De-Biasing Covariance-Regularized Discriminant Analysis

Conference Paper - July 2018
DOI: 10.24963/ijcai.2018/401

6 authors, including:

Haoyi Xiong
Baidu Online Network Technology
54 PUBLICATIONS 614 CITATIONS

Wei Cheng
University of North Carolina at Chapel Hill
37 PUBLICATIONS 135 CITATIONS

Wenqing Hu
Missouri University of Science and Technology (formerly University of ...
25 PUBLICATIONS 48 CITATIONS

Jiang Bian
Missouri University of Science and Technology
7 PUBLICATIONS 0 CITATIONS

Some of the authors of this publication are also working on these related projects:

Discriminant Learning for Human Behavior Understanding View project

Mobile Crowd Sensing: Algorithms and Systems View project

All content following this page was uploaded by Haoyi Xiong on 24 June 2018.
The user has requested enhancement of the downloaded file.
De-Biasing Covariance-Regularized Discriminant Analysis

Haoyi Xiong†,‡, Wei Cheng‡, Yanjie Fu†,‡, Wenhong Hu†, Jiang Bian†,‡, Zhishan Guo†
†Baidu Inc., Beijing, China
‡National Engineering Laboratory of Deep Learning Technology and Application, Beijing, China
§Missouri University of Science and Technology, MO, United States
§NEC Laboratories America, NJ, United States

Abstract

Fisher’s Linear Discriminant Analysis (FLD) is a well-known technique for linear classification, feature extraction and dimension reduction. The empirical FLD relies on two key estimations from the data – the mean vector for each class and the (inverse) covariance matrix. To improve the accuracy of FLD under the High Dimension Low Sample Size (HDLSS) settings, Covariance-Regularized FLD (CRLD) has been proposed to use shrunken covariance estimators, such as Graphical Lasso, to strike a balance between biases and variances. Though CRLD could obtain better classification accuracy, it usually incurs bias and converges to the optimal result with a slower asymptotic rate. Inspired by the recent progress in de-biased Lasso, we propose a novel FLD classifier, DBLD, which improves classification accuracy of CRLD through de-biasing. Theoretical analysis shows that DBLD possesses better asymptotic properties than CRLD. We conduct experiments on both synthetic datasets and real application datasets to confirm the correctness of our theoretical analysis and demonstrate the superiority of DBLD over classical FLD, CRLD and other downstream competitors under HDLSS settings.

1 Introduction

Fisher’s Linear Discriminant Analysis (FLD) [Duda et al., 2001] is a well-known technique for feature extraction and dimension reduction [Kulis and others, 2013]. It has been widely used in many applications, such as face recognition [Peck and Van Ness, 1982], image retrieval, etc. An intrinsic limitation of classical FLD is that its objective function relies on the well-estimated and non-singular covariance matrices. For many applications, such as the micro-array data analysis, all scatter matrices can be singular or ill-posed since the data is often with high dimension but low sample size (HDLSS) [Cai et al., 2016].

The classical FLD classifier relies on two key parameters – the mean vector of each type and the precision matrix. Under the HDLSS settings, the sample precision matrix (a.k.a., the inverse of sample covariance matrix) used in FLD is usually ill-estimated and quite different from the inverse of population/true covariance matrix [Cai et al., 2016]. 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 [Marčenko and Pastur, 1967]. Such inconsistency between the true and the estimated precision matrices degrades the accuracy of FLD classifiers under the HDLSS settings [Zolotov and Dougherty, 2013].

A plethora of excellent work has been conducted to address such HDLSS data classification problem for FLD. For example, Krzanowski et al. [Krzanowski et al., 1995] suggested to use pseudo-inverse to approximate the inverse covariance matrix, when the sample covariance matrix is singular. However, the precision of pseudo-inverse FLD is usually low and not well guaranteed. Other techniques include the two-stage algorithm PCA+FLD [Ye et al., 2004], FLD based on Kernels [Zhang and others, 2003] and/or other non-parametric statistics [Kaski and Peltonen, 2003]. To overcome the singularity of the sample covariance matrices, instead of estimating inverse covariance matrix and mean vectors separately, [Cai and Liu, 2011] proposed to estimate the projection vector for discrimination directly. More popularly, regularized FLD approaches [Krzanowski et al., 1995; Witten and Tibshirani, 2009] are proposed to solve the problem. These methods can improve the performance of FLD either empirically or theoretically [Durrant and Kabán, 2015; Bickel et al., 2004], while few of them can directly address the ill-estimated inverse covariance matrix estimation issue.

One representative regularization approach is Covariance-Regularized FLD [Witten and Tibshirani, 2009] that replaces the precision matrix used in FLD with a shrunken estimator, such as Graphical Lasso [Friedman et al., 2008], so as to achieve a “superior prediction”. Intuitively, through replacing precision matrix used in FLD with a sparse regularized estimation, the ill-posed problem caused by the HDLSS settings can be well addressed. The sparse estimators usually converge to the inverse of true/population covariance matrix faster than the sample estimators [Cai et al., 2016]. With the asymptotic properties, the sparse FLD should be close to the optimal FLD. 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 literature.

Further, with induced sparsity, the inverse covariance estimator becomes biased [Zhang and Zhang, 2014]. The performance of sparse FLD is frequently bottlenecked due to the bias of the sparse estimators. Recently, researchers tried to de-bias the Lasso estimator [Zhang and Zhang, 2014], through adjusting the $\ell_1$-penalty for the regularized estimation, so as to achieve a better regression performance. Inspired by this line of research, we propose to improve sparse FLD through de-biasing (i.e., de-sparsifying) in this paper.

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

1. Inspired by De-biased Lasso [Javanmard and Montanari, 2014], we study the problem of de-biasing the Covariance-Regularized FLD (CRLD), which has been widely-used for empirical sparse FLD estimation, for performance improvement. To the best of our knowledge, this is the first work aiming at de-biasing CRLD.

2. We propose a novel algorithm DBLD — De-Biased Fisher’s Linear Discriminant Analysis on top of CRLD. DBLD leverages yet-another De-Biased Estimator for linear classification problem, to re-balance the variances and biases of the estimation, through de-sparsifying the projection vector obtained by CRLD.

3. Our theoretical analysis shows, under certain mild assumptions, DBLD converges faster than CRLD with sharp asymptotic rate. We also conduct extensive experiments to demonstrate the advantage of the proposed algorithms over other competitors. The results validate the correctness of our theoretical analysis.

**Notations.** Following key notations are used in the rest of this paper: Given a $p$-dimensional vector $v \in \mathbb{R}^p$, we denote the $\ell_p$ vector-norm as $|v|_p = \left(\sum_{i=1}^m|v_i|^p\right)^{1/p}$ ($P$ is a non-negative integer) and the $\ell_\infty$ vector-norm as $|v|_\infty = \max_{1 \leq i \leq m} \{|v_i|\}$. Given a matrix $A \in \mathbb{R}^{m \times m}$, we denote the $\ell_P$ matrix-norm as $|A|_P = \max_{v \in \mathbb{R}^p} \{|Av|_p/|v|_p\}$. Note that the symbol $p$ refers to the number of dimensions of the data while $m$ refers to the number of samples. The operator $O_p(\cdot)$ refers to the big-O-notation in high probability.

2 **Preliminaries**

In this section, we first briefly introduce the binary classifier using FLD, then present the practice of CRLD based on Graphical Lasso.

2.1 **FLD for Binary Classification**

To use the Fisher’s Linear Discriminant Analysis (FLD), given the $i.i.d.$ labeled data pairs $(x_1, \ell_1) \ldots (x_m, \ell_m)$, we first estimate the sample covariance matrix $\Sigma$ using the pooled sample covariance matrix estimator with respect to the two classes [Duda et al., 2001]. Then estimate the sample precision matrix as $\bar{\Theta} = \Sigma^{-1}$. Further, $\mu_+$ and $\mu_-$ are estimated as the mean vectors of the positive samples and the negative samples in the $m$ training samples, respectively.

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

$$\tilde{f}(x) = \arg\max_{\ell \in \{-, +\}} \delta(x, \bar{\Theta}, \mu_\ell, \pi_\ell), \text{ where}$$

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

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

2.2 **Covariance-Regularized FLD via Graphical Lasso**

This algorithm, referred to as the Covariance-Regularized FLD (CRLD) via Graphical Lasso, was derived from the Scout family of FLD introduced by Witten et al. in [Witten and Tibshirani, 2009]. Compared to the classical FLD, this baseline algorithm leverages Graphical Lasso estimator to replace the precision matrix estimated using sample covariance matrix. The proposed algorithm is implemented using the discriminant function defined in Eq. 1, as:

$$\hat{f}(x) = \arg\max_{\ell \in \{-, +\}} \delta(x, \hat{\Theta}, \hat{\mu}_\ell, \pi_\ell),$$

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

$$\hat{\Theta} = \arg\min_{\Theta > 0} \left(\text{tr}(\Sigma \Theta) - \log \det(\Theta) + \lambda \sum_{j \neq k} |\Theta_{jk}|\right).$$

Note that, as a linear classifier, the CRLD decision rule introduced in Eq. 2 can be re-formulated in a linear model, such as:

$$\hat{f}(x) = \text{sign} \left(\delta(x, \hat{\Theta}, \hat{\mu}_+, \pi_+) - \delta(x, \hat{\Theta}, \hat{\mu}_-, \pi_-)\right)$$

$$= \text{sign} \left(x^T \hat{\beta}^G + c^G\right),$$

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

In this paper, we present the analytical results (i.e., statistical rate of convergence that $\hat{\beta}^G$ approximates to the optimal projection vector with varying number of samples $n$ and dimensions $p$) of CRLD in Theorem 1.

3 **The Proposed Algorithm**

In this section, we introduce our proposed algorithm DBLD — De-Biased Fisher’s Linear Discriminant Analysis (via Graphical Lasso), then present the theoretical analysis on the theoretical properties of the proposed algorithms.

3.1 **DBLD: The De-Biased Estimation for Covariance-Regularized FLD**

Given the $i.i.d.$ labeled data pairs $(x_1, \ell_1) \ldots (x_m, \ell_m)$ drawn from the two classes with certain priors, as shown in Algorithm 1. The algorithm first (i) estimates the sample estimation
of covariance matrices and the mean vectors, then (ii) lever-
age CRLD to estimate the shrunked projection vector $\tilde{\beta}^G$. Further, DBLD (iii) proposes a de-biased estimator (denoted as DeBiasing function) to de-bias $\beta^G$ and obtain the projection vector $\tilde{\beta}^D$. Finally, we introduce a decision rule that enables classification using the estimated $\tilde{\beta}^D$.

Algorithm 1 DBLD Estimation Algorithm

1: procedure DBLD($(x_1, \ell_1) \ldots (x_m, \ell_m)$)
2: /*(i) Sample Estimators for Mean and Covariance */
3: $X_+ \leftarrow$ PositiveSampleSet($(x_1, \ell_1) \ldots (x_m, \ell_m)$);
4: $X_- \leftarrow$ NegativeSampleSet($(x_1, \ell_1) \ldots (x_m, \ell_m)$);
5: $\bar{\mu}_+ \leftarrow \frac{1}{|X_+|} \cdot \sum_{x \in X_+} x$, $\bar{\mu}_- \leftarrow \frac{1}{|X_-|} \cdot \sum_{x \in X_-} x$;
6: $\bar{\Sigma}_+ \leftarrow \frac{1}{|X_+|} \cdot \sum_{x \in X_+} (x - \bar{\mu}_+)(x - \bar{\mu}_+)^T$;
7: $\bar{\Sigma}_- \leftarrow \frac{1}{|X_-|} \cdot \sum_{x \in X_-} (x - \bar{\mu}_-)(x - \bar{\mu}_-)^T$;
8: $\bar{\mu} \leftarrow \frac{|X_+| \cdot \bar{\mu}_+ + |X_-| \cdot \bar{\mu}_-}{|X_+| + |X_-|}$, $\bar{\Sigma} \leftarrow \frac{|X_+| \cdot \bar{\Sigma}_+ + |X_-| \cdot \bar{\Sigma}_-}{|X_+| + |X_-|}$;
9: /*(ii) CRLD Estimator (to obtain $\tilde{\beta}^G$) */
10: $\bar{\Theta} \leftarrow$ GraphicalLasso($\bar{\Sigma}$, $\lambda$);
11: $\beta^G \leftarrow \bar{\Theta}(\bar{\mu}_+ - \bar{\mu}_-) ;$
12: /*(iii) DBLD Estimator (to obtain $\tilde{\beta}^D$) */
13: $X \leftarrow [x_1, x_2, \ldots, x_m]$; $p \times m$ matrix /*
14: $L \leftarrow [\ell_1, \ell_2, \ldots, \ell_m]$; $\ell \times 1$ matrix /*
15: $U \leftarrow [\bar{\mu}, \bar{\mu}, \ldots, \bar{\mu}]$;
16: $\Theta U$ is an $m \times p$ matrix, every column is $\bar{\mu}^n$/*
17: $c \leftarrow -\bar{\mu}^T \beta^G$;
18: $C \leftarrow [c, c, \ldots, c]^T$;
19: /*$C$ is a $m \times 1$ matrix, every row is $c$*/
20: $\tilde{\beta}^D \leftarrow \tilde{\beta}^G + \frac{1}{m} \bar{\Theta}(X - U)(2 \cdot L - X^T \beta^G - C)$;
21: return $\tilde{\beta}^D$;

In the following section, we present the design of the De-
Biased Estimator (denoted as DeBiasing function in Algo-
rithm 1) to obtain $\tilde{\beta}^D$, then introduce the decision rule for
optimal classification. Later we analyze the theoretical prop-
erties of $\tilde{\beta}^D$.

The De-Biased Estimator

Inspired by the De-biased Lasso [Javanmard and Monta-
nari, 2014], we propose to improve the performance of
CRLD through de-biasing $\beta^G$. 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 $\bar{\Theta}$ on the data and the CRLD model
(i.e., $\beta^G$), we propose a novel de-biased estimator of $\beta^D$ that
takes the form as

$$\tilde{\beta}^D \leftarrow \tilde{\beta}^G + \frac{1}{m} \bar{\Theta}(X - U)(2 \cdot L - X^T \beta^D - C),$$

where we denote $X$ as an $p \times m$ matrix where $1 \leq i \leq m$
and the $i$th column is $x_i$; $L$ as an $m \times 1$ matrix (i.e., vector)
whose $i$th row is $\ell_i \in \{\pm 1\}$; $U$ is a $p \times m$ matrix where
each column is $\bar{\mu}$ (as line 7 in Algorithm 1); and $C$ is an $m \times 1$
matrix where each row is $c$ (as line 16 in Algorithm 1).

The DBLD Classifier.

Given the de-biased estimator $\tilde{\beta}^D$, the DBLD classifies the
input vector $x$ using the following rule:

$$\tilde{f}^D(x) = \text{sign} \left( x^T - \frac{\mu_+ + \mu_-}{2} \right) \tilde{\beta}^D + \log(\pi_+ / \pi_-).$$

In the following section, we present the analytical results
of DBLD, including the computational complexity of de-
biasing and statistical rate of convergence.

3.2 Complexity Analysis of DBLD

In this section, we analyze the computational complexity for
the three steps of Algorithm 1. The step (i) estimates the sample
covariance matrices and mean vectors, which consumes at most $O(p^2 \cdot m)$ operations. The step (ii) performs Graphical
Lasso and matrix multiplication, where the complexity based
on standard implementation [Friedman et al., 2008] is upper-
bounded by $O(p^3)$. The step (iii) de-biasing is implemented as
an exact formula with $O(p^2)$ complexity.

Remark 1. All three steps of Algorithm 1 are scalable on
both the number of dimensions ($p$) and the number of train-
samples ($m$). The overall complexity of the three steps is
$O(p^3 + p^2 \cdot m)$. Under the HDLSS setting $p > m$, the computational complexity of DBLD is upper-bounded by
$O(p^2 \cdot m)$. On the other hand, with large sample setting where
$m \geq p$, the worst case computational complexity of DBLD
is bounded by $O(p^2 \cdot m)$. Obviously, the proposed de-biasing
estimator (i.e., step (iii)) with complexity $O(p^2)$ would not
bound the performance, when compared to the first two steps.

3.3 Convergence Analysis of DBLD

In order to analyze the performance of DBLD, we first define
the linear projection vector of the optimal FLD as $\beta^*$. Given $m$
samples $(x_1, \ell_1), (x_2, \ell_2), \ldots, (x_m, \ell_m)$ drawn i.i.d. from $\mathcal{N}(\mu^*_+, \Sigma^*)$ and $\mathcal{N}(\mu^*_-, \Sigma^*)$ with the equal priors for training, the optimal
projection vector should be $\beta^* = \Theta^*(\mu^*_+ - \mu^*_-)$. Thus
we intend to understand how close $\beta^G$ and $\beta^D$ approximate to
the optimal estimation $\beta^*$.

Assumption 1. We follow the assumptions made in [Rothman
et al., 2010] that a positive constant $K$ having

$$1/K \leq \lambda_{\min}(\Sigma^*) \leq \lambda_{\max}(\Sigma^*) \leq K$$

exists. The operators $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest
and largest eigenvalues respectively. In this way, there exists
$\|\Sigma^*\|_2 \leq K$ and $\|\Theta^*\|_2 \leq K$.

Assumption 2. We further follow the assumption that, the
data vectors for training are all realized from a random vector
$X$ and there exists an constant $B$ having $|X|_2 \leq B$. Thus
there has $|\mu^*_+|_2 \leq B$ and $|\mu^*_-|_2 \leq B$.

Theorem 1. With appropriate setting of tuning parameter
$\lambda \propto \sqrt{\log p/m}$ (in Eq 3), the $\ell_2$-vector-norm convergence
rate of CRLD $\beta^G$ approximating to the optimal estimation $\beta^*$ is:

$$|\beta^G - \beta^*|_2 = O_p \left( \sqrt{\frac{(p + d) \log p}{m}} \right),$$

where $d$ is some positive constant.
where \( d = \max_{1 \leq i \leq p} \{ j : \Sigma_{\beta_i}^{-1} \neq 0 \} \) refers to the maximal degree of the graph (i.e., population inverse covariance matrix).

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

\[
| \hat{\beta}^G - \beta^* |_2 = | \hat{\Theta} (\hat{\mu}_+ - \hat{\mu}_-) - \Theta^* (\mu^*_+ - \mu^*_-) |_2.
\]  

(8)

Considering the inequities \(| x + y | \leq |x| + |y| \) and \(| Ax | \leq |A| |x| \), we have

\[
| \hat{\beta}^G - \beta^* |_2 \leq \| (\hat{\Theta} - \Theta^* ) \|_2 : \| \hat{\mu}_+ - \mu_+ \|_2 + \| \Theta^* \|_2 : (\| \mu_+ - \mu_+^* \|_2 + \| \mu_- - \mu_-^* \|_2).
\]  

(9)

According to [Rothman et al., 2010], when \( \lambda \propto \sqrt{\log p/m} \), we consider the spectral-norm convergence rate \( \| \hat{\Theta} - \Theta^* \|_2 \leq \| \hat{\Theta} - \Theta^* \|_p \). \( \Theta^* \) is the asymptotic rate of sample mean vector [DasGupta, 2008] is \( \| \mu_+ - \mu_+^* \|_2 = \| \mu_- - \mu_-^* \|_2 = \mathcal{O}_p(\sqrt{p/m}) \), with the increasing number of dimensions \( p \) and number of samples \( m \).

Further, there has \( \| \Theta^* \|_2 \leq K \) (Assumption 1) and \( \ell_2 \)-norms of all mean vectors are bounded by \( B \). In this way, there must exist positive constants \( C_1 \) and \( C_2 \) having:

\[
| \hat{\beta}^G - \beta^* |_2 \leq C_1 \cdot 2B \sqrt{\frac{(p + d) \log p}{m}} + C_2K \sqrt{\frac{p}{m}}.
\]  

(10)

Thus, according to the definition of asymptotic rate, we conclude the convergence rate as:

\[
| \hat{\beta}^G - \beta^* |_2 = \mathcal{O}_p \left( \sqrt{\frac{(p + d) \log p}{m}} \right).
\]  

(11)

**Theorem 2.** With appropriate setting of tuning parameter \( \lambda \) (in Eq 3), the \( \ell_2 \)-vector-norm convergence rate of DBLD \( \hat{\beta}^G \) approximating to the optimal estimation \( \beta^* \) is:

\[
| \hat{\beta}^D - \beta^* |_2 = \mathcal{O}_p \left( \sqrt{\frac{\log p}{m}} \right).
\]  

(12)

**Proof.** Here, we prove the upper bound of \(| \hat{\beta}^D - \beta^* |_\infty \). Consider the definition of the de-biased FLD estimator \( \hat{\beta}^D \) introduced in Eq. 5, we have

\[
\hat{\beta}^D = \hat{\beta}^G + \frac{2}{m} \cdot \hat{\Theta} \mathbf{XL} - \frac{2}{m} \cdot \hat{\Theta} \mathbf{UL} - \frac{1}{m} \cdot \hat{\Theta} (\mathbf{X} - \mathbf{U}) (\mathbf{X} - \mathbf{U})^\top \hat{\beta}^G.
\]  

(13)

With the assumption of equal priors \((\pi_+ = \pi_- = 0.5) \), \( \mathbf{L} \) is a \( m \times 1 \) label matrix that half of its elements are +1 while the rest are all -1. \( \mathbf{X} \) refers to a \( p \times m \) matrix, where each column is a sample of data e.g., \( x_1, x_2, \ldots, x_m \). As was defined \( \hat{\beta}^G = \hat{\Theta} (\hat{\mu}_+ - \hat{\mu}_-) = \frac{2}{m} \cdot \hat{\Theta} \mathbf{XL} \). As \( \mathbf{U} \) is a matrix in which each column is a constant vector \((\mu_+ + \mu_-)/2 \) and \( \mathbf{L} \) is a vector with half elements as 1 and half elements as -1, thus \( \frac{2}{m} \cdot \hat{\Theta} \mathbf{UL} = \frac{1}{m} \cdot \hat{\Theta} (\mathbf{UL}) = 0 \). As each column of \( \mathbf{X} \) refers to a sample drawn from the original data distribution, thus \( \frac{1}{m} (\mathbf{X} - \mathbf{U}) (\mathbf{X} - \mathbf{U})^\top = \hat{\Sigma}_s \) is the sample covariance matrix estimator. With all above in mind, we have

\[
\hat{\beta}^D = \hat{\beta}^G + (\mathbf{I} - \hat{\Theta} \hat{\Sigma}_s) \hat{\beta}^G,
\]  

(14)

where \( \mathbf{I} \) refers to a \( p \times p \) identity matrix. Note that \((\mathbf{I} - \hat{\Theta} \hat{\Sigma}_s) \hat{\beta}^G \) can be considered as the de-sparsification term that de-biases \( \hat{\beta}^G \). Thus, considering the asymptotic rate of sample mean vector [DasGupta, 2008] is \( \| \hat{\mu}_+ - \mu_+^* \|_2 = \mathcal{O}_p(\sqrt{p/m}) \) and \( \| \hat{\mu}_- - \mu_-^* \|_2 = \mathcal{O}_p(\sqrt{p/m}) \), with the increasing number of dimensions \( p \) and number of samples \( m \).

According to [Jankova et al., 2015], with appropriate setting of \( \lambda \), the spectral-norm convergence rate of the de-sparsified estimator \( \hat{\Theta}^D = \left( 2 \cdot \hat{\Theta} \cdot \hat{\Theta} \hat{\Sigma}_s \hat{\Theta} \right) \) under mild conditions should be \( \| \hat{\Theta}^D - \Theta^* \|_\infty = \mathcal{O}_p(\sqrt{\log p/m}) \), then there exists

\[
| \hat{\Theta}^D - \Theta^* |_2 = \mathcal{O}_p \left( \sqrt{\frac{\log p}{m}} \right), \text{ with the varying number of dimensions } p \text{ and number of samples } m.
\]  

In this way, with high probability, we conclude the convergence rate:

\[
| \hat{\beta}^D - \beta^* |_2 = \mathcal{O}_p \left( \sqrt{\frac{\log p}{m}} \right).
\]  

(15)

**Remark 2.** Compared to CRLD’s projection vector \( \hat{\beta}^G \), our method DBLD recovers the linear projection vector \( \hat{\beta}^D \) with a faster asymptotic rate, i.e., \( \sqrt{p \log p/m} \) versus \( \sqrt{(p + d) \log p/m} \) in a mild condition. Thus, it would benefit to some applications, such as dimensionality reduction and feature selection. Our later experimental results show that DBLD outperforms CRLD with higher classification accuracy, due to the faster statistical rate of convergence.

**Remark 3.** The proposed algorithm provides a sub-optimal solution, when compared to [Cai and Liu, 2011]. Our work intends to propose an estimator of \( \beta^* \) through approximating \( \Sigma^* \), \( \mu^*_+ \) and \( \mu^*_- \) separately, while [Cai and Liu, 2011] approximates \( \beta^* \) straightforwardly via so-called “direct estimation”.

4 Experiments

In this section, we first validate different properties of DBLD on the synthesized data. Then, we experimentally evaluate the performance of DBLD using several real-world datasets. Experiments show the superiority of DBLD.

4.1 Synthesized Data Evaluation

To validate our algorithms, we evaluate our algorithms on a synthesized dataset (imported from [Cai and Liu, 2011]), which is obtained through a pseudo-random 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 $\mu^*_+, \mu^*_-$ and $\Sigma^*$ are as follows: $\Sigma^*$ is a $p \times p$ symmetric and positive-definite matrix, where each element $\Sigma^*_{i,j} = 0.8^{i-j}1$, $1 \leq i \leq p$ and $1 \leq j \leq p$. $\mu^*_+$ and $\mu^*_-$ are both $p$-dimensional vectors, where $\mu^*_+ = (1, 1, \ldots, 1, 0, 0, \ldots, 0)^T$ (the first 10 elements are all 1’s, while the rest $p - 10$ elements are 0’s) and $\mu^*_- = 0$. In our experiment, we set $p = 200$. To simulate the HDLSS settings, we train CRLD and DBLD, with 20 to 200 samples randomly drawn from the distributions with equal priors, and test the two algorithms using 500 randomly generated samples. For each setting, we repeat the experiments for 100 times and report the averaged results, in a cross-validation manner.

In this experiment, we compare DBLD, CRLD and FLD (with pseudo inverse). The results of FLD is not included here, as it performs extremely worse than both CRLD and DBLD under the HDLSS settings. Figure. 1(a) presents the comparison between DBLD and CRLD, in terms of accuracy, where each algorithm is fine-tuned with the best parameter $\lambda$. A detailed example of parameter tuning is reported in Figure. 1(b), where we run both algorithms, with training set size as 160, when varying $\lambda$ from 1 to 70. From Figure. 1(a), it is obvious that DBLD outperforms CRLD marginally. The $\lambda$ tuning comparison addressed in Figure. 1(b) shows that, given a small $\lambda$, both CRLD and DBLD cannot perform well, as the sparse approximation of $\hat{\beta}^G$ and $\hat{\beta}^D$ cannot be well recovered in such case [Witten and Tibshirani, 2009]. When $\lambda \geq 6$, DBLD starts outperforming CRLD, while the advantage of DBLD to CRLD decreases when increasing $\lambda$. However, even with an extremely large $\lambda$, DBLD still outperforms CRLD. In Figure 2(a), 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. The proportion of positive training samples is varying from 10% to 40%. It is obvious that all algorithms achieve their best performance when the proportion of positive training sample is 10% (the most unbalanced case).

To further verify our algorithms, we propose the optimal FLD classifier $\hat{\beta}^* = \Theta(\hat{\mu}^*_+ - \hat{\mu}^-)$, which is all based on the population parameters. We compare the $\hat{\beta}^D$, $\hat{\beta}^G$ and $\hat{\beta}$ estimated by DBLD, CRLD and FLD (with pseudo-inverse) to $\beta^*$. Figure. 2(b) presents the comparison among $|\hat{\beta}^D - \beta^*|_\infty$, $|\hat{\beta}^G - \beta^*|_\infty$ and $|\hat{\beta} - \beta^*|_\infty$. It is obvious that $\hat{\beta}^D$ is more close to $\beta^*$ than $\hat{\beta}^G$ and $\hat{\beta}$. This observation further verifies the Theorem 1 and 2. We also compare the accuracy of $\beta^*$ to CRLD, DBLD and FLD. $\beta^*$ outperforms these algorithms and the accuracy of $\beta^*$ is around 84.4% It is reasonable to conclude that DBLD outperforms CRLD, because $\hat{\beta}^D$ is more close to $\beta^*$.

### 4.2 Benchmark Evaluation Results

In Figure. 3(a), we compare DBLD and other FLD algorithms, including FLD with pseudo-inverse, Sparse FLD via Graphical Lasso (CRLD) and Ye-FLD derived from [Ye et al., 2004], on the Web datasets [Lin, 2017]. 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, reported result is averaged over 100 randomly selected subsets of the training/testing data with equal priors. CRLD and DBLD are fine-tuned with the best $\lambda$. The experimental settings show that DBLD consistently outperforms other competitors in different settings. The non-monotonic trend of FLD with the increasing training set size is partially due to the poor performance of pseudo inverse used in FLD.

In addition to FLD classifiers, we also compared DBLD 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 Figure. 3(b). All algorithms are fine-tuned with the best parameters under our experiment settings.

### 4.3 Early Detection of Diseases on EHR Datasets

To demonstrate the effectiveness of DBLD in handling the real problems, we evaluate DBLD on the real-world Electronic Health Records (EHR) data for early detection of di-
Figure 2: More Performance Comparison based on Pseudo-Random Synthesized Data

Figure 3: Performance Comparison on Benchmark Datasets ($p = 300$ and $p \gg m$, D-Tree: Decision Tree, R-Forest: Random Forest, K-SVM: Kernel SVM, and L-SVM: Linear SVM)

Table 1: Early Detection of Diseases Accuracy Comparison between DBLD and Baselines.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLD</td>
<td>0.659±0.022</td>
<td>0.677±0.028</td>
<td>0.691±0.024</td>
<td>0.692±0.023</td>
<td>0.690±0.021</td>
<td>0.696±0.024</td>
<td>0.701±0.023</td>
</tr>
<tr>
<td>FLD</td>
<td>0.543±0.034</td>
<td>0.586±0.033</td>
<td>0.616±0.022</td>
<td>0.642±0.029</td>
<td>0.642±0.022</td>
<td>0.657±0.025</td>
<td>0.658±0.026</td>
</tr>
<tr>
<td>Ye-FLD</td>
<td>0.627±0.050</td>
<td>0.620±0.077</td>
<td>0.652±0.063</td>
<td>0.620±0.067</td>
<td>0.655±0.062</td>
<td>0.637±0.064</td>
<td>0.670±0.045</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.621±0.046</td>
<td>0.649±0.031</td>
<td>0.652±0.041</td>
<td>0.655±0.030</td>
<td>0.671±0.028</td>
<td>0.665±0.031</td>
<td>0.668±0.040</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>0.615±0.026</td>
<td>0.628±0.030</td>
<td>0.647±0.023</td>
<td>0.666±0.029</td>
<td>0.666±0.021</td>
<td>0.670±0.030</td>
<td>0.675±0.029</td>
</tr>
<tr>
<td>Kernel SVM</td>
<td>0.635±0.032</td>
<td>0.669±0.027</td>
<td>0.674±0.039</td>
<td>0.678±0.021</td>
<td>0.668±0.038</td>
<td>0.688±0.024</td>
<td>0.682±0.029</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.631±0.035</td>
<td>0.630±0.039</td>
<td>0.620±0.028</td>
<td>0.622±0.027</td>
<td>0.621±0.022</td>
<td>0.617±0.025</td>
<td>0.626±0.070</td>
</tr>
<tr>
<td>CRLD</td>
<td>0.658±0.023</td>
<td>0.676±0.024</td>
<td>0.682±0.028</td>
<td>0.686±0.022</td>
<td>0.683±0.021</td>
<td>0.692±0.025</td>
<td>0.695±0.018</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.590±0.035</td>
<td>0.602±0.035</td>
<td>0.653±0.031</td>
<td>0.602±0.040</td>
<td>0.674±0.024</td>
<td>0.666±0.026</td>
<td>0.658±0.032</td>
</tr>
</tbody>
</table>

In this application, each patient’s EHR data is represented by a $p = 295$ dimensional vector, referring to the outpatient record on the physical disorders diagnosed. Patients are labeled with either “positive” or “neg-
**4.4 Leukemia and Colon Cancer Datasets**

We evaluate DBLD, CRLD and other baseline algorithms, including Decision Tree, Random Forest and SVM, using leukemia and colon cancer datasets (derived from [Lin, 2017; Tibshirani et al., 2002]) under HDLSS settings (i.e., $p = 7, 128$ and $2,000$ vs. $m = 20$).

Table ?? presents the description of two datasets [Lin, 2017; Tibshirani et al., 2002] that we used to evaluate the proposed and baseline algorithms. “Leukemia” refers to the leukemia cancer dataset [Tibshirani et al., 2002] that includes 7,128 features and totally 72 samples (for training and testing). In this datasets, 47 samples are labeled as “ALL” class while 25 samples are identified as “AML”. On the other hand, “Colon” refers to the colon cancer datasets [Lin, 2017] that are with 2,000 features and totally 62 samples, where 40 samples are negative and 22 samples are identified as positive. Both datasets are with a ultra-large number of dimensions but with extremely low sample sizes (i.e., $p \gg m$).

To accurately estimate the performance of algorithms using these datasets under HDLSS settings, we use cross-validation to limit the potential over-fitting. In each round of cross-validation, we first randomly drawn 20 samples with equal prior from the datasets as the training set, and randomly drawn 20 samples with equal prior from the disjoint set of training set as the testing set. For each round of cross validation, there are no common samples shared by the two sets. We use the training set to train each classifier (i.e., $p = 7, 128$ or $2,000$ and $m = 20$), so as to simulate the extremely HDLSS settings, then test the trained classifiers using the testing set. For each experiment, we repeat the cross-validation for 100 rounds. All algorithms (including baselines and DBLD) are tuned to have the best accuracy. The experiment results are shown in Table ?? . All results show that DBLD significantly improves CRLD, and it outperforms all baseline algorithms with the highest accuracy and F1-score. Please note that though we trained classifiers using less training data, baselines in our experiments perform comparably with the test errors reported in [Tibshirani et al., 2002].

**4.5 Summary of Experiment Results**

We evaluate DBLD with a limited number of samples for training i.e., $p > m$ or $m \gg p$, to understand its performance under HDLSS setting. For large sample scenario, i.e., when $m \gg p$, the sample-based estimators may provide a robust estimation of LDA. In this case, singularity issues might not exist, then regularization and further the de-biasing procedures are not mandatory.
et al. Witten and Tibshirani, 2009] proposed to first estimate the DBLD performance of covariance-regularized FLD (CRLD) through re-balancing the biases and variances of the projection vector obtained by CRLD. Our analysis shows that DBLD is with better asymptotic properties, compared to CRLD, and can obtain higher classification accuracy, under HDLSS settings. The experimental results on synthesized and real-world datasets show that DBLD outperformed all baseline algorithms. Further, the empirical studies on estimator comparison validate our theoretical analysis.

5 Related Work and Discussion

In this section, we review several most relevant studies of our research. To address the HDLSS issues for FLD, a line of research [Shao et al., 2011; Cai and Liu, 2011] 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 Van Ness, 1982; Bickel and Levina, 2008; Witten and Tibshirani, 2009] proposed to first estimate the inverse covariance matrix through shrinkage covariance estimators, 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 [Zoljanvari and Dougherty, 2013]. Moreover, the performance of FLD has been previously studied in [Durrant and Kabán, 2015; Bickel et al., 2004].

In our paper, we focus on improving covariance-regularized FLD [Witten and Tibshirani, 2009], through de-biasing the projection vector estimated with Graphical Lasso [Witten et al., 2011]. Our work is distinct due to the following reasons: (1) Our work is the first to study the problem of de-biasing the sparse FLD [Zhang and Zhang, 2014]; (2) Compared to the existing solution to the de-biased linear regression models [Javanmard and Montanari, 2014], we proposed a novel de-biased estimator (using a different formulation in Eq 5) for the covariance-regularized sparse Linear Discriminant Analysis [Witten and Tibshirani, 2009; Witten et al., 2011]; (3) We analyzed the de-biased estimator and obtained its asymptotic properties; (4) We validate our algorithms through comparing a wide range of baselines on both synthesized and real-world datasets, where the evaluation result endorses our theory (e.g., asymptotic properties proved in Theorem 1 and 2 vs. the curve shown in Fig 2(b)).

Discussion and Future Work. In this research, we compare DBLD with CRLD, and common FLD (sample FLD, pseudo-inverse FLD, Ye-FLD). We do not make further comparison with other sparse FLD [Cai and Liu, 2011], as we focus on the covariance-regularization. In future work, we plan to study the de-biased estimators for these sparse FLD.

6 Conclusion

In this paper, we studied the problem of improving the performance of covariance-regularized FLD (CRLD) through re-balancing the biases and variances of the projection vector estimation. Inspired by the de-biased estimator of Lasso [Javanmard and Montanari, 2014], we proposed DBLD—a novel De-Biased estimator for CRLD that lowers the estimation error with faster asymptotic rate, through de-biasing the projection vector obtained by CRLD. Our analysis shows that DBLD is with better asymptotic properties, compared to CRLD, and can obtain higher classification accuracy, under HDLSS settings. The experimental results on synthesized and real-world datasets show that DBLD outperformed all baseline algorithms. Further, the empirical studies on estimator comparison validate our theoretical analysis.

References


[Krzanowski et al., 1995] WJ Krzanowski, Philip Jonathan, WV McCarthy, and MR Thomas. Discriminant analysis


