as to understand the accuracy of *De-Sparse* and *CRDA*. Since the divergence of the datasets to the nearest Gaussian distributions should be the same for both algorithms, we mainly compared the Gaussian error terms for *De-Sparse* and *CRDA*, i.e., $\varepsilon(\hat{T}^{-1}, \Sigma^*)$ vs. $\varepsilon(\hat{\Theta}^{-1}, \Sigma^*)$. More specifically, [14] demonstrated the connections between the Gaussian data error terms and the spectral properties of \hat{T} and $\hat{\Theta}$. Considering **Lemma 2**, we hope to understand (1) how close the matrices $\hat{T}\Sigma^*$ and $\hat{\Theta}\Sigma^*$ would approach to I matrix and (2) how the spectrum of these matrices behaves [55], such that

$$\begin{aligned} \|\hat{T}\Sigma^* - I\|_2 &= \|(\hat{T} - \Theta^*)\Sigma^*\|_2 \\ &\leq \lambda_{\max}(\Sigma^*) \|\hat{T} - \Theta^*\|_2, \text{ and} \\ \|\hat{\Theta}\Sigma^* - I\|_2 &= \|(\hat{\Theta} - \Theta^*)\Sigma^*\|_2 \\ &\leq \lambda_{\max}(\Sigma^*) \|\hat{\Theta} - \Theta^*\|_2, \end{aligned} \quad (15)$$

where $\lambda_{\max}(\cdot)$ refers to the largest eigenvalue of the input matrix. Obviously, the terms of $\lambda_{\max}(\Sigma^*)$, $\|\hat{T} - \Theta^*\|_2$ and

$\|\hat{\Theta} - \Theta^*\|_2$ are non-negative. When $\|\hat{T} - \Theta^*\|_2 \rightarrow 0$ and $\|\hat{\Theta} - \Theta^*\|_2 \rightarrow 0$, then optimal error rates would be achieved asymptotically. In this way, we are wondering whether \hat{T} would converge to Θ^* faster than $\hat{\Theta}$ in the spectrum-norm distance.

Considering Corollaries 3 and 5, the sharp spectrum-norm statistical convergence rate of Graphical Lasso is not known in any of the previous studies [13, 17, 27, 56], while the spectrum-norm statistical convergence rate of de-sparsified Graphical Lasso could be easily derived from the ℓ_∞ -norm rate. In this way, we compare CRDA and *De-Sparse* through the ℓ_2 -norm statistical convergence rate of their inverse covariance matrix estimators. Thus, we can derive the ℓ_2 -norm convergence rate as:

$$\|\hat{T} - \Theta^*\|_2 = \mathcal{O}\left(\sqrt{\frac{d \log d}{N}}\right). \quad (16)$$

On the other side, [17] demonstrated that the ℓ_2 -norm convergence rate for the Graphical Lasso estimator $\hat{\Theta}$ is

$$\|\hat{\Theta} - \Theta^*\|_2 = \mathcal{O}\left(\sqrt{\frac{(d+s) \log d}{N}}\right),$$

where s has been defined in Corollary 3. We conclude *the convergence rate of \hat{T} is faster than $\hat{\Theta}$* . In this way, we consider *De-Sparse* would outperform CRDA as it adopts a better inverse covariance matrix estimator.

4 Experiments

In this section, we first introduce the design of the experiments to evaluate the superiority of the proposed *De-Sparse* framework. Then, we present the experimental results, including the performance comparison between the *De-Sparse* framework, existing LDA baselines, and other predictive models, followed by a comparison between inverse covariance matrix to support our theoretical analysis of *De-Sparse*.

4.1 Experiment setups

In this study, to evaluate *De-Sparse*, we use predictive analytics of disease based on Electronic Health Records (EHR) data.

- **Predictive Analytics of Diseases** - Given N training samples (i.e., the EHR data of each patient) along with corresponding labels i.e., $(x_1, l_1), (x_2, l_2) \dots (x_N, l_N)$ where $l_i \in \{-1, +1\}$ refers to whether the patient i is diagnosed with the target disease or not (i.e., positive sample or negative sample), the predictive analytics task is to determine whether a new patient would

develop into the target disease via classification of the vector x to $+1$ (diagnosed as the positive result) or -1 (diagnosed as the negative result).

- **Performance Metrics** - To demonstrate the effectiveness of predictive analytics of diseases, we compared all these methods with other competitors using Accuracy and F1-score. Specifically, Accuracy characterizes the proportion of patients who are accurately classified by the algorithms. F1-Score measures both correctness and completeness of the prediction. Of-course, we also include other metrics such as sensitivity and specificity to evaluate the performance of predictive analytics through addressing the medical concerns.
- **Data for Evaluation** - In the experiments, we use the de-identified EHR data of 200,000 students from ten U.S. universities [28]. Among all diseases recorded, we choose mental health disorders, including *anxiety disorders, mood disorders, depression disorders, and other related disorders*, as one targeted disease for early detection [57]. We represent each patient using his/her diagnosis-frequency vector based on the clustered codeset ($d = 295$).

Note that to prepare the training and testing sets, we use the complete EHR data of the patients who haven't been diagnosed with any of mental health disorders (negative samples). For patients having been diagnosed with any mental health disorders (positive samples), we collect their EHR data from the first visit to the last visit that was 90 days before the diagnosis of mental health disorders. Thus, we can simulate the early detection of diseases with 90 days in advance.

4.2 Design of experiment

To understand the performance impact of *De-Sparse* beyond classic LDA, we first propose four LDA baseline approaches to compare against *De-Sparse*, then, three discriminative learning models are used for the comparison:

- **LDA Derivatives: *LDA*, *Shrinkage*, *DIAG* and *Ye-LDA*** - The first three algorithms are all based on the common implementation of generalized Fisher's discriminant analysis listed in (2). Specifically, *LDA* uses the sample covariance estimation, and inverts the covariance matrix using pseudo-inverse [58] when the matrix inverse is not available; *Shrinkage* is based on *LDA*, using a sparse estimation of sample covariance as follows,

$$\Sigma_\beta = \beta * \bar{\Sigma} + (1 - \beta) * \text{diag}(\bar{\Sigma}) \text{ and } \Theta_\beta = \Sigma_\beta^{-1}, \quad (17)$$

where $\text{diag}(\bar{\Sigma})$ refers to the diagonal matrix of the sample estimation $\bar{\Sigma}$. *DIAG* is a special *Shrinkage* approach with $\beta = 0.0$. *Ye-LDA* is derived from [7,

58]. In our research, we focus on studying the improvement of LDA classification caused by (inverse) covariance matrix regularization, thus we don't compare our method to linear-coefficient-regularized LDA classifiers [4, 10, 11] or heuristic LDA derivation [6].

- Downstream Classifiers: *Support Vector Machine (SVM)*, *Logistic Regression (Logit. Reg.)* and *AdaBoost* – Inspired by the previous studies [21, 59] in EHR data mining, we use a linear binary SVM classifier with fine-tuned parameters as well as the Logistic Regression classifier. Further, we compare our algorithm to AdaBoost, where AdaBoost-10 refers to the AdaBoost classifier using 10 Logistic Regression instances and AdaBoost-50 leverages 50 instances.

With the seven baseline algorithms, we perform experiments with training sets of varying sizes and cross-validation. To train the classifiers, we randomly selected 50, 100, 150, 200, and 250 positive patients, and randomly selected the same number of negative patients. Then, we test the classifiers, using a testing set with 1000 randomly selected positive patients and the same number of negative patients. Note that there is no over-lap between training set and the paired testing set. All algorithms used in our work are implemented with JSAT¹ and glasso in R².

4.3 Overall comparison

We include the comparison results of *De-Sparse* evaluation in Tables 1, 2, 3, 4, and 5 for models learned from 50 ~ 250 × 2 labeled samples respectively. All experiments are done with cross validation using random sampling without replacement and repeated 10 times. Specifically, we compare the performance using various experimental settings, such as the varying parameters for model training and number of days in advance for early detection (e.g., 30 days, 60 days and 90 days). We carry out the experiments with varying *Days in Advance* settings, so as to evaluate the performance of algorithms for predictive analytics. As was addressed, we actually need to use the past EHR data (before the diagnoses of mental disorders) as the features for prediction. More specific, for positive samples in both training and testing datasets, we backtracked their EHR data from their prediction dates. For every positive sample, the prediction date is set as 30, 60 and 90 days before the medicare visit that the patient received his/her first diagnoses of “anxiety disorders, mood disorders, depression disorders, and other related disorders”. In this way, we carry out the experiments in three categories according to the varying *days in advance*.

¹<https://github.com/EdwardRaff/JSAT>

²<https://cran.r-project.org/package=glasso>

4.3.1 Comparisons on accuracy and F1-score

As can be seen from the results in Tables 1, 2, 3, 4, and 5, *De-Sparse* clearly outperforms the baseline algorithms in terms of overall accuracy and F1-score. Specifically, *De-Sparse* achieves 18.6%–21.3% increase in accuracy and 22.9%–32% increase in F1-score over LDA; *De-Sparse* achieves 17.9% increase in accuracy and 31.5%–40.6% increase in F1-score over DIAG. Compared to Shrinkage and CRDA, the accuracy and F1-score of *De-Sparse* in most parameter settings are 0.3%–18.9% higher and 0.14%–71.8% higher, respectively. Compared to SVM, Logistic Regression, and AdaBoost, *De-Sparse* can achieve 2.3%–19.4% higher accuracy and 7.5%–43.5% higher F1-score. In this case, we can conclude that the classic LDA model cannot perform as well as many other predictive models such as SVM and AdaBoost. However, *De-Sparse* significantly outperforms these methods in all settings. Thus, we can conclude that *De-Sparse* overall outperforms the baseline algorithms in all experimental settings. Note that, though *De-Sparse* outperforms CRDA marginally, *De-Sparse* enjoys a more tight upper bound of error rate.

4.3.2 Trade-off between sensitivity and specificity

We also intend to compare *De-Sparse* with baseline methods with respect to the needs of medicines. Specifically, in addition to accuracy and F1-score, we focus on two more metrics [60]:

- *Sensitivity* - In medical diagnosis, the sensitivity measures the ability of the prediction algorithms to *correctly identify the patients with the disease (true positive rate)*. More specific, we estimate sensitivity as

$$\text{Sensitivity} = \frac{\# \text{Patients with the diseases} \cap \text{Patients predicted as positive}}{\# \text{Patients with the diseases}}. \quad (18)$$

- *Specificity* - In contrast, the specificity metric characterizes the ability of the algorithms to *correctly identify ones without the disease (true negative rate)*. More specific, we estimate specificity as

$$\text{Specificity} = \frac{\# \text{Patients without the diseases} \cap \text{Patients predicted as negative}}{\# \text{Patients without the diseases}}. \quad (19)$$

Please see also Tables 1, 2, 3, 4, and 5. In terms of specificity, the baseline algorithms outperform *De-Sparse*, in the most of cases. However, in terms of sensitivity and specificity trade-off, *De-Sparse* on average gains 19.5% higher sensitivity while sacrificing 8.2% specificity, when

Table 1 Performance Comparison with Training Set: 50×2 , Testing Set: 2000×2

	Accuracy	F1-Score	Sensitivity	Specificity
Days in Advance: 30				
AdaBoost ($\times 10$)	0.637 ± 0.028	0.571 ± 0.057	0.491 ± 0.085	0.783 ± 0.053
AdaBoost ($\times 50$)	0.640 ± 0.024	0.570 ± 0.061	0.487 ± 0.093	0.792 ± 0.053
CRDA ($\lambda = 1.0$)	0.662 ± 0.017	0.692 ± 0.028	0.762 ± 0.069	0.563 ± 0.058
CRDA ($\lambda = 10.0$)	0.670 ± 0.017	0.713 ± 0.010	0.819 ± 0.023	0.520 ± 0.047
CRDA ($\lambda = 100.0$)	0.664 ± 0.020	0.713 ± 0.008	0.834 ± 0.033	0.494 ± 0.068
LDA	0.555 ± 0.026	0.565 ± 0.033	0.579 ± 0.048	0.531 ± 0.040
Logistic Regression	0.615 ± 0.055	0.469 ± 0.206	0.395 ± 0.200	0.835 ± 0.094
De-Sparse ($\lambda = 1.0$)	0.658 ± 0.019	0.677 ± 0.034	0.723 ± 0.073	0.592 ± 0.050
De-Sparse ($\lambda = 10.0$)	0.672 ± 0.015	0.713 ± 0.010	0.813 ± 0.025	0.532 ± 0.042
De-Sparse ($\lambda = 100.0$)	0.668 ± 0.018	0.714 ± 0.008	0.830 ± 0.026	0.506 ± 0.056
SVM	0.611 ± 0.026	0.619 ± 0.034	0.632 ± 0.050	0.590 ± 0.029
DIAG	0.568 ± 0.014	0.515 ± 0.026	0.460 ± 0.042	0.676 ± 0.046
Shrinkage ($\beta = 0.25$)	0.574 ± 0.014	0.538 ± 0.025	0.499 ± 0.041	0.649 ± 0.045
Shrinkage ($\beta = 0.5$)	0.560 ± 0.033	0.438 ± 0.220	0.413 ± 0.210	0.708 ± 0.152
Shrinkage ($\beta = 0.75$)	0.560 ± 0.025	0.480 ± 0.163	0.448 ± 0.158	0.672 ± 0.118
Days in Advance: 60				
AdaBoost ($\times 10$)	0.646 ± 0.021	0.596 ± 0.054	0.531 ± 0.095	0.762 ± 0.057
AdaBoost ($\times 50$)	0.639 ± 0.027	0.569 ± 0.083	0.491 ± 0.111	0.788 ± 0.060
CRDA ($\lambda = 1.0$)	0.654 ± 0.016	0.690 ± 0.016	0.774 ± 0.067	0.535 ± 0.088
CRDA ($\lambda = 10.0$)	0.653 ± 0.019	0.706 ± 0.010	0.833 ± 0.053	0.474 ± 0.083
CRDA ($\lambda = 100.0$)	0.643 ± 0.024	0.701 ± 0.028	0.844 ± 0.098	0.443 ± 0.124
LDA	0.556 ± 0.028	0.550 ± 0.042	0.547 ± 0.072	0.565 ± 0.065
Logistic Regression	0.631 ± 0.031	0.535 ± 0.108	0.447 ± 0.132	0.814 ± 0.073
De-Sparse ($\lambda = 1.0$)	0.655 ± 0.012	0.675 ± 0.023	0.723 ± 0.070	0.587 ± 0.074
De-Sparse ($\lambda = 10.0$)	0.661 ± 0.016	0.708 ± 0.009	0.823 ± 0.051	0.499 ± 0.077
De-Sparse ($\lambda = 100.0$)	0.649 ± 0.021	0.705 ± 0.020	0.844 ± 0.082	0.454 ± 0.110
SVM	0.627 ± 0.019	0.625 ± 0.027	0.625 ± 0.053	0.629 ± 0.056
DIAG	0.565 ± 0.011	0.514 ± 0.046	0.468 ± 0.076	0.662 ± 0.072
Shrinkage ($\beta = 0.25$)	0.568 ± 0.012	0.530 ± 0.040	0.492 ± 0.069	0.644 ± 0.063
Shrinkage ($\beta = 0.5$)	0.567 ± 0.013	0.528 ± 0.038	0.489 ± 0.067	0.646 ± 0.059
Shrinkage ($\beta = 0.75$)	0.561 ± 0.025	0.477 ± 0.164	0.444 ± 0.163	0.677 ± 0.120
Days in Advance: 90				
AdaBoost ($\times 10$)	0.627 ± 0.034	0.572 ± 0.063	0.507 ± 0.091	0.747 ± 0.054
AdaBoost ($\times 50$)	0.632 ± 0.035	0.575 ± 0.054	0.504 ± 0.077	0.759 ± 0.058
CRDA ($\lambda = 1.0$)	0.641 ± 0.018	0.663 ± 0.041	0.716 ± 0.106	0.566 ± 0.091
CRDA ($\lambda = 10.0$)	0.651 ± 0.018	0.693 ± 0.034	0.797 ± 0.093	0.505 ± 0.096
CRDA ($\lambda = 100.0$)	0.634 ± 0.040	0.675 ± 0.101	0.808 ± 0.188	0.459 ± 0.173
LDA	0.546 ± 0.025	0.532 ± 0.038	0.518 ± 0.058	0.574 ± 0.046
Logistic Regression	0.597 ± 0.058	0.423 ± 0.217	0.351 ± 0.207	0.843 ± 0.096
De-Sparse ($\lambda = 1.0$)	0.642 ± 0.022	0.663 ± 0.035	0.710 ± 0.078	0.574 ± 0.060
De-Sparse ($\lambda = 10.0$)	0.658 ± 0.016	0.696 ± 0.022	0.787 ± 0.073	0.528 ± 0.084
De-Sparse ($\lambda = 100.0$)	0.641 ± 0.031	0.683 ± 0.081	0.808 ± 0.164	0.475 ± 0.148
SVM	0.597 ± 0.034	0.600 ± 0.036	0.606 ± 0.047	0.587 ± 0.046
DIAG	0.568 ± 0.023	0.514 ± 0.048	0.464 ± 0.074	0.672 ± 0.066
Shrinkage ($\beta = 0.25$)	0.569 ± 0.020	0.530 ± 0.041	0.490 ± 0.065	0.648 ± 0.054
Shrinkage ($\beta = 0.5$)	0.565 ± 0.021	0.519 ± 0.041	0.473 ± 0.059	0.657 ± 0.044
Shrinkage ($\beta = 0.75$)	0.559 ± 0.019	0.511 ± 0.040	0.465 ± 0.061	0.653 ± 0.050

Bold entries show the best performance among different models

Table 2 Performance Comparison with Training Set: 100× 2, Testing Set: 2000× 2

	Accuracy	F1-Score	Sensitivity	Specificity
Days in Advance: 30				
AdaBoost (× 10)	0.632 ± 0.029	0.541 ± 0.095	0.452 ± 0.117	0.812 ± 0.065
AdaBoost (× 50)	0.631 ± 0.032	0.538 ± 0.099	0.447 ± 0.120	0.814 ± 0.062
CRDA ($\lambda = 1.0$)	0.674 ± 0.012	0.708 ± 0.019	0.792 ± 0.043	0.556 ± 0.029
CRDA ($\lambda = 10.0$)	0.675 ± 0.006	0.722 ± 0.008	0.844 ± 0.022	0.507 ± 0.017
CRDA ($\lambda = 100.0$)	0.664 ± 0.010	0.718 ± 0.004	0.858 ± 0.031	0.469 ± 0.048
LDA	0.594 ± 0.016	0.592 ± 0.019	0.591 ± 0.027	0.597 ± 0.018
Logistic Regression	0.593 ± 0.054	0.394 ± 0.200	0.305 ± 0.180	0.881 ± 0.075
<i>De-Sparse</i> ($\lambda = 1.0$)	0.674 ± 0.018	0.700 ± 0.025	0.765 ± 0.050	0.582 ± 0.026
<i>De-Sparse</i> ($\lambda = 10.0$)	0.681 ± 0.006	0.724 ± 0.006	0.838 ± 0.018	0.524 ± 0.020
<i>De-Sparse</i> ($\lambda = 100.0$)	0.668 ± 0.009	0.720 ± 0.006	0.854 ± 0.028	0.481 ± 0.041
SVM	0.636 ± 0.016	0.642 ± 0.024	0.655 ± 0.044	0.618 ± 0.025
DIAG	0.594 ± 0.019	0.562 ± 0.034	0.524 ± 0.050	0.663 ± 0.033
Shrinkage ($\beta = 0.25$)	0.600 ± 0.020	0.582 ± 0.031	0.559 ± 0.045	0.641 ± 0.022
Shrinkage ($\beta = 0.5$)	0.581 ± 0.044	0.467 ± 0.235	0.449 ± 0.228	0.714 ± 0.144
Shrinkage ($\beta = 0.75$)	0.599 ± 0.014	0.582 ± 0.020	0.559 ± 0.029	0.639 ± 0.022
Days in Advance: 60				
AdaBoost (× 10)	0.633 ± 0.024	0.537 ± 0.076	0.439 ± 0.110	0.827 ± 0.067
AdaBoost (× 50)	0.623 ± 0.024	0.507 ± 0.065	0.396 ± 0.089	0.850 ± 0.052
CRDA ($\lambda = 1.0$)	0.676 ± 0.016	0.711 ± 0.015	0.797 ± 0.041	0.555 ± 0.052
CRDA ($\lambda = 10.0$)	0.672 ± 0.019	0.719 ± 0.015	0.837 ± 0.025	0.508 ± 0.039
CRDA ($\lambda = 100.0$)	0.668 ± 0.017	0.716 ± 0.013	0.838 ± 0.038	0.498 ± 0.054
LDA	0.603 ± 0.024	0.599 ± 0.026	0.595 ± 0.033	0.610 ± 0.032
Logistic Regression	0.613 ± 0.042	0.462 ± 0.164	0.362 ± 0.147	0.863 ± 0.069
<i>De-Sparse</i> ($\lambda = 1.0$)	0.679 ± 0.011	0.707 ± 0.014	0.776 ± 0.041	0.582 ± 0.043
<i>De-Sparse</i> ($\lambda = 10.0$)	0.676 ± 0.016	0.720 ± 0.012	0.834 ± 0.026	0.518 ± 0.039
<i>De-Sparse</i> ($\lambda = 100.0$)	0.671 ± 0.017	0.718 ± 0.012	0.838 ± 0.029	0.504 ± 0.045
SVM	0.644 ± 0.016	0.645 ± 0.020	0.650 ± 0.038	0.637 ± 0.037
DIAG	0.596 ± 0.015	0.562 ± 0.033	0.522 ± 0.058	0.670 ± 0.054
Shrinkage ($\beta = 0.25$)	0.600 ± 0.016	0.580 ± 0.024	0.554 ± 0.040	0.645 ± 0.038
Shrinkage ($\beta = 0.5$)	0.596 ± 0.035	0.532 ± 0.178	0.513 ± 0.174	0.680 ± 0.113
Shrinkage ($\beta = 0.75$)	0.596 ± 0.039	0.532 ± 0.179	0.513 ± 0.175	0.678 ± 0.115
Days in Advance: 90				
AdaBoost (× 10)	0.626 ± 0.022	0.519 ± 0.061	0.412 ± 0.093	0.840 ± 0.058
AdaBoost (× 50)	0.631 ± 0.017	0.523 ± 0.056	0.413 ± 0.087	0.849 ± 0.053
CRDA ($\lambda = 1.0$)	0.674 ± 0.013	0.709 ± 0.020	0.796 ± 0.052	0.552 ± 0.047
CRDA ($\lambda = 10.0$)	0.674 ± 0.010	0.721 ± 0.006	0.845 ± 0.021	0.502 ± 0.034
CRDA ($\lambda = 100.0$)	0.666 ± 0.015	0.719 ± 0.006	0.856 ± 0.025	0.477 ± 0.052
LDA	0.605 ± 0.017	0.607 ± 0.026	0.612 ± 0.045	0.598 ± 0.028
Logistic Regression	0.611 ± 0.036	0.453 ± 0.130	0.345 ± 0.136	0.876 ± 0.067
<i>De-Sparse</i> ($\lambda = 1.0$)	0.675 ± 0.013	0.700 ± 0.026	0.764 ± 0.061	0.587 ± 0.045
<i>De-Sparse</i> ($\lambda = 10.0$)	0.682 ± 0.007	0.725 ± 0.007	0.840 ± 0.025	0.523 ± 0.030
<i>De-Sparse</i> ($\lambda = 100.0$)	0.669 ± 0.013	0.721 ± 0.006	0.853 ± 0.023	0.486 ± 0.046
SVM	0.632 ± 0.017	0.638 ± 0.023	0.649 ± 0.039	0.616 ± 0.026
DIAG	0.597 ± 0.015	0.574 ± 0.039	0.549 ± 0.072	0.644 ± 0.063
Shrinkage ($\beta = 0.25$)	0.593 ± 0.034	0.531 ± 0.179	0.517 ± 0.182	0.668 ± 0.120
Shrinkage ($\beta = 0.5$)	0.602 ± 0.015	0.589 ± 0.028	0.575 ± 0.053	0.628 ± 0.043
Shrinkage ($\beta = 0.75$)	0.599 ± 0.015	0.586 ± 0.025	0.570 ± 0.045	0.629 ± 0.037

Bold entries show the best performance among different models

Table 3 Performance Comparison with Training Set: 150 × 2, Testing Set: 2000 × 2

	Accuracy	F1-Score	Sensitivity	Specificity
Days in Advance: 30				
AdaBoost (× 10)	0.615 ± 0.010	0.484 ± 0.033	0.363 ± 0.039	0.867 ± 0.024
AdaBoost (× 50)	0.615 ± 0.007	0.482 ± 0.025	0.359 ± 0.032	0.871 ± 0.023
CRDA ($\lambda = 1.0$)	0.682 ± 0.008	0.723 ± 0.008	0.829 ± 0.021	0.534 ± 0.019
CRDA ($\lambda = 10.0$)	0.671 ± 0.013	0.721 ± 0.008	0.851 ± 0.016	0.490 ± 0.035
CRDA ($\lambda = 100.0$)	0.662 ± 0.014	0.718 ± 0.007	0.861 ± 0.020	0.464 ± 0.044
LDA	0.613 ± 0.012	0.611 ± 0.018	0.610 ± 0.038	0.615 ± 0.037
Logistic Regression	0.581 ± 0.045	0.352 ± 0.189	0.255 ± 0.142	0.908 ± 0.053
<i>De-Sparse</i> ($\lambda = 1.0$)	0.681 ± 0.009	0.712 ± 0.012	0.790 ± 0.028	0.572 ± 0.020
<i>De-Sparse</i> ($\lambda = 10.0$)	0.681 ± 0.007	0.727 ± 0.006	0.849 ± 0.013	0.512 ± 0.019
<i>De-Sparse</i> ($\lambda = 100.0$)	0.667 ± 0.013	0.720 ± 0.007	0.857 ± 0.020	0.478 ± 0.041
SVM	0.650 ± 0.012	0.660 ± 0.014	0.680 ± 0.024	0.620 ± 0.023
DIAG	0.619 ± 0.014	0.610 ± 0.031	0.600 ± 0.056	0.637 ± 0.037
Shrinkage ($\beta = 0.25$)	0.599 ± 0.051	0.500 ± 0.251	0.503 ± 0.256	0.696 ± 0.156
Shrinkage ($\beta = 0.5$)	0.611 ± 0.039	0.562 ± 0.189	0.566 ± 0.195	0.656 ± 0.121
Shrinkage ($\beta = 0.75$)	0.615 ± 0.009	0.611 ± 0.024	0.608 ± 0.051	0.623 ± 0.045
Days in Advance: 60				
AdaBoost (× 10)	0.625 ± 0.039	0.512 ± 0.131	0.424 ± 0.156	0.826 ± 0.081
AdaBoost (× 50)	0.637 ± 0.024	0.554 ± 0.072	0.466 ± 0.113	0.809 ± 0.068
CRDA ($\lambda = 1.0$)	0.677 ± 0.017	0.717 ± 0.015	0.818 ± 0.028	0.536 ± 0.032
CRDA ($\lambda = 10.0$)	0.671 ± 0.012	0.721 ± 0.008	0.848 ± 0.022	0.494 ± 0.038
CRDA ($\lambda = 100.0$)	0.662 ± 0.014	0.718 ± 0.006	0.861 ± 0.031	0.463 ± 0.055
LDA	0.623 ± 0.014	0.621 ± 0.023	0.619 ± 0.040	0.627 ± 0.023
Logistic Regression	0.600 ± 0.054	0.412 ± 0.217	0.331 ± 0.195	0.869 ± 0.090
<i>De-Sparse</i> ($\lambda = 1.0$)	0.681 ± 0.016	0.711 ± 0.016	0.787 ± 0.033	0.574 ± 0.036
<i>De-Sparse</i> ($\lambda = 10.0$)	0.678 ± 0.011	0.724 ± 0.009	0.843 ± 0.017	0.513 ± 0.023
<i>De-Sparse</i> ($\lambda = 100.0$)	0.667 ± 0.014	0.720 ± 0.007	0.856 ± 0.028	0.477 ± 0.050
SVM	0.649 ± 0.017	0.654 ± 0.025	0.665 ± 0.042	0.633 ± 0.024
DIAG	0.615 ± 0.018	0.597 ± 0.032	0.574 ± 0.054	0.656 ± 0.045
Shrinkage ($\beta = 0.25$)	0.618 ± 0.018	0.605 ± 0.031	0.587 ± 0.051	0.649 ± 0.039
Shrinkage ($\beta = 0.5$)	0.608 ± 0.039	0.548 ± 0.184	0.533 ± 0.181	0.683 ± 0.110
Shrinkage ($\beta = 0.75$)	0.618 ± 0.015	0.602 ± 0.027	0.581 ± 0.045	0.655 ± 0.033
Days in Advance: 90				
AdaBoost (× 10)	0.630 ± 0.023	0.531 ± 0.075	0.436 ± 0.123	0.824 ± 0.082
AdaBoost (× 50)	0.630 ± 0.023	0.534 ± 0.078	0.441 ± 0.126	0.820 ± 0.083
CRDA ($\lambda = 1.0$)	0.674 ± 0.012	0.708 ± 0.017	0.794 ± 0.045	0.553 ± 0.039
CRDA ($\lambda = 10.0$)	0.671 ± 0.011	0.720 ± 0.007	0.845 ± 0.021	0.498 ± 0.035
CRDA ($\lambda = 100.0$)	0.663 ± 0.013	0.718 ± 0.004	0.857 ± 0.025	0.470 ± 0.050
LDA	0.611 ± 0.020	0.610 ± 0.025	0.608 ± 0.039	0.614 ± 0.024
Logistic Regression	0.614 ± 0.045	0.463 ± 0.174	0.374 ± 0.180	0.853 ± 0.098
<i>De-Sparse</i> ($\lambda = 1.0$)	0.672 ± 0.018	0.693 ± 0.030	0.745 ± 0.065	0.600 ± 0.042
<i>De-Sparse</i> ($\lambda = 10.0$)	0.678 ± 0.010	0.722 ± 0.009	0.836 ± 0.026	0.521 ± 0.033
<i>De-Sparse</i> ($\lambda = 100.0$)	0.668 ± 0.010	0.720 ± 0.005	0.851 ± 0.022	0.485 ± 0.039
SVM	0.639 ± 0.015	0.645 ± 0.020	0.657 ± 0.035	0.622 ± 0.026
DIAG	0.610 ± 0.012	0.602 ± 0.022	0.590 ± 0.042	0.631 ± 0.031
Shrinkage ($\beta = 0.25$)	0.613 ± 0.011	0.608 ± 0.019	0.601 ± 0.036	0.626 ± 0.027
Shrinkage ($\beta = 0.5$)	0.602 ± 0.036	0.547 ± 0.183	0.540 ± 0.183	0.665 ± 0.114
Shrinkage ($\beta = 0.75$)	0.601 ± 0.036	0.545 ± 0.183	0.536 ± 0.182	0.665 ± 0.113

Bold entries show the best performance among different models

Table 4 Performance Comparison with Training Set: 200×2 , Testing Set: 2000×2

	Accuracy	F1-Score	Sensitivity	Specificity
Days in Advance: 30				
AdaBoost ($\times 10$)	0.618 ± 0.026	0.485 ± 0.082	0.373 ± 0.115	0.863 ± 0.064
AdaBoost ($\times 50$)	0.618 ± 0.022	0.491 ± 0.064	0.377 ± 0.092	0.859 ± 0.052
CRDA ($\lambda = 1.0$)	0.688 ± 0.006	0.725 ± 0.007	0.824 ± 0.017	0.553 ± 0.016
CRDA ($\lambda = 10.0$)	0.680 ± 0.005	0.725 ± 0.005	0.847 ± 0.013	0.513 ± 0.013
CRDA ($\lambda = 100.0$)	0.669 ± 0.011	0.721 ± 0.003	0.855 ± 0.026	0.483 ± 0.047
LDA	0.637 ± 0.006	0.644 ± 0.010	0.655 ± 0.021	0.620 ± 0.020
Logistic Regression	0.598 ± 0.046	0.411 ± 0.175	0.313 ± 0.159	0.883 ± 0.070
<i>De-Sparse</i> ($\lambda = 1.0$)	0.686 ± 0.007	0.717 ± 0.007	0.794 ± 0.017	0.578 ± 0.019
<i>De-Sparse</i> ($\lambda = 10.0$)	0.684 ± 0.006	0.729 ± 0.005	0.850 ± 0.007	0.519 ± 0.010
<i>De-Sparse</i> ($\lambda = 100.0$)	0.673 ± 0.009	0.723 ± 0.004	0.852 ± 0.024	0.494 ± 0.038
SVM	0.660 ± 0.012	0.671 ± 0.012	0.693 ± 0.014	0.626 ± 0.015
DIAG	0.623 ± 0.013	0.603 ± 0.024	0.575 ± 0.041	0.671 ± 0.029
Shrinkage ($\beta = 0.25$)	0.628 ± 0.013	0.621 ± 0.023	0.610 ± 0.039	0.646 ± 0.024
Shrinkage ($\beta = 0.5$)	0.619 ± 0.042	0.565 ± 0.190	0.560 ± 0.190	0.678 ± 0.110
Shrinkage ($\beta = 0.75$)	0.633 ± 0.012	0.629 ± 0.019	0.624 ± 0.034	0.642 ± 0.022
Days in Advance: 60				
AdaBoost ($\times 10$)	0.605 ± 0.023	0.445 ± 0.085	0.325 ± 0.074	0.885 ± 0.033
AdaBoost ($\times 50$)	0.616 ± 0.010	0.479 ± 0.038	0.356 ± 0.048	0.876 ± 0.032
CRDA ($\lambda = 1.0$)	0.684 ± 0.006	0.721 ± 0.006	0.818 ± 0.019	0.549 ± 0.023
CRDA ($\lambda = 10.0$)	0.674 ± 0.008	0.722 ± 0.006	0.844 ± 0.019	0.505 ± 0.026
CRDA ($\lambda = 100.0$)	0.673 ± 0.010	0.721 ± 0.006	0.845 ± 0.021	0.502 ± 0.035
LDA	0.626 ± 0.009	0.622 ± 0.013	0.616 ± 0.028	0.635 ± 0.031
Logistic Regression	0.589 ± 0.038	0.380 ± 0.151	0.270 ± 0.113	0.908 ± 0.038
<i>De-Sparse</i> ($\lambda = 1.0$)	0.684 ± 0.010	0.710 ± 0.012	0.773 ± 0.027	0.595 ± 0.023
<i>De-Sparse</i> ($\lambda = 10.0$)	0.682 ± 0.006	0.726 ± 0.007	0.844 ± 0.017	0.520 ± 0.014
<i>De-Sparse</i> ($\lambda = 100.0$)	0.675 ± 0.008	0.722 ± 0.006	0.843 ± 0.022	0.508 ± 0.031
SVM	0.651 ± 0.006	0.659 ± 0.010	0.675 ± 0.026	0.626 ± 0.028
DIAG	0.627 ± 0.012	0.615 ± 0.023	0.597 ± 0.045	0.657 ± 0.039
Shrinkage ($\beta = 0.25$)	0.618 ± 0.041	0.562 ± 0.188	0.553 ± 0.187	0.683 ± 0.111
Shrinkage ($\beta = 0.5$)	0.620 ± 0.040	0.565 ± 0.189	0.557 ± 0.187	0.683 ± 0.110
Shrinkage ($\beta = 0.75$)	0.616 ± 0.039	0.557 ± 0.186	0.544 ± 0.183	0.688 ± 0.109
Days in Advance: 90				
AdaBoost ($\times 10$)	0.626 ± 0.033	0.507 ± 0.107	0.411 ± 0.153	0.840 ± 0.088
AdaBoost ($\times 50$)	0.632 ± 0.028	0.533 ± 0.092	0.441 ± 0.135	0.823 ± 0.080
CRDA ($\lambda = 1.0$)	0.682 ± 0.008	0.722 ± 0.008	0.825 ± 0.017	0.540 ± 0.020
CRDA ($\lambda = 10.0$)	0.664 ± 0.012	0.718 ± 0.006	0.856 ± 0.025	0.472 ± 0.044
CRDA ($\lambda = 100.0$)	0.656 ± 0.016	0.715 ± 0.005	0.865 ± 0.029	0.447 ± 0.058
LDA	0.631 ± 0.014	0.630 ± 0.018	0.631 ± 0.034	0.630 ± 0.032
Logistic Regression	0.605 ± 0.060	0.424 ± 0.232	0.353 ± 0.222	0.857 ± 0.107
<i>De-Sparse</i> ($\lambda = 1.0$)	0.684 ± 0.010	0.714 ± 0.014	0.789 ± 0.031	0.579 ± 0.020
<i>De-Sparse</i> ($\lambda = 10.0$)	0.676 ± 0.008	0.724 ± 0.004	0.852 ± 0.019	0.500 ± 0.030
<i>De-Sparse</i> ($\lambda = 100.0$)	0.658 ± 0.015	0.716 ± 0.005	0.863 ± 0.029	0.452 ± 0.057
SVM	0.657 ± 0.009	0.669 ± 0.015	0.693 ± 0.031	0.621 ± 0.024
DIAG	0.625 ± 0.013	0.614 ± 0.029	0.601 ± 0.055	0.648 ± 0.045
Shrinkage ($\beta = 0.25$)	0.627 ± 0.014	0.617 ± 0.030	0.604 ± 0.056	0.651 ± 0.043
Shrinkage ($\beta = 0.5$)	0.626 ± 0.013	0.616 ± 0.027	0.603 ± 0.051	0.650 ± 0.042
Shrinkage ($\beta = 0.75$)	0.626 ± 0.014	0.617 ± 0.023	0.604 ± 0.045	0.649 ± 0.043

Bold entries show the best performance among different models

Table 5 Performance Comparison with Training Set: 250 × 2, Testing Set: 2000 × 2

	Accuracy	F1-Score	Sensitivity	Specificity
Days in Advance: 30				
AdaBoost (× 10)	0.620 ± 0.037	0.484 ± 0.110	0.380 ± 0.147	0.860 ± 0.076
AdaBoost (× 50)	0.625 ± 0.033	0.499 ± 0.097	0.394 ± 0.138	0.856 ± 0.074
CRDA ($\lambda = 1.0$)	0.689 ± 0.010	0.726 ± 0.009	0.824 ± 0.021	0.553 ± 0.025
CRDA ($\lambda = 10.0$)	0.677 ± 0.012	0.722 ± 0.009	0.840 ± 0.020	0.513 ± 0.029
CRDA ($\lambda = 100.0$)	0.666 ± 0.014	0.719 ± 0.007	0.853 ± 0.027	0.479 ± 0.050
LDA	0.644 ± 0.009	0.645 ± 0.012	0.648 ± 0.023	0.640 ± 0.020
Logistic Regression	0.605 ± 0.057	0.424 ± 0.204	0.339 ± 0.200	0.870 ± 0.089
<i>De-Sparse</i> ($\lambda = 1.0$)	0.690 ± 0.007	0.719 ± 0.007	0.791 ± 0.022	0.589 ± 0.027
<i>De-Sparse</i> ($\lambda = 10.0$)	0.684 ± 0.009	0.726 ± 0.008	0.837 ± 0.015	0.531 ± 0.012
<i>De-Sparse</i> ($\lambda = 100.0$)	0.671 ± 0.012	0.721 ± 0.008	0.848 ± 0.026	0.494 ± 0.039
SVM	0.663 ± 0.013	0.673 ± 0.015	0.694 ± 0.024	0.632 ± 0.023
DIAG	0.633 ± 0.011	0.619 ± 0.028	0.599 ± 0.055	0.668 ± 0.046
Shrinkage ($\beta = 0.25$)	0.625 ± 0.044	0.569 ± 0.192	0.562 ± 0.193	0.689 ± 0.108
Shrinkage ($\beta = 0.5$)	0.626 ± 0.044	0.569 ± 0.192	0.561 ± 0.192	0.691 ± 0.106
Shrinkage ($\beta = 0.75$)	0.639 ± 0.011	0.633 ± 0.022	0.624 ± 0.039	0.653 ± 0.025
Days in Advance: 60				
AdaBoost (× 10)	0.635 ± 0.026	0.539 ± 0.087	0.449 ± 0.141	0.820 ± 0.091
AdaBoost (× 50)	0.634 ± 0.027	0.536 ± 0.089	0.445 ± 0.144	0.823 ± 0.091
CRDA ($\lambda = 1.0$)	0.692 ± 0.006	0.729 ± 0.006	0.827 ± 0.014	0.557 ± 0.015
CRDA ($\lambda = 10.0$)	0.682 ± 0.008	0.730 ± 0.004	0.860 ± 0.019	0.504 ± 0.031
CRDA ($\lambda = 100.0$)	0.674 ± 0.014	0.726 ± 0.005	0.864 ± 0.025	0.483 ± 0.051
LDA	0.642 ± 0.011	0.643 ± 0.015	0.645 ± 0.025	0.638 ± 0.017
Logistic Regression	0.623 ± 0.048	0.489 ± 0.184	0.411 ± 0.195	0.835 ± 0.105
<i>De-Sparse</i> ($\lambda = 1.0$)	0.691 ± 0.008	0.717 ± 0.009	0.781 ± 0.019	0.601 ± 0.017
<i>De-Sparse</i> ($\lambda = 10.0$)	0.689 ± 0.004	0.733 ± 0.004	0.854 ± 0.013	0.524 ± 0.017
<i>De-Sparse</i> ($\lambda = 100.0$)	0.676 ± 0.013	0.727 ± 0.005	0.863 ± 0.024	0.488 ± 0.048
SVM	0.662 ± 0.008	0.668 ± 0.012	0.681 ± 0.023	0.642 ± 0.017
DIAG	0.634 ± 0.012	0.613 ± 0.026	0.582 ± 0.053	0.687 ± 0.049
Shrinkage ($\beta = 0.25$)	0.627 ± 0.044	0.565 ± 0.189	0.545 ± 0.185	0.709 ± 0.101
Shrinkage ($\beta = 0.5$)	0.642 ± 0.010	0.634 ± 0.015	0.620 ± 0.028	0.663 ± 0.028
Shrinkage ($\beta = 0.75$)	0.641 ± 0.010	0.636 ± 0.012	0.627 ± 0.021	0.655 ± 0.022
Days in Advance: 90				
AdaBoost (× 10)	0.633 ± 0.027	0.536 ± 0.089	0.447 ± 0.140	0.818 ± 0.086
AdaBoost (× 50)	0.631 ± 0.026	0.535 ± 0.087	0.445 ± 0.137	0.818 ± 0.085
CRDA ($\lambda = 1.0$)	0.686 ± 0.006	0.721 ± 0.009	0.813 ± 0.029	0.558 ± 0.026
CRDA ($\lambda = 10.0$)	0.675 ± 0.007	0.720 ± 0.006	0.838 ± 0.021	0.512 ± 0.028
CRDA ($\lambda = 100.0$)	0.671 ± 0.009	0.719 ± 0.004	0.844 ± 0.028	0.497 ± 0.043
LDA	0.648 ± 0.009	0.648 ± 0.018	0.651 ± 0.037	0.644 ± 0.025
Logistic Regression	0.628 ± 0.028	0.520 ± 0.095	0.427 ± 0.146	0.828 ± 0.090
<i>De-Sparse</i> ($\lambda = 1.0$)	0.687 ± 0.009	0.713 ± 0.014	0.778 ± 0.033	0.597 ± 0.022
<i>De-Sparse</i> ($\lambda = 10.0$)	0.683 ± 0.006	0.725 ± 0.008	0.839 ± 0.021	0.527 ± 0.018
<i>De-Sparse</i> ($\lambda = 100.0$)	0.673 ± 0.008	0.720 ± 0.005	0.841 ± 0.024	0.505 ± 0.037
SVM	0.666 ± 0.009	0.672 ± 0.014	0.687 ± 0.030	0.644 ± 0.023
DIAG	0.635 ± 0.015	0.621 ± 0.030	0.601 ± 0.053	0.668 ± 0.032
Shrinkage ($\beta = 0.25$)	0.638 ± 0.012	0.631 ± 0.027	0.621 ± 0.051	0.656 ± 0.032
Shrinkage ($\beta = 0.5$)	0.642 ± 0.011	0.635 ± 0.026	0.626 ± 0.050	0.657 ± 0.032
Shrinkage ($\beta = 0.75$)	0.641 ± 0.010	0.635 ± 0.024	0.628 ± 0.046	0.655 ± 0.030

Bold entries show the best performance among different models

compared to typical LDA. On opposite side of the trade-off, when compared to CRDA (based on graphical lasso), *De-Sparse* on average gains 2.3% higher specificity while sacrificing 1.4% sensitivity.

4.3.3 Discussion on the performance comparison

We consider testing accuracy and F1-score as two primary metrics for the evaluation, as these two metrics well characterize the performance of classifiers. Thus, we conclude that *De-Sparse* overall outperforms the baseline algorithms, including both LDA, SVM, Logistic Regression, and other classifiers, in all experimental settings. In terms of the trade-off between sensitivity and specificity, we argue that *De-Sparse* still outperforms the original LDA classifier and CRDA classifiers, considering the requests of predictive analytics of diseases and the early diagnosis. While the original LDA classifier well-balances the sensitivity and specificity, both CRDA and *De-Sparse* would incorporate slightly higher sensitivity, compared to the original LDA, while having lower specificity. In this way, CRDA and *De-Sparse* could discover more patients potentially with the diseases, but also slightly raise the frequency of false alarms. We believe, compared to the marginal increase of false alarms, the improvement of sensitivity should be appreciated in medical contexts. Compared *De-Sparse* to CRDA, the de-sparsified Graphical Lasso here helps *De-Sparse* achieve higher overall accuracy and F1-score with a more balanced pair of sensitivity and specificity.

4.4 Empirical convergence of parameter estimation

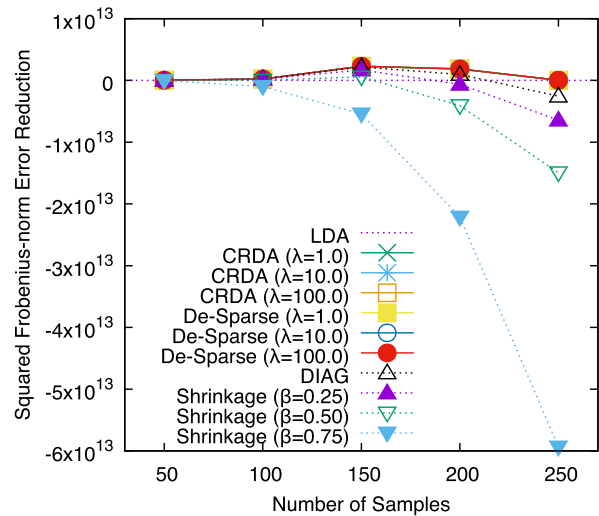
We hypothesize *De-Sparse* improves LDA because that the de-sparsified Graphical Lasso used in *De-Sparse* approaches to the inverse of population covariance matrix has a tighter error bound than the inverse of sample covariance matrix used in simple LDA models, when the training sample size is limited. In order to verify our hypothesis, we compare the inverse covariance matrix estimators used in *De-Sparse*, CRDA, and other LDA baselines, using the EHR data. Specifically, we (1) learned a “ground truth” covariance matrix Σ_{GT} (and its inverse $\Theta_{GT} = \Sigma_{GT}^{-1}$) using diagnosis-frequency vectors of 10,000 patients (w/o the target disease, balanced) randomly retrieved from all patients of 22 U.S. university healthcare systems, (2) randomly selected another 50 to 250 samples (w/o the target disease, balanced) to train LDA, *De-Sparse*, CRDA and Shrinkage, (3) estimated the error between the inverse covariance matrix (denoted as Θ , $\Theta = \hat{\Theta}$ for CRDA, $\Theta = \hat{T}$ for *De-Sparse*, and $\Theta = \Theta_\beta$ for Shrinkage LDA) learned in each classifier versus the inverse of “ground truth” covariance matrix Σ_{GT} , all in ℓ_2 -norm, and (4) further estimated the error reduction

of Θ from the inverse of sample covariance estimation (i.e., $\bar{\Theta} = \bar{\Sigma}^{-1}$) as

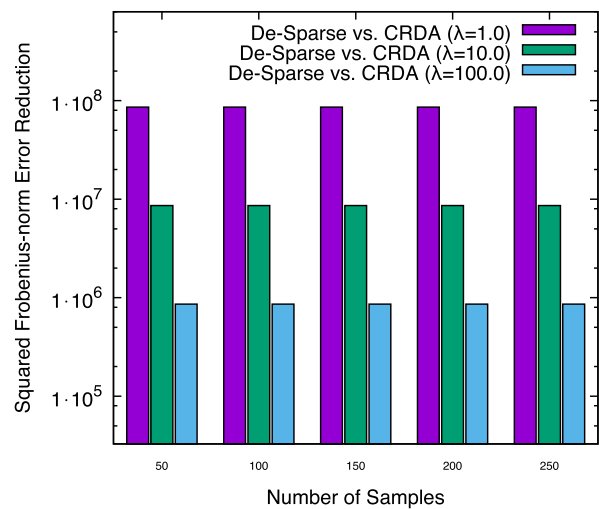
$$R(\Theta) = \|\Theta - \bar{\Theta}\|_2^2 - \|\Theta_{GT} - \bar{\Theta}\|_2^2, \quad (20)$$

where $\Theta = \hat{\Theta}$ for CRDA, $\Theta = \hat{T}$ for *De-Sparse*, and $\Theta = \Theta_\beta$ for Shrinkage LDA. We repeated above (1)–(4) steps for 100 times, and illustrated the average error reduction $R(\Theta)$ in Fig. 2a, with varying parameters and settings.

Figure 2a demonstrates that, estimators used in *De-Sparse* (\hat{T}) and CRDA ($\hat{\Theta}$) outperform the sample estimation in all settings, while DIAG and Shrinkage estimators (i.e., Θ_β and $\beta = 0.0, 0.25, 0.5$, and 0.75) may cause even higher estimation error (with negative error reduction) when the number of samples increases. Figure 2b illustrates the trend of error reduction with CRDA and *De-Sparse*. Though the difference between these two



(a) Error Reduction of All Estimators Beyond Sample Estimation



(b) Zoom-in on *De-Sparse* vs. CRDA

Fig. 2 Comparison on ℓ_2 -norm Estimator Error Reduction of Inverse Covariance Matrices Estimation (the higher the better)

algorithms is not visible in such scale, we can observe that these two algorithms achieve the maximal error reduction when number of samples is 150 in our experiments, while the error reduction is low when the number of samples is relatively small (50) or large (250). Because, when the sample size is small, both sample-based estimation ($\hat{\Theta}$) and the regularized estimation (\hat{T} and $\hat{\Theta}$) work poorly, though \hat{T} and $\hat{\Theta}$ still outperform $\hat{\Theta}$. With the increasing sample size, the advantage of CRDA and *De-Sparse* becomes more and more significant. However, when sample size is large, both sample-based estimation and the regularized estimation converge well, thus the error reduction becomes marginal.

5 Conclusion

In this paper, we study the long existing problem of covariance-regularized discriminant analysis for classification under high-dimensional low sample sizes (HDLSS) settings. More specific, we take care of the applications to the predictive analytics of diseases using Electronic Health Records (EHRs) data and common diagnosis-frequency data representation. To understand the performance of LDA, we extend the existing theory [14, 25], and propose a novel analytical model characterizing the error rate of LDA classification under the uncertainty of parameter estimation. Based on the analytical model, we propose *De-Sparse* – a novel LDA classifier using de-sparsified Graphical Lasso. Our analysis shows that the proposed algorithm could outperform the existing Covariance-regularized discriminant analysis (CRDA) based on common Graphical Lasso. The experimental results on real-world Electronic Health Record (EHR) datasets show *De-Sparse* outperforms all baseline algorithms. We interpret the comparison of results and demonstrate the advantage of proposed methods in medicare settings. Further, the empirical studies on estimator comparison validate our analysis.

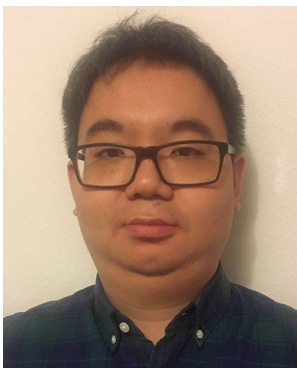
References

1. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, Hoboken
2. Peck R, Ness JV (1982) The use of shrinkage estimators in linear discriminant analysis. *IEEE Trans Pattern Anal Mach Intell* 5:530–537
3. Xiong H, Cheng W, Bian J, Hu W, Sun Z, Guo Z (2018) DBSDA Lowering the bound of misclassification rate for sparse linear discriminant analysis via model debiasing. *IEEE Trans Neural Netwo Learning Sys* 30(3):707–717
4. Buhlmann P, Van De Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Berlin
5. Krzanowski WJ, Jonathan P, McCarthy WV, Thomas MR (1995) Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Appl Stat*, pp 101–115
6. Belhumeur PN, Hespanha JP, Kriegman DJ (1996) Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In: *ECCV* (1), vol 1064. Springer, pp 45–58
7. Ye J, Janardan R, Li Q (2004) Two-dimensional linear discriminant analysis. In: *NIPS*, Cambridge, MA, USA, pp 1569–1576
8. Tikhonov AN (1943) On the stability of inverse problems. In: *Dokl. Akad. Nauk SSSR*, vol 39, pp 195–198
9. Witten DM, Tibshirani R (2009) Covariance-regularized regression and classification for high dimensional problems. *J Royal Stat Soc: Series B (Statistical Methodology)* 71(3):615–636
10. Clemmensen L, Hastie T, Witten D, Ersbøll B (2011) Sparse discriminant analysis. *Technometrics*, 53(4)
11. Shao J, Wang Y, Deng X, Wang S, et al. (2011) Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann Stat* 39(2):1241–1265
12. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441
13. Cai TT, Ren Z, Zhou HH, et al. (2016) Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics* 10(1):1–59
14. Zollanvari A, Dougherty ER (2013) Random matrix theory in pattern classification An application to error estimation. In: 2013 Asilomar Conference on Signals, Systems and Computers
15. Marčenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* 1(4):457
16. Jain M (2001) Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, pp 295–327
17. Rothman AJ, Bickel PJ, Levina E, Zhu J, et al. (2008) Sparse permutation invariant covariance estimation. *Electron J Stat* 2:494–515
18. Yadav P, Steinbach M, Kumar V, Simon G (2018) Mining electronic health records (ehrs) a survey. *ACM Computing Surveys (CSUR)* 50(6):1–40
19. Wang F, Sun J (2015) Psf: A unified patient similarity evaluation framework through metric learning with weak supervision. *IEEE J Biomed Health Informatics* 19(3):1053–1060
20. Sun J, Wang F, Hu J, Edabollahi S (2012) Supervised patient similarity measure of heterogeneous patient records. *ACM SIGKDD Explorations Newsletter* 14(1):16–24
21. Ng K, Sun J, Hu J, Wang F (2015) Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summit on Clinical Research Informatics (CRI)*
22. Zhang J, Xiong H, Huang Y, Wu H, Leach K, Barnes L (2015) MSEQ Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data. In: 2015 International Conference on Big Data (Workshop), IEEE
23. Jensen S, SPSS UK (2001) Mining medical data for predictive and sequential patterns: Pkdd 2001. In: *Proceedings of the 5th European conference on principles and practice of knowledge discovery in databases*
24. Liu C, Wang F, Hu J, Xiong H (2015) Temporal phenotyping from longitudinal electronic health records: A graph based framework. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15*. ACM, New York, pp 705–714
25. Lachenbruch PA, Mickey RM (1968) Estimation of error rates in discriminant analysis. *Technometrics* 10(1):1–11

26. Bian J, Barnes L, Chen G, Xiong H (2017) Early detection of diseases using electronic health records data and covariance-regularized linear discriminant analysis. In: IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), p 2017
27. Jankova J, van de Geer S et al (2015) Confidence intervals for high-dimensional inverse covariance estimation. *Electronic J Stat* 9(1):1205–1229
28. Turner JC, Keller A (2015) College Health Surveillance Network: Epidemiology and Health Care Utilization of College Students at U.S. 4-Year Universities. *Journal of American College Health*, pp 530–538
29. Van Vleck TT, Elhadad N (2010) Corpus-based problem selection for ehr note summarization. In: AMIA Annual symposium proceedings, vol 2010, p 817. American Medical Informatics Association
30. Yu S, Berry D, Bisbal J (2011) Performance analysis and assessment of a tf-idf based archetype-snomed-ct binding algorithm. In: 2011 24th International Symposium on Computer-Based Medical Systems (CBMS). IEEE, pp 1–6
31. Shen F, Sohn S, Rastegar-Mojarad M, Liu S, Pankratz JJ, Hatton MA, Sowada N, Shrestha OK, Shurson SL, Liu H (2017) Populating physician biographical pages based on EMR data. *AMIA Summits on Translational Science Proceedings* 2017:522
32. Luhn HP (1957) A statistical approach to mechanized encoding and searching of literary information. *IBM J Res Dev* 1(4):309–317
33. Jones KS (1972) A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*
34. Aizawa A (2003) An information-theoretic perspective of tf-idf measures. *Information Processing & Management* 39(1):45–65
35. Dubberke ER, Reske KA, McDonald LC, Fraser VJ (2006) Icd-9 codes and surveillance for clostridium difficile-associated disease. *Emerging Infectious Diseases* 12(10):1576
36. Kowsari K, Meimandi KJ, Heidarysafa M, Mendu S, Barnes L, Brown D (2019) Text classification algorithms: A survey. *Information* 10(4):150
37. Choi E, Bahadori MT, Searles E, Coffey C, Thompson M, Bost J, Tejedor-Sojo J, Sun J (2016) Multi-layer representation learning for medical concepts. In: Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining, pp 1495–1504
38. Zhang J, Kowsari K, Harrison JH, Lobo JM, Barnes LE (2018) Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access* 6:65333–65346
39. Choi E, Bahadori MT, Le S, Stewart WF, Sun J (2017) Gram: graph-based attention model for healthcare representation learning. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 787–795
40. Bai T, Zhang S, Egleston BL, Vucetic S (2018) Interpretable representation learning for healthcare via capturing disease progression through time. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 43–51
41. Ma T, Xiao C, Wang F (2018) Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. In: Proceedings of the 2018 SIAM International Conference on Data Mining. SIAM, pp 261–269
42. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, et al. (2018) Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 1(1):18
43. Shickel B, Tighe PJ, Bihorac A, Rashidi P (2017) Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE J Biomed Health Informatics* 22(5):1589–1604
44. Solares JRA, Raimondi FED, Zhu Y, Rahimian F, Canoy D, Tran J, Gomes ACP, Payberah AH, Zottoli M, Nazarzadeh M, et al. (2020) Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J Biomed Inform* 101:103337
45. HCUP (2014) Appendix a - clinical classification software-diagnoses
46. Sun L, Zhang X, Qian Y, Xu J, Zhang S (2019) Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. *Inform Sci* 502:18–41
47. Sun L, Zhang X, Qian Y, Xu J, Zhang S, Tian Y (2019) Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. *Appl Intell* 49(4):1245–1259
48. Chen L, Wang S (2012) Automated feature weighting in naive bayes for high-dimensional data classification. In: Proceedings of the 21st ACM International conference on information and knowledge management, pp 1243–1252
49. Wan H, Wang H, Guo G, Wei X (2017) Separability-oriented subclass discriminant analysis. *IEEE Trans Pattern Anal Mach Intell* 40(2):409–422
50. Yang X, Jiang X, Tian C, Wang P, Zhou F, Fujita H (2020) Inverse projection group sparse representation for tumor classification: A low rank variation dictionary approach. *Knowl.-Based Syst* 196(21):105768. <https://doi.org/10.1016/j.knsys.2020.105768>
51. Xiao Q, Dai J, Luo J, Fujita H (2019) Multi-view manifold regularized learning-based method for prioritizing candidate disease miRNAs. *Knowl.-Based Syst* 175:118–129. <https://www.sciencedirect.com/science/article/pii/S0950705119301480>
52. Marozzi M (2015) Multivariate multidistance tests for high-dimensional low sample size case-control studies. *Stat Med* 34(9):1511–1526
53. Field C (1982) Small sample asymptotic expansions for multivariate m-estimates. *Ann Stat*, 672–689
54. Blanchard G, Kawanabe M, Sugiyama M, Spokoiny V (2006) Klaus-Robert MÄßler In search of non-gaussian components of a high-dimensional distribution. *J Mach Learn Res* 7(Feb):247–282
55. Zollanvari A, Braga-Neto UM, Dougherty ER (2011) Analytic study of performance of error estimators for linear discriminant analysis. *IEEE Trans Signal Process* 59(9):4238–4255
56. Banerjee O, El Ghaoui L, d'Aspremont A (2008) Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J Mach Learn Res* 9(Mar):485–516
57. Kendler KS, Hetttema JM, Butera F, Gardner CO, Prescott CA (2003) Life event dimensions of loss, humiliation, entrapment, and danger in the prediction of onsets of major depression and generalized anxiety. *Arch Gen Psychiatry* 60(8):789–796
58. Ye J, Janardan R, Park CH, Park H (2004) An optimization criterion for generalized discriminant analysis on undersampled problems. *IEEE Trans Pattern Anal Mach Intell* 26(8):982–994
59. Huang SH, LePendur P, Iyer SV, Tai-Seale M, Carrell D, Shah NH (2014) Toward personalizing treatment for depression: predicting diagnosis and severity. *J Am Med Inform Assoc* 21(6):1069–1075
60. Altman DG, Bland JM (1994) Diagnostic tests. 1: Sensitivity and specificity. *Br Med J* 308(6943):1552



Sijia Yang received MSc in Information, Communications and Technology Business Management from Telecom Ecole de Management, Paris, France 2015 and BEng Degree in Food Science and Engineering from Zhejiang Gongshang University, Zhejiang, China 2011. She is currently working towards her PhD degree in Cyberspace Security in Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include cyber security, data analytics, and machine learning.



Haoyi Xiong received the Ph.D. degree in computer science from Telecom SudParis, University of Paris VI, Paris, France, in 2015. From 2016 to 2018, he was an Assistant Professor with the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA (formerly known as University of Missouri at Rolla). From 2015 to 2016, he was a Post-Doctoral Research Associate with the Department of Systems and Informa-

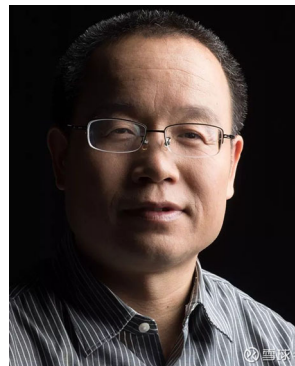
tion Engineering, University of Virginia, Charlottesville, VA, USA. He is currently a Principal R&D Architect and Researcher with Big Data Laboratory, Baidu Research, Beijing, China.

His current research interests include automated deep learning (AutoDL), ubiquitous computing, artificial intelligence, and cloud computing. He has published more than 60 papers in top computer science conferences and journals, such as ICML, ICLR, UbiComp, RTSS, AAAI, IJCAI, ICDM, PerCom, IEEE Transactions on Mobile Computing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Computers, ACM Transactions on Intelligent Systems and Technology, ACM Transactions on Knowledge Discovery from Databases, and etc. He gave keynote speeches in a series of academic and industrial activities, such as the industrial session of the 19th IEEE International Conference on Data Mining (ICDM'19), and served as Poster Co-chair for the 2019 IEEE International Conference on Big Data (IEEE Big Data'19). Dr. Xiong was a recipient of the Best Paper Award from IEEE UIC 2012, the Outstanding Ph.D. Thesis Runner Up Award from CNRS SAMOVAR 2015, and the Best Service Award from IEEE UIC 2017. He was the co-recipient of Science & Technology Advancement Award (First Prize) from the Chinese Institute of Electronics 2019.

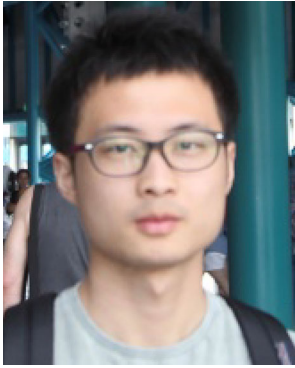


Dr. Kaibo Xu received his Bachelor degree (1998) in Computer Science from Beijing University of Chemical Technology and his Master (2005) and PhD (2010) in Computer Science from the University of the West of Scotland. He worked as a Teaching Assistant (1998-2004), Lecturer (2004-2009), Associate Professor (2009-2017) at Beijing Union University. He has supervised more than 20 master and doctoral students who are successful in their academic and industrial careers.

As the principal investigator, he has received 7 governmental funds and 5 industrial funds with the total amount of 5M in the Chinese dollar. Dr. Kaibo Xu has also consulted extensively and been involved in many industrial projects. He worked as the Chief-Information-Officer (CIO) of Yunbai Clothing Retail Group, China (2016-2019). Currently, he is serving as the vice president and principal scientist of MiningLamp Tech. His research interests include graph mining, knowledge graph and knowledge reasoning.



Licheng Wang received the B.S. degree in engineering from Northwest Normal University, Lanzhou, China, in 1995, the M.S. degree in mathematics from Nanjing University, Nanjing, China, in 2001, and the Ph.D. degree in engineering from Shanghai Jiao Tong University, Shanghai, China, in 2007. He is currently the Full Professor with the Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include cryptography, blockchain, and future Internet architecture.



Jiang Bian received the B.Eng degree of Logistics Systems Engineering in Huazhong University of Science and Technology, Wuhan, China, in 2014, and the M.Sc degree of Industrial Systems Engineering in University of Florida at Gainesville, FL, USA, in 2016. He is currently pursuing the Ph.D. degree in the Department of Computer and Electrical Engineering, University of Central Florida, Orlando, FL, USA, under co-supervision of Dr. Zhishan

Guo and Dr. Haoyi Xiong. From 2016 to 2018, he spent the first two years of his Ph.D study in the Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA. His research interests include Human-Subject Data Learning, Ubiquitous Computing, Intelligent Cyber-Physical Systems.




Dr. Zeyi Sun received the B.Eng. degree in material science and engineering from Tongji University, Shanghai, China, in 2002, the M.Eng. degree in manufacturing from the University of Michigan Ann Arbor, Ann Arbor, MI, USA, in 2010, and the Ph.D. degree in industrial engineering and operations research from the University of Illinois at Chicago, Chicago, IL, USA, in 2015. He served as an Assistant Professor with the Department of Engineer-

ing Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO, USA, from 2015 to 2020. Currently, he is a senior research scientist with MiningLamp Academy of Sciences, MiningLamp Technology, Beijing, China.

His research interest is mainly focused on using reinforcement learning algorithms to solve dynamic decision-making problem formulated by Markov Decision Process.

Affiliations

Sijia Yang¹ · Haoyi Xiong² · Kaibo Xu³ · Licheng Wang¹ · Jiang Bian⁴ · Zeyi Sun³ 

Sijia Yang
ysjhhh@gmail.com

Haoyi Xiong
haoyi.xiong.fr@ieee.org

Kaibo Xu
xukaibo@mininglamp.com

Jiang Bian
bj1119@knights.ucf.edu

¹ School of Cyberspace Security, State Key Laboratory of Networking and Switching, Beijing University of Posts and Telecommunications, Haidian, Beijing, China

² Department of Computer Science, Missouri University of Science and Technology, Rolla, MO 65409, USA

³ Mininglamp Academy of Sciences, Mininglamp Technology, Beijing, 100084, China

⁴ Department of Electrical and Computer Engineering, University of Central Florida, Orland, FL 32816, USA