

Statistica Sinica Preprint No: SS-2020-0374	
Title	Communication-Efficient Distributed Linear Discriminant Analysis for Binary Classification
Manuscript ID	SS-2020-0374
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0374
Complete List of Authors	Mengyu Li and Junlong Zhao
Corresponding Author	Junlong Zhao
E-mail	zhaojunlong928@126.com
Notice: Accepted version subject to English editing.	

Communication-Efficient Distributed Linear Discriminant Analysis for Binary Classification

Mengyu Li^{1,2} and Junlong Zhao^{2*}

Renmin University of China, Beijing, China¹

Beijing Normal University, Beijing, China²

Abstract: Large-scale data are commonly encountered nowadays where the sample size n is large, and the data are often stored on k different local machines. Distributed statistical learning is an efficient way to deal with this type of data. In this paper, we consider the binary classification problem for massive data based on the linear discriminant analysis (LDA) in the frame of distributed learning. The classical centralized LDA requires the transmission of some p by p summary matrices to the hub, where p is the dimension of variates under consideration. This can be a burden when p is large or the communication costs between nodes are expensive. We consider two distributed LDA estimators, named two-round and one-shot estimator respectively, which are communication-efficient without transmitting p by p matrices. We study the asymptotic relative efficiency of distributed LDA estimators compared to centralized LDA using the random matrix theory under different settings of k . It is shown that when k is in a suitable range such as $k = o(n/p)$, these two distributed estimators achieve the same efficiency

*Corresponding author. (Email: zhaojunlong928@126.com)

as that of the centralized estimator under mild conditions. Moreover, the two-round estimator can relax the restriction on k allowing $kp/n \rightarrow c \in [0, 1)$ under some conditions. Simulations confirm the theoretical results.

Key words and phrases: Deterministic equivalent, distributed learning, linear discriminant analysis (LDA), random matrix, relative efficiency.

1. Introduction

With the rapid development of information technology, modern statistical inference often needs to deal with massive data. In many cases, the size of data is too large to be conveniently handled by a single data hub. Moreover, individual agents (e.g. local governments, hospitals, research labs) collect data independently and have constraints in communication due to the issues of costs, privacy, ownership, security, etc. Consequently, data has to be stored and also processed on many local computers connected to a central server, thus forming a distributed system. In this way, we are able to break a large-scale computation problem into many small pieces, then solve them with divide-and-conquer procedures and communicate only certain summary statistics. In recent years, distributed statistical inference has received considerable attention, covering a wide range of topics including M-estimation (Chen and Xie, 2014; Rosenblatt and Nadler, 2016;

Lee et al., 2017; Battey et al., 2018; Shi, Lu, and Song, 2018; Jordan et al., 2018; Banerjee, Durot, and Sen, 2019; Fan, Guo, and Wang, 2019), hypothesis test (Lalitha, Sarwate, and Javidi, 2014; Battey et al., 2018), confidence intervals (Jordan, Lee, and Yang, 2018; Chen, Liu, and Zhang, 2018; Dobriban and Sheng, 2018; Wang et al., 2019), principal component analysis (Garber, Shamir, and Srebro, 2017; Fan et al., 2019), nonparametric regression (Zhang, Duchi, and Wainwright, 2015; Chang, Lin, and Zhou, 2017; Shang and Cheng, 2017; Han et al., 2018; Szabó and Van Zanten, 2019), Bayesian methods (Xu et al., 2014; Jordan et al., 2018), quantile regression (Volgushev, Chao, and Cheng, 2019; Chen, Liu, and Zhang, 2019), bootstrap inference (Kleiner et al., 2014; Han and Liu, 2016), and so on.

Linear Discriminant Analysis (LDA) is a classical method for classification in statistics, and how to implement LDA in the distributed system has attracted the attention of some researchers recently. Suppose that $\{(\mathbf{X}_i, C_i), 1 \leq i \leq n\}$ are *i.i.d.* observations, where $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top$ is the p -dimensional covariate and C_i is the label. For ease of description, the classical LDA estimator is referred to as centralized LDA. In distributed systems, data are stored on k local machines, and for simplicity, we assume that the size of subsample for each machine is the same, denoted as $n^{(l)} \equiv n/k$ for $l = 1, \dots, k$. For a distributed LDA estimator, one can consider its

relative efficiency by comparing its classification accuracy with that of centralized LDA. Macua, Belanovic, and Zazo (2011) developed a distributed algorithm for LDA on a single-hop network in the classical regime with fixed dimension p , but the relative efficiency of their algorithm is unknown. Tian and Gu (2017) proposed a communication-efficient distributed sparse LDA estimator in the high dimensional regime, where the dimension p can be much larger than the sample size n . To ensure their distributed estimator attaining the same efficiency as the centralized one, the authors showed that k has the order $k = O(\sqrt{n/\log p}/\max(s, s'))$, where s and s' represent the sparsity of some parameters.

In this paper, we focus on distributed LDA for binary classification, under the setting of $p/n \rightarrow 0$ without the sparsity assumption on parameters. When $p/n \rightarrow 0$, we show that the centralized LDA can still be effective under mild conditions in Section 3. However, centralized LDA needs the transmission of local summary matrices of size p by p , which can be a burden when p is large or the communication costs between nodes are expensive. In response to this problem, we propose two communication-efficient distributed LDA estimators, named two-round estimator and one-shot one respectively according to their communication costs, without transmitting p by p summary matrices. We study the relative efficiency of these two

estimators. It is shown that both estimators achieve the same efficiency as the centralized one when k is in a suitable range such as $k = o(n/p)$. Moreover, under some conditions, the two-round estimator can relax the restriction on k , allowing $kp/n \rightarrow c \in [0, 1)$. When $c > 0$, the sample covariance matrix constructed with only data on the local machine is not a consistent estimator of the true covariance matrix, which brings challenges for the theoretical analysis. We successfully establish the efficiency of two-round estimator using the tools from random matrix theory. Interestingly, when the prior probabilities of two classes are equal (i.e. both are $1/2$), the two-round estimator still has the same efficiency as the centralized one even if $c > 0$.

The rest of this paper is organized as follows. In Section 2, we give the distributed LDA estimators and calculate corresponding classification accuracy. Section 3 studies the relative efficiency and derives their asymptotic properties. Section 4 provides numerical experiments to back up the developed theory. In Section 5, we discuss our results together with potential future directions.

Here we summarize notations to be used throughout the paper. We adopt the common convention of using boldface letters for vectors only, while regular font is used for both scalars and matrices. For a vector

$\mathbf{x} = (x_1, \dots, x_p)^\top \in \mathbb{R}^p$, and $0 < q < \infty$, define the ℓ_q norm by $\|\mathbf{x}\|_q = (\sum_{i=1}^p |x_i|^q)^{1/q}$. For a symmetric matrix $M \in \mathbb{R}^{p \times p}$, $\text{tr}(M)$ denotes the trace of M , and $\lambda_{\max}(M), \lambda_{\min}(M)$ represent the maximal eigenvalue and the minimal eigenvalue respectively. For a matrix $M \in \mathbb{R}^{n \times p}$, the nuclear norm is defined by $\|M\|_* = \text{tr}[(M^\top M)^{1/2}] = \sum_{i=1}^{\min\{n,p\}} \sigma_i(M)$, and the matrix ℓ_2 norm is defined as $\|M\|_2 = \sqrt{\lambda_{\max}(M^\top M)} = \sigma_1(M)$, where $\sigma_i(M)$ represents the i -th largest singular value. Besides, for two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \leq C|b_n|$ for all $n \geq 1$, and write $a_n = o(b_n)$ if $\lim_{n \rightarrow \infty} a_n/b_n = 0$. For two sequences of random variables $\{X_n\}$, $\{Y_n\}$, and a random variable X , write $X_n \rightarrow_{a.s.} X$ if $\{X_n\}$ converges to X almost surely, and $X_n \rightarrow_p X$ if $\{X_n\}$ converges to X in probability. In addition, write $X_n = O_p(Y_n)$ if X_n/Y_n is bounded in probability.

2. Communication-Efficient Distributed Linear Discriminant Analysis

2.1 Centralized linear discriminant analysis in the distributed system

In this paper, we focus on binary classification problems, assuming that the two classes are of normal distributions with the same covariance matrix,

2.1 Centralized linear discriminant analysis in the distributed system

specifically, $N_p(\boldsymbol{\mu}_1, \Sigma)$ for class 1 and $N_p(\boldsymbol{\mu}_2, \Sigma)$ for class 2, where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are p -dimensional mean vectors and the covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ is a positive symmetrical matrix. Denote $\boldsymbol{\mu}_a = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\Theta = \Sigma^{-1}$ being the precision matrix (a.k.a. the inverse covariance matrix). For a new observation $\mathbf{H} \in \mathbb{R}^p$ with prior probabilities π_1 and π_2 from class 1 and class 2 respectively, the Fisher's linear discriminant rule takes the form

$$\psi(\mathbf{H}) = \mathbb{1} \{ (\mathbf{H} - \boldsymbol{\mu}_a)^\top \Theta \boldsymbol{\mu}_d > \log(\pi_2/\pi_1) \}, \quad (2.1)$$

where $\mathbb{1}(\cdot)$ represents the indicator function. A new observation \mathbf{H} is classified into class 1 if $\psi(\mathbf{H}) = 1$, and class 2 otherwise. Clearly, there are two types of errors. Specifically, \mathbf{H} is from class 1 but is classified into class 2, and vice versa, with their probabilities denoted as follows

$$p_{21} = P(\psi(\mathbf{H}) = 0 \mid \mathbf{H} \in \text{class 1}), \quad p_{12} = P(\psi(\mathbf{H}) = 1 \mid \mathbf{H} \in \text{class 2}).$$

Then the efficiency of the LDA rule measured by classification accuracy is defined as

$$A_{cen} = 1 - \pi_1 p_{21} - \pi_2 p_{12}.$$

When $\mathbf{H} \sim N_p(\boldsymbol{\mu}_1, \Sigma)$, it holds that $(\mathbf{H} - \boldsymbol{\mu}_a)^\top \Theta \boldsymbol{\mu}_d \sim N(\delta^2/2, \delta^2)$, where $\delta^2 = \boldsymbol{\mu}_d^\top \Theta \boldsymbol{\mu}_d$ is the squared Mahalanobis distance between two populations.

2.1 Centralized linear discriminant analysis in the distributed system

Thus

$$p_{21} = \Phi\left(-\frac{\delta}{2} + \frac{\log(\pi_2/\pi_1)}{\delta}\right), \quad p_{12} = \Phi\left(-\frac{\delta}{2} - \frac{\log(\pi_2/\pi_1)}{\delta}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal. Then it follows that

$$A_{cen} = \pi_1 \Phi\left(\frac{\delta}{2} - \frac{\log(\pi_2/\pi_1)}{\delta}\right) + \pi_2 \Phi\left(\frac{\delta}{2} + \frac{\log(\pi_2/\pi_1)}{\delta}\right). \quad (2.2)$$

Particularly, when $\pi_1 = \pi_2 = 1/2$, we have $p_{21} = p_{12} = \Phi(-\delta/2)$, and then $A_{cen} = \Phi(\delta/2)$.

In practice, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, Σ and π_1 , π_2 are unknown and can be estimated from data. Suppose $\{\mathbf{X}_i : 1 \leq i \leq n_1\}$ and $\{\mathbf{Y}_i : 1 \leq i \leq n_2\}$ are independent and identically distributed observations from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$, respectively, where $n_1 + n_2 = n$. In this paper, we do not impose sparsity assumptions on the parameters. The centralized estimators of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Θ are

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_i, \quad \hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{Y}_i, \quad \hat{\Theta} = \hat{\Sigma}^{-1}, \quad (2.3)$$

respectively, where

$$\hat{\Sigma} = \frac{1}{n} \left[\sum_{i=1}^{n_1} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)^\top + \sum_{i=1}^{n_2} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_2)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_2)^\top \right]$$

is the pooled sample covariance matrix with $n = n_1 + n_2$. π_1 and π_2 can be simply estimated by $\hat{\pi}_1 = n_1/n$ and $\hat{\pi}_2 = n_2/n$. Plugging these estimators

2.1 Centralized linear discriminant analysis in the distributed system

into (2.1) gives rise to the empirical version of $\psi(\mathbf{H})$ as follows

$$\hat{\psi}(\mathbf{H}) = \mathbb{1} \left\{ (\mathbf{H} - \hat{\boldsymbol{\mu}}_a)^\top \hat{\Theta} \hat{\boldsymbol{\mu}}_d > \log(n_2/n_1) \right\}, \quad (2.4)$$

where $\hat{\boldsymbol{\mu}}_a = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$, $\hat{\boldsymbol{\mu}}_d = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$. For $\mathbf{H} \sim N_p(\boldsymbol{\mu}_j, \Sigma)$, $j = 1, 2$, it holds that

$$(\mathbf{H} - \hat{\boldsymbol{\mu}}_a)^\top \hat{\Theta} \hat{\boldsymbol{\mu}}_d \sim N \left((\boldsymbol{\mu}_j - \hat{\boldsymbol{\mu}}_a)^\top \hat{\Theta} \hat{\boldsymbol{\mu}}_d, (\hat{\Theta} \hat{\boldsymbol{\mu}}_d)^\top \Sigma \hat{\Theta} \hat{\boldsymbol{\mu}}_d \right).$$

Then given the samples $\{\mathbf{X}_i\}$ and $\{\mathbf{Y}_i\}$, the conditional misclassification rates of (2.4) are as follows (Cai and Liu, 2011)

$$\begin{aligned} \hat{p}_{12} &= 1 - \Phi \left(\frac{(\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_2)^\top \hat{\Theta} \hat{\boldsymbol{\mu}}_d + \log(n_2/n_1)}{\sqrt{(\hat{\Theta} \hat{\boldsymbol{\mu}}_d)^\top \Sigma \hat{\Theta} \hat{\boldsymbol{\mu}}_d}} \right), \\ \hat{p}_{21} &= 1 - \Phi \left(-\frac{(\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1)^\top \hat{\Theta} \hat{\boldsymbol{\mu}}_d + \log(n_2/n_1)}{\sqrt{(\hat{\Theta} \hat{\boldsymbol{\mu}}_d)^\top \Sigma \hat{\Theta} \hat{\boldsymbol{\mu}}_d}} \right). \end{aligned}$$

Thus the classification accuracy of centralized LDA is given by

$$\begin{aligned} \hat{A}_{cen} &= 1 - \hat{\pi}_1 \hat{p}_{21} - \hat{\pi}_2 \hat{p}_{12} \\ &= \frac{n_1}{n} \Phi \left(-\frac{(\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1)^\top \hat{\Theta} \hat{\boldsymbol{\mu}}_d + \log(n_2/n_1)}{\sqrt{(\hat{\Theta} \hat{\boldsymbol{\mu}}_d)^\top \Sigma \hat{\Theta} \hat{\boldsymbol{\mu}}_d}} \right) \\ &\quad + \frac{n_2}{n} \Phi \left(\frac{(\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_2)^\top \hat{\Theta} \hat{\boldsymbol{\mu}}_d + \log(n_2/n_1)}{\sqrt{(\hat{\Theta} \hat{\boldsymbol{\mu}}_d)^\top \Sigma \hat{\Theta} \hat{\boldsymbol{\mu}}_d}} \right). \end{aligned} \quad (2.5)$$

When the data are stored on k machines, the implementation of centralized LDA is still feasible with the price of considerable communication

2.1 Centralized linear discriminant analysis in the distributed system

costs. Let $\mathbb{X}^{(l)}$ be the data from class 1 stored on the l -th local machine (i.e. the collection of \mathbf{X}_i 's that are stored on the l -th machine), and similarly, let $\mathbb{Y}^{(l)}$ be the data from class 2 stored on the l -th machine, $l = 1, \dots, k$. For clarity, we denote

$$\mathbb{X}^{(l)} = \{\mathbf{X}_i^{(l)}, i = 1, \dots, n_{1l}\}, \quad \mathbb{Y}^{(l)} = \{\mathbf{Y}_i^{(l)}, i = 1, \dots, n_{2l}\}, \quad l = 1, \dots, k, \quad (2.6)$$

where $n_{1l} > 0$ and $n_{2l} > 0$ are the cardinalities of $\mathbb{X}^{(l)}$ and $\mathbb{Y}^{(l)}$, respectively. Thus, the l -th machine (or worker) has access to only a subset of $n^{(l)} = n_{1l} + n_{2l}$ observations out of the total n observations. Obviously, it holds that

$$n = \sum_{l=1}^k n^{(l)} = n_1 + n_2, \quad n_j = \sum_{l=1}^k n_{jl}, \quad j = 1, 2. \quad (2.7)$$

Denote $B_x = \sum_{i=1}^{n_1} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_1)^\top$ and $B_y = \sum_{i=1}^{n_2} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_2)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_2)^\top$. Then $\hat{\Sigma} = n^{-1}(B_x + B_y)$. Let $B_x^{(l)} = \sum_{i=1}^{n_{1l}} \mathbf{X}_i^{(l)} \mathbf{X}_i^{(l)\top}$, and $\hat{\boldsymbol{\mu}}_1^{(l)}$ be the sample mean obtained with data $\mathbb{X}^{(l)}$. It is easy to see that

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1} \sum_{l=1}^k n_{1l} \hat{\boldsymbol{\mu}}_1^{(l)}, \quad B_x = \sum_{i=1}^{n_1} \mathbf{X}_i \mathbf{X}_i^\top - n_1 \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1^\top = \sum_{l=1}^k B_x^{(l)} - n_1 \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1^\top,$$

Since both $B_x^{(l)}$ and $\hat{\boldsymbol{\mu}}_1^{(l)}$ are computed locally with data on the l -th machine, we see that B_x can be obtained by transmitting some summary matrices and vectors (i.e. $B_x^{(l)}$'s and $\hat{\boldsymbol{\mu}}_1^{(l)}$'s) to the hub. Computing B_y similarly, one can obtain $\hat{\Sigma}$ and consequently $\hat{\Theta}$. $\hat{\boldsymbol{\mu}}_a$ and $\hat{\boldsymbol{\mu}}_d$ can be computed in a

2.2 Distributed linear discriminant analysis by averaging

similar fashion by transmitting $\hat{\boldsymbol{\mu}}_j^{(l)}$'s, $j = 1, 2$. However, the centralized estimator requires the transmission of p -dimensional mean vectors $\hat{\boldsymbol{\mu}}_1^{(l)}, \hat{\boldsymbol{\mu}}_2^{(l)}$ and p by p matrices $B_x^{(l)}, B_y^{(l)}$, where $l = 1, \dots, k$. When p and k are large, transmitting these p by p matrices to the central hub can be a burden in communication, while transmitting p -dimensional mean vectors is much easier. In the following Section 2.2, we propose two distributed estimators without transmitting these p by p matrices.

2.2 Distributed linear discriminant analysis by averaging

In this subsection, we considered the communication-efficient LDA estimators. Recall that $\mathbb{X}^{(l)} = \{\mathbf{X}_i^{(l)}, i = 1, \dots, n_{1l}\}$ and $\mathbb{Y}^{(l)} = \{\mathbf{Y}_i^{(l)}, i = 1, \dots, n_{2l}\}$ are the data on the l -th local machine, where n_{jl} 's satisfy (2.7), $j = 1, 2$ and $l = 1, \dots, k$. Suppose that $\{\mathbf{X}_i^{(l)}, i = 1, \dots, n_{1l}, l = 1, \dots, k\}$ are *i.i.d.* observations from $N_p(\boldsymbol{\mu}_1, \Sigma)$, and that $\{\mathbf{Y}_i^{(l)}, i = 1, \dots, n_{2l}, l = 1, \dots, k\}$ are *i.i.d.* observations from $N_p(\boldsymbol{\mu}_2, \Sigma)$. Assume that $n_{jl} \geq 2, j = 1, 2$ for all l . Denote the estimators of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ using data on the l -th machine as follows

$$\hat{\boldsymbol{\mu}}_1^{(l)} = \frac{1}{n_{1l}} \sum_{i=1}^{n_{1l}} \mathbf{X}_i^{(l)}, \quad \hat{\boldsymbol{\mu}}_2^{(l)} = \frac{1}{n_{2l}} \sum_{i=1}^{n_{2l}} \mathbf{Y}_i^{(l)}.$$

As argued at the end of Section 2.1, we prefer an estimator without transmitting the p by p matrices. We consider two types of distributed LDA

2.2 Distributed linear discriminant analysis by averaging

estimators. The first one is named two-round distributed LDA estimator, which estimates the mean vectors using full data with two rounds of communication. And the second one is called one-shot estimator that estimates the means based on the local data with just one round of communication.

(1) We introduce the two-round distributed LDA estimator. By aggregating the local estimators, we estimate $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and Θ as follows

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1} \sum_{l=1}^k n_{1l} \hat{\boldsymbol{\mu}}_1^{(l)}, \quad \hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2} \sum_{l=1}^k n_{2l} \hat{\boldsymbol{\mu}}_2^{(l)}, \quad \bar{\Theta} = \frac{1}{n} \sum_{l=1}^k n^{(l)} \hat{\Theta}_{two}^{(l)}, \quad (2.8)$$

where $\hat{\Theta}_{two}^{(l)} = (\hat{\Sigma}_{two}^{(l)})^{-1}$, and

$$\hat{\Sigma}_{two}^{(l)} = \frac{1}{n^{(l)}} \left[\sum_{i=1}^{n_{1l}} (\mathbf{X}_i^{(l)} - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_i^{(l)} - \hat{\boldsymbol{\mu}}_1)^\top + \sum_{i=1}^{n_{2l}} (\mathbf{Y}_i^{(l)} - \hat{\boldsymbol{\mu}}_2)(\mathbf{Y}_i^{(l)} - \hat{\boldsymbol{\mu}}_2)^\top \right].$$

It is easy to see that $\hat{\Theta}_{two}^{(l)}$ can be obtained using data on the l -th machine after giving $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$. Recall that $\hat{\boldsymbol{\mu}}_a = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$ and $\hat{\boldsymbol{\mu}}_d = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$.

Then we define the discriminant rule of two-round distributed LDA as

$$\bar{\psi}_{two}(\mathbf{H}) = \mathbf{1} \{ (\mathbf{H} - \hat{\boldsymbol{\mu}}_a)^\top \bar{\Theta} \hat{\boldsymbol{\mu}}_d > \log(n_2/n_1) \}. \quad (2.9)$$

As shown in the following Algorithm 1, $\bar{\psi}_{two}(\mathbf{H})$ can be computed in a communication-efficient way with only the p -dimensional mean vectors being transmitted for two rounds. Comparing $\bar{\psi}_{two}(\mathbf{H})$ with its centralized counterpart $\hat{\psi}(\mathbf{H})$, one can see that the only difference between these two estimators lies in the different estimation of Θ . For the centralized estima-

2.2 Distributed linear discriminant analysis by averaging

tor, Θ is estimated by $\hat{\Theta} = \hat{\Sigma}^{-1}$ with $\hat{\Sigma}$ being obtained by transmitting p by p matrices $B_x^{(l)}$'s and $B_y^{(l)}$'s to the hub.

Algorithm 1 Two-Round Distributed LDA

- Input:** Observation \mathbf{H} and data matrices $\mathbb{X}^{(l)}, \mathbb{Y}^{(l)}$ on the l -th machine, $l \in \{1, \dots, k\}$, do the following steps.
- 1: Compute local sample means $\hat{\boldsymbol{\mu}}_1^{(l)}, \hat{\boldsymbol{\mu}}_2^{(l)}$ on local machines, and then transmit them to the hub.
 - 2: Compute on the hub the estimator $\hat{\boldsymbol{\mu}}_j$ by (2.8), $j \in \{1, 2\}$, and then compute $\hat{\boldsymbol{\mu}}_a$ and $\hat{\boldsymbol{\mu}}_d$.
 - 3: Broadcast $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\mu}}_d$ and $\hat{\boldsymbol{\mu}}_a$ to each local machine. Compute $\hat{\Theta}_{two}^{(l)}$ by (2.8), and obtain $V_l = \hat{\boldsymbol{\mu}}_a^\top \hat{\Theta}_{two}^{(l)} \hat{\boldsymbol{\mu}}_d$ and $\mathbf{U}_l = \hat{\Theta}_{two}^{(l)} \hat{\boldsymbol{\mu}}_d$ with data on the l -th machine, $l \in \{1, \dots, k\}$.
 - 4: Send V_l 's and \mathbf{U}_l 's to the hub, and compute their averages $\bar{\mathbf{U}} = n^{-1} \sum_{l=1}^k n^{(l)} \mathbf{U}_l$ and $\bar{V} = n^{-1} \sum_{l=1}^k n^{(l)} V_l$. Then define the distributed LDA estimator $\bar{\psi}_{two}(\mathbf{H}) = \mathbf{H}^\top \bar{\mathbf{U}} - \bar{V}$.
 - 5: **return** Classification result $\bar{\psi}_{two}(\mathbf{H})$
-

There are two rounds of communication in Algorithm 1. First, the local estimators $\hat{\boldsymbol{\mu}}_1^{(l)}$'s and $\hat{\boldsymbol{\mu}}_2^{(l)}$'s are transmitted to the hub to compute $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\mu}}_a$ and $\hat{\boldsymbol{\mu}}_d$, and then broadcast the vector $(\hat{\boldsymbol{\mu}}_1^\top, \hat{\boldsymbol{\mu}}_2^\top, \hat{\boldsymbol{\mu}}_a^\top, \hat{\boldsymbol{\mu}}_d^\top)^\top \in \mathbb{R}^{4p}$ to each

2.2 Distributed linear discriminant analysis by averaging

local node. The second round is to send \mathbf{U}_l 's and V_l 's to the central hub. Note that in each round, we only transmit vectors with dimension no more than $4p$, avoiding the transmission of p by p matrices in the centralized estimator. Thus the estimator is computationally efficient.

Similar to the centralized estimator, we define the conditional misclassification rates of two-round distributed LDA. Let

$$\begin{aligned}\bar{p}_{12} &= 1 - \Phi \left(\frac{(\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_2)^\top \bar{\Theta} \hat{\boldsymbol{\mu}}_d + \log(n_2/n_1)}{\sqrt{(\bar{\Theta} \hat{\boldsymbol{\mu}}_d)^\top \Sigma \bar{\Theta} \hat{\boldsymbol{\mu}}_d}} \right), \\ \bar{p}_{21} &= 1 - \Phi \left(-\frac{(\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1)^\top \bar{\Theta} \hat{\boldsymbol{\mu}}_d + \log(n_2/n_1)}{\sqrt{(\bar{\Theta} \hat{\boldsymbol{\mu}}_d)^\top \Sigma \bar{\Theta} \hat{\boldsymbol{\mu}}_d}} \right),\end{aligned}$$

which are the counterparts of \hat{p}_{12} and \hat{p}_{21} , respectively. Hence, the classification accuracy of two-round estimator is equal to

$$\begin{aligned}\hat{A}_{two} &= 1 - \hat{\pi}_1 \bar{p}_{21} - \hat{\pi}_2 \bar{p}_{12} \\ &= \frac{n_1}{n} \Phi \left(-\frac{(\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1)^\top \bar{\Theta} \hat{\boldsymbol{\mu}}_d + \log(n_2/n_1)}{\sqrt{(\bar{\Theta} \hat{\boldsymbol{\mu}}_d)^\top \Sigma \bar{\Theta} \hat{\boldsymbol{\mu}}_d}} \right) \\ &\quad + \frac{n_2}{n} \Phi \left(\frac{(\hat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_2)^\top \bar{\Theta} \hat{\boldsymbol{\mu}}_d + \log(n_2/n_1)}{\sqrt{(\bar{\Theta} \hat{\boldsymbol{\mu}}_d)^\top \Sigma \bar{\Theta} \hat{\boldsymbol{\mu}}_d}} \right).\end{aligned}\tag{2.10}$$

Define the relative efficiency of two-round estimator as $\hat{R}_{two} = \hat{A}_{two} / \hat{A}_{cen}$.

(2) When communication between nodes is prohibitively expensive, we consider the one-shot estimator, where only one round of communication is

2.2 Distributed linear discriminant analysis by averaging

required. Denote the estimator of Θ with data on the l -th machine as

$$\hat{\Theta}^{(l)} = (\hat{\Sigma}^{(l)})^{-1},$$

where

$$\hat{\Sigma}^{(l)} = \frac{1}{n^{(l)}} \left[\sum_{i=1}^{n_{1l}} (\mathbf{X}_i^{(l)} - \hat{\boldsymbol{\mu}}_1^{(l)})(\mathbf{X}_i^{(l)} - \hat{\boldsymbol{\mu}}_1^{(l)})^\top + \sum_{i=1}^{n_{2l}} (\mathbf{Y}_i^{(l)} - \hat{\boldsymbol{\mu}}_2^{(l)})(\mathbf{Y}_i^{(l)} - \hat{\boldsymbol{\mu}}_2^{(l)})^\top \right].$$

Different from $\hat{\Sigma}_{two}^{(l)}$ that estimates the means by $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$, $\hat{\Sigma}^{(l)}$ here uses estimators $\hat{\boldsymbol{\mu}}_1^{(l)}$ and $\hat{\boldsymbol{\mu}}_2^{(l)}$. The discriminant rule of the one-shot estimator is defined as follows

$$\bar{\psi}_{one}(\mathbf{H}) = \mathbb{1} \left\{ \frac{1}{n} \sum_{l=1}^k n^{(l)} \left(\mathbf{H} - \hat{\boldsymbol{\mu}}_a^{(l)} \right)^\top \hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)} > \log(n_2/n_1) \right\}, \quad (2.11)$$

where $\hat{\boldsymbol{\mu}}_a^{(l)} = (\hat{\boldsymbol{\mu}}_1^{(l)} + \hat{\boldsymbol{\mu}}_2^{(l)})/2$, $\hat{\boldsymbol{\mu}}_d^{(l)} = \hat{\boldsymbol{\mu}}_1^{(l)} - \hat{\boldsymbol{\mu}}_2^{(l)}$. Note that

$$\begin{aligned} & \frac{1}{n} \sum_{l=1}^k n^{(l)} \left(\mathbf{H} - \hat{\boldsymbol{\mu}}_a^{(l)} \right)^\top \hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)} \\ &= \mathbf{H}^\top \left(\frac{1}{n} \sum_{l=1}^k n^{(l)} \hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)} \right) - \frac{1}{n} \sum_{l=1}^k n^{(l)} \hat{\boldsymbol{\mu}}_a^{(l)\top} \hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)} \end{aligned}$$

and that the p -dimensional vector $\hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)}$ and the scalar $\hat{\boldsymbol{\mu}}_a^{(l)\top} \hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)}$ can be computed directly using data on the l -th machine. Thus, we only need to transmit vectors of dimension $p+1$ to the hub in just one-round communication. For $\mathbf{H} \sim N_p(\boldsymbol{\mu}_j, \Sigma)$, $j = 1, 2$, it holds that $n^{-1} \sum_{l=1}^k n^{(l)} (\mathbf{H} - \hat{\boldsymbol{\mu}}_a^{(l)})^\top \hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)}$ follows a normal distribution with mean $n^{-1} \sum_{l=1}^k n^{(l)} (\boldsymbol{\mu}_j - \hat{\boldsymbol{\mu}}_a^{(l)})^\top \hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)}$ and variance

$$\left(\sum_{l=1}^k \frac{n^{(l)}}{n} \hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)} \right)^\top \Sigma \left(\sum_{l=1}^k \frac{n^{(l)}}{n} \hat{\Theta}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)} \right).$$

Thus the corresponding classification accuracy of one-shot distributed LDA equals

$$\begin{aligned} \hat{A}_{one} = & \frac{n_1}{n} \Phi \left(- \frac{n^{-1} \sum_{l=1}^k n^{(l)} (\hat{\boldsymbol{\mu}}_a^{(l)} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\Theta}}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)} + \log(n_2/n_1)}{\sqrt{(n^{-1} \sum_{l=1}^k n^{(l)} \hat{\boldsymbol{\Theta}}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)})^\top \Sigma (n^{-1} \sum_{l=1}^k n^{(l)} \hat{\boldsymbol{\Theta}}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)})}} \right) \\ & + \frac{n_2}{n} \Phi \left(\frac{n^{-1} \sum_{l=1}^k n^{(l)} (\hat{\boldsymbol{\mu}}_a^{(l)} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\Theta}}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)} + \log(n_2/n_1)}{\sqrt{(n^{-1} \sum_{l=1}^k n^{(l)} \hat{\boldsymbol{\Theta}}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)})^\top \Sigma (n^{-1} \sum_{l=1}^k n^{(l)} \hat{\boldsymbol{\Theta}}^{(l)} \hat{\boldsymbol{\mu}}_d^{(l)})}} \right). \end{aligned}$$

Define $\hat{R}_{one} = \hat{A}_{one}/\hat{A}_{cen}$ as the relative efficiency of one-shot estimator. In the following Section 3, we study the conditions under which the distributed estimators reach the same efficiency as that of the centralized one. It is shown that the two-round estimator requires a weaker assumption on k , compared with the one-shot case.

3. Theoretical Properties

3.1 Deterministic equivalent of the sample covariance matrix

In this section, we compare the efficiency of the distributed and centralized LDA. Denote $\gamma_p = p/n$, $\gamma_p^{(l)} = p/n^{(l)}$, $l = 1, \dots, k$. For simplicity, we assume that data are evenly distributed to each machine, that is,

$$n_{11} = \dots = n_{1k} = n_1/k, \quad n_{21} = \dots = n_{2k} = n_2/k. \quad (3.12)$$

From (3.12), it follows that $n^{(l)} \equiv n/k$ and $\gamma_p^{(l)} \equiv kp/n = k\gamma_p$ for all l . Here the assumption (3.12) is assumed to reduce the complexity of notations.

3.1 Deterministic equivalent of the sample covariance matrix

The results in this section can be extended without difficulty to the case where n_{jl} 's are different but with the same order, $j = 1, 2$.

In this paper, we consider the case of $\gamma_p \rightarrow 0$ but $\gamma_p^{(l)} \rightarrow c \in [0, 1)$. For distributed estimators, when $c \neq 0$, the sample covariance matrix constructed with data on the l -th machine will not be a consistent estimator of Σ . Consequently, $\hat{\Theta}_{two}^{(l)}$ and $\hat{\Theta}^{(l)}$ are not consistent estimators of Θ , which brings challenges for the theoretical analysis. We study the asymptotic properties of \hat{A}_{two} , \hat{A}_{one} and \hat{A}_{cen} based on random matrix theory. Specifically, we will use the technique of deterministic equivalents (Couillet and Debbah, 2011, Chap. 6) from random matrix theory to obtain limits of some random quantities. The notion of equivalence is defined as follows.

Definition 1. (Dobriban and Sheng, 2018) The (deterministic or random) matrix sequences A_n , B_n of growing dimensions are equivalent, and write

$A_n \asymp B_n$ if

$$\lim_{n \rightarrow \infty} |\text{tr}[C_n(A_n - B_n)]| = 0$$

almost surely, for any sequence C_n of not necessarily symmetric matrices with bounded nuclear norm, i.e., such that $\lim_{n \rightarrow \infty} \sup \|C_n\|_* < \infty$.

Dobriban and Sheng (2018) studied the deterministic equivalent of the sample covariance matrix in elliptical models, which is a consequence of

3.1 Deterministic equivalent of the sample covariance matrix

generalized Marchenko-Pastur theorem (Rubio and Mestre, 2011). For the elliptical model, observations take the form $\{\mathbf{z}_i = g_i^{1/2} \Sigma^{1/2} \mathbf{u}_i, 1 \leq i \leq m\}$, where $\mathbf{u}_i \in \mathbb{R}^p$ is a vector with *i.i.d.* entries, g_i is a datapoint-specific scale parameter allowing observations having different scales, and $\Sigma \in \mathbb{R}^{p \times p}$ is the covariance matrix of \mathbf{z}_i . A special case of the elliptical model is \mathbf{z}_i following a normal distribution, where we have $g_i = 1$. Arrange the samples as rows of matrix Z , that has the form

$$Z = \Gamma^{1/2} U \Sigma^{1/2} \in \mathbb{R}^{m \times p}, \quad (3.13)$$

where $\Gamma = \text{diag}(g_1, \dots, g_m) \in \mathbb{R}^{m \times m}$ is the diagonal scaling matrix containing the scales g_i of samples, and $U = (\mathbf{u}_1, \dots, \mathbf{u}_m)^\top \in \mathbb{R}^{m \times p}$ has *i.i.d.* entries. Suppose that $E(U) = 0$ and let $\tilde{\Sigma}_* = m^{-1} Z^\top Z$ be the sample covariance matrix. Under some conditions, the random matrix $\tilde{\Sigma}_*^{-1}$ has a deterministic equivalent

$$\tilde{\Sigma}_*^{-1} \asymp e_p \Theta. \quad (3.14)$$

Here $e_p = e_p(m, p, \Gamma) > 0$ is the unique solution of the fixed-point equation

$$1 = \frac{1}{m} \text{tr} [e_p \Gamma (I_m + \gamma_p e_p \Gamma)^{-1}]. \quad (3.15)$$

To study our problem, define the following pooled sample covariance

3.1 Deterministic equivalent of the sample covariance matrix

matrices with known $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$

$$\begin{aligned}\tilde{\Sigma} &= \frac{1}{n} \left[\sum_{i=1}^{n_1} (\mathbf{X}_i - \boldsymbol{\mu}_1)(\mathbf{X}_i - \boldsymbol{\mu}_1)^\top + \sum_{i=1}^{n_2} (\mathbf{Y}_i - \boldsymbol{\mu}_2)(\mathbf{Y}_i - \boldsymbol{\mu}_2)^\top \right], \\ \tilde{\Sigma}^{(l)} &= \frac{1}{n^{(l)}} \left[\sum_{i=1}^{n_{1l}} (\mathbf{X}_i^{(l)} - \boldsymbol{\mu}_1)(\mathbf{X}_i^{(l)} - \boldsymbol{\mu}_1)^\top + \sum_{i=1}^{n_{2l}} (\mathbf{Y}_i^{(l)} - \boldsymbol{\mu}_2)(\mathbf{Y}_i^{(l)} - \boldsymbol{\mu}_2)^\top \right],\end{aligned}$$

where $l = 1, \dots, k$. To give the deterministic equivalents of $\tilde{\Sigma}^{-1}$ and $(\tilde{\Sigma}^{(l)})^{-1}$, we introduce the following conditions firstly.

(C1) Assume that (i) $0 < c_1 < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < c_2$ and that (ii)

$\gamma_p = p/n \rightarrow 0$, where c_1 and c_2 are constants independent of p .

(C2) As $n \rightarrow \infty$, $k = k_n$ satisfies the following conditions: (i) $n^{(l)} \equiv n/k \rightarrow$

∞ ; (ii) $\gamma_p^{(l)} \equiv pk/n \rightarrow c \in [0, 1)$, $l \in \{1, \dots, k\}$.

Condition (i) of (C1) is commonly assumed in the literature. Condition $\gamma_p \rightarrow 0$ implies that the sample covariance matrix obtained using full data will be a consistent estimator of Σ (Wainwright, 2019, Chap. 11). So the inverse sample covariance matrix can be a consistent estimator of Θ , which guarantees the effectiveness of centralized LDA. But for the distributed system, as the number k of local machines increases, namely $k \rightarrow \infty$, it may occur that $\gamma_p^{(l)} = kp/n \rightarrow c > 0$. The local sample covariance matrix $\hat{\Sigma}_{two}^{(l)}$ and its inverse $\hat{\Theta}_{two}^{(l)}$ based on data on the l -th machine will be inconsistent (Bai and Silverstein, 2010, Chap. 3).

3.2 Relative efficiency

Proposition 1. *Under the condition (C1), for the sample covariance matrix $\tilde{\Sigma}$ with known μ_1 and μ_2 , we have*

$$\tilde{\Sigma}^{-1} \asymp \Theta. \quad (3.16)$$

Under the condition (i) of (C1) and the condition (C2), for the sample covariance matrix $\tilde{\Sigma}^{(l)}$ with known μ_1 and μ_2 on the l -th machine, we have

$$(\tilde{\Sigma}^{(l)})^{-1} \asymp \frac{1}{1 - \gamma_p^{(l)}} \Theta, \quad l = 1, \dots, k. \quad (3.17)$$

Particularly, if taking $c = 0$ in (C2), we have $(\tilde{\Sigma}^{(l)})^{-1} \asymp \Theta, l = 1, \dots, k$.

This important conclusion will serve as the basis of the following theorems.

3.2 Relative efficiency

As defined at the end of Section 2.2, the relative efficiency of distributed LDA compared to the centralized case is the ratio of their classification accuracy. Then the relative efficiency of two-round distributed LDA equals

$$\hat{R}_{two} = \frac{n_1 \Phi \left(-\frac{(\hat{\mu}_a - \mu_1)^\top \bar{\Theta} \hat{\mu}_d + \log(n_2/n_1)}{\hat{\Delta}_p} \right) + n_2 \Phi \left(\frac{(\hat{\mu}_a - \mu_2)^\top \bar{\Theta} \hat{\mu}_d + \log(n_2/n_1)}{\hat{\Delta}_p} \right)}{n_1 \Phi \left(-\frac{(\hat{\mu}_a - \mu_1)^\top \hat{\Theta} \hat{\mu}_d + \log(n_2/n_1)}{\hat{\Delta}_p} \right) + n_2 \Phi \left(\frac{(\hat{\mu}_a - \mu_2)^\top \hat{\Theta} \hat{\mu}_d + \log(n_2/n_1)}{\hat{\Delta}_p} \right)}, \quad (3.18)$$

where $\bar{\Delta}_p^2 = (\bar{\Theta} \hat{\mu}_d)^\top \Sigma \bar{\Theta} \hat{\mu}_d$, and $\hat{\Delta}_p^2 = (\hat{\Theta} \hat{\mu}_d)^\top \Sigma \hat{\Theta} \hat{\mu}_d$. To study the properties of \hat{R}_{two} , we define its population version as

$$R_{two} = A_{two}/A_{cen},$$

3.2 Relative efficiency

where A_{cen} is defined in (2.2), and

$$A_{two} = \pi_1 \Phi \left(\frac{\delta}{2} - \frac{(1-c) \log(\pi_2/\pi_1)}{\delta} \right) + \pi_2 \Phi \left(\frac{\delta}{2} + \frac{(1-c) \log(\pi_2/\pi_1)}{\delta} \right), \quad (3.19)$$

with $\delta^2 = \boldsymbol{\mu}_d^\top \Theta \boldsymbol{\mu}_d$. It is easy to see that $R_{two} \leq 1$ for any $c \in [0, 1]$.

Particularly, we have

$$\begin{cases} R_{two} = 1, & c = 0; \\ R_{two} = 1, & c \in (0, 1), \pi_1 = 1/2; \\ R_{two} < 1, & c \in (0, 1), \pi_1 \neq 1/2. \end{cases} \quad (3.20)$$

The following Theorem 1 establishes the properties of two-round distributed LDA.

Theorem 1. *Under (C1) and (C2), as $n \rightarrow \infty$, it holds that $\hat{A}_{two} \rightarrow_p A_{two}$ and $\hat{A}_{cen} \rightarrow_p A_{cen}$. Consequently, $\hat{R}_{two} \rightarrow_p R_{two}$.*

According to (3.20), we discuss Theorem 1 in three cases. (1) $c = 0$; (2) $c \in (0, 1)$ and $\pi_1 = 1/2$; (3) $c \in (0, 1)$ and $\pi_1 \neq 1/2$. For Case (1), when $c = 0$ or equivalently k satisfies $k = o(n/p)$, the two-round estimator has the same efficiency as the centralized estimator. This coincides with our expectation, since $\hat{\Sigma}_{two}^{(l)}$ is a good estimator of Σ when k is small. Case (2) is an interesting result beyond expectation. When $c \in (0, 1)$, we see that k has the order as n/p . By the well known results of random matrix

3.2 Relative efficiency

(Bai and Silverstein, 2010, Chap. 3), the local sample covariance matrix $\hat{\Sigma}_{two}^{(l)}$ is not a consistent estimator of Σ . However, Theorem 1 shows that, as long as $\pi_1 = 1/2$, the distributed estimator has the same efficiency as the centralized one, regardless of the value of c . In other words, even if each local estimator $\hat{\Sigma}_{two}^{(l)}$ of the sample covariance matrix is inconsistent, the distributed estimator loses no information when $\pi_1 = 1/2$. For Case (3), when k has the same order as n/p but $\pi_1 \neq 1/2$, the two-round distributed LDA will lose efficiency. The following Theorem 2 gives the results on \hat{R}_{one} .

Theorem 2. *Suppose that (C1) and (C2) hold but with $c = 0$ in (C2). As $n \rightarrow \infty$, both \hat{A}_{one} and \hat{A}_{cen} converge to A_{cen} in probability. Consequently, \hat{R}_{one} converges to 1 in probability.*

Condition $c = 0$ in (C2) in Theorem 2 implies that $k = o(n/p)$, that is, when k is small, the one-shot estimator can achieve the same efficiency as the centralized one. We briefly discuss the difference between the two-round estimator and the one-shot one. Two-round estimator replaces local sample means $\hat{\mu}_i^{(l)}$'s with global sample means $\hat{\mu}_i$'s through an extra round of communication, which relaxes the restriction on k , allowing k being the same order as n/p . Particularly, when $\pi_1 = 1/2$, the two-round estimator loses no information even if k is exactly the same order of n/p (i.e. $c \in (0, 1)$). However, we do not have similar results for the one-shot estimator.

In fact, when $\gamma_p^{(l)} \rightarrow c \in (0, 1)$, $\hat{\boldsymbol{\mu}}_i^{(l)}$'s are no longer consistent estimators of $\boldsymbol{\mu}_i$'s in terms of ℓ_2 norm. In this case, from the proof of Theorem 2 (see Section S4 in Supplementary Material), it is easy to see that \hat{A}_{one} converges to the quantity

$$\begin{aligned} & \pi_1 \Phi \left(\frac{\delta^2(1 + E_1) - 2(1 - c) \log(\pi_2/\pi_1)}{2\sqrt{\delta^2(1 + E_2)}} \right) \\ & + \pi_2 \Phi \left(\frac{\delta^2(1 + E_3) + 2(1 - c) \log(\pi_2/\pi_1)}{2\sqrt{\delta^2(1 + E_2)}} \right), \end{aligned}$$

where E_i 's are random variables representing the addition bias caused by the local estimators $\boldsymbol{\mu}_i^{(l)}$'s, satisfying $E_i = O_p(c)$, $i = 1, 2, 3$. When $c > 0$, \hat{A}_{one} may not have a constant limit as $n \rightarrow \infty$.

4. Simulations

In this section, we illustrate the performance of distributed LDA methods by comparing it with centralized LDA. To begin with, we introduce the setup in the simulation study. The training data is generated as follows. We first withdraw independent and identically distributed observations of size n from normal distributions $N_p(\boldsymbol{\mu}_1, \Sigma)$ (class 1) and $N_p(\boldsymbol{\mu}_2, \Sigma)$ (class 2) with each class having $n/2$ observations, and then distribute the samples in each class equally at random on k machines. Moreover, we generate $N/2$ observations in each class as the testing set. In the following simulation, we

set $N = 1000$, $\boldsymbol{\mu}_1 = (0, \dots, 0)^\top \in \mathbb{R}^p$ and $\boldsymbol{\mu}_2 = (0.2, \dots, 0.2)^\top \in \mathbb{R}^p$. The covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ is generated as follows.

Example 1. (Toeplitz matrix) $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = (2 - |i - j|)_+$ for $1 \leq i, j \leq p$.

Example 2. (Approximately sparse matrix) $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0.8^{|i-j|}$ for $1 \leq i, j \leq p$.

We consider the following four cases.

Case 1a. Σ is from Example 1. Fix $k = 5$ and set $p = \lceil n^{1/2} \rceil$, where $\lceil a \rceil$ denotes the integral part of a constant a . It is seen that $\gamma_p = p/n \rightarrow 0$ and $\gamma_p^{(l)} = pk/n \rightarrow 0$. Then we set $n \in \{100 + (i - 1) \times 10^3, i = 1, \dots, 11\}$.

Case 1b. Σ is from Example 1. Set $k = \lceil cn^{3/5} \rceil$ and $p = \lceil n^{2/5} \rceil$, where $c \in \{0.1, 0.3, 0.6\}$. It is seen that $\gamma_p = p/n \rightarrow 0$ and $\gamma_p^{(l)} = pk/n \rightarrow c$. Then we let $n \in \{100 + (i - 1) \times 10^3, i = 1, \dots, 11\}$.

Case 2a. Σ is from Example 2. Other settings are the same as *Case 1a*.

Case 2b. Σ is from Example 2. Other settings are the same as *Case 1b*.

For each case, we perform distributed LDA and centralized LDA on the training set to estimate the classification rule and compute relative efficiency based on the testing set. Then we repeat the procedure 100 times to calculate the average value of relative efficiency. In the following Figure 1, we report the average values of \hat{R}_{two} for two-round distributed LDA, and those of \hat{R}_{one} for one-shot distributed LDA.

For Case 1a and Case 2a where $c = 0$, as n and p increase, both \hat{R}_{two} and \hat{R}_{one} converge to 1 quickly, coinciding with theoretical findings, showing that both distributed estimators perform as well as the centralized one. Then we turn to Case 1b and Case 2b, where $c > 0$. When c is small (e.g. $c = 0.1$), the values of \hat{R}_{two} and \hat{R}_{one} are very close to 1, even n is small such as 100, and there is no significant difference between two estimators. However, when c is large (e.g. $c = 0.3$ or 0.6), we see that \hat{R}_{two} is still very close to 1 for large n , but \hat{R}_{one} is much worse than that of \hat{R}_{two} , especially when $c = 0.6$. This supports our theoretical findings.

5. Discussion

In this paper, we study the Fisher's linear discriminant analysis in distributed systems for binary classification, proposing two communication-efficient estimators. The classification accuracy is calculated for distributed

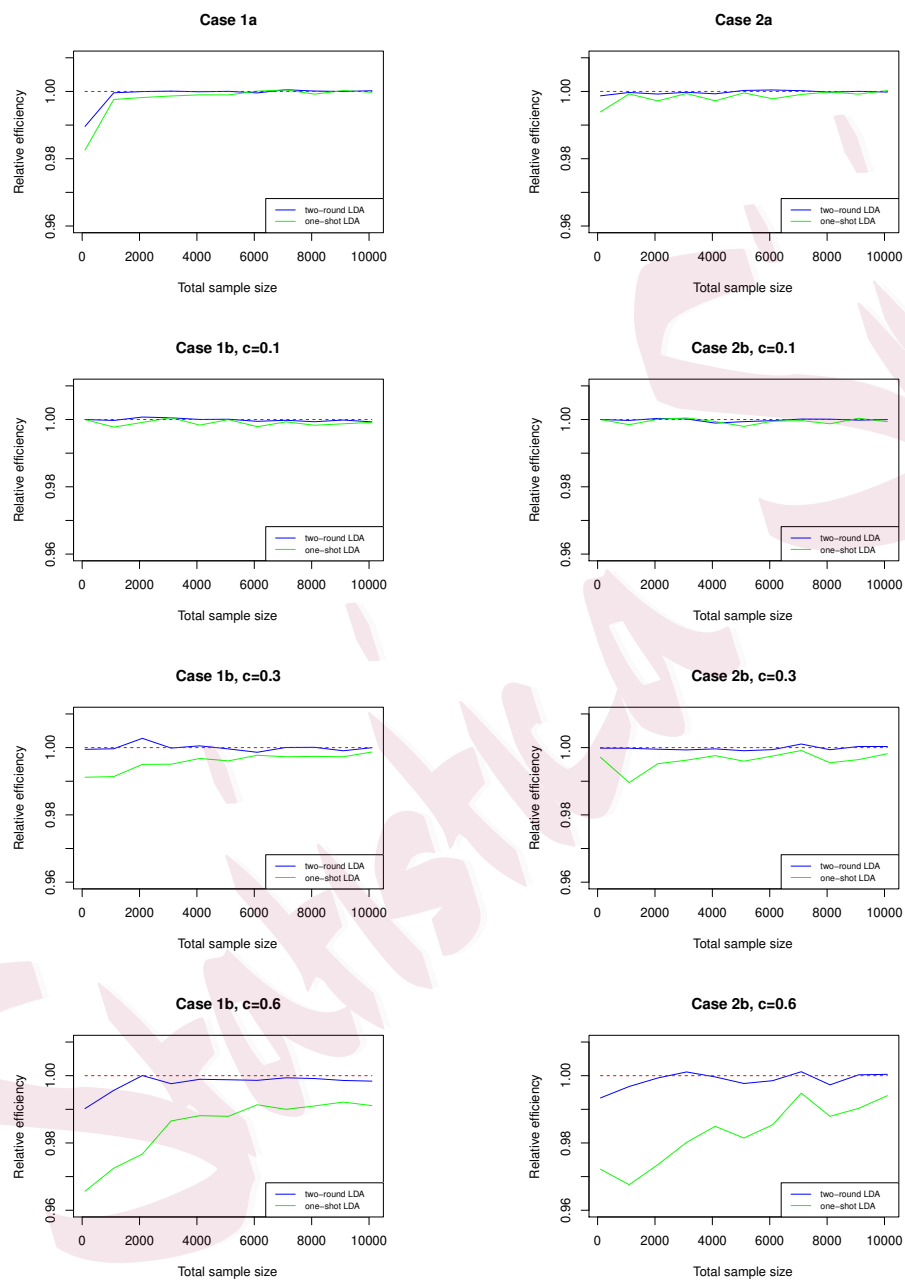


Figure 1: Relative efficiency of distributed LDA, where the red dotted line is the theoretical limits of relative efficiency for distributed LDA.

LDA. Using the technique of deterministic equivalents from random matrix theory, we show the relative efficiency compared to centralized LDA can reach 1, that is, the proposed distributed methods can achieve the same classification accuracy as the centralized case under suitable conditions. The numerical results support the theoretical findings. In the follow-up work, we could consider multi-class LDA to solve more general classification problems. Also, it is possible to relax the normality assumption on sample distributions. Cai and Liu (2011) considered classification accuracy for the elliptical distribution (Fang and Anderson, 1990, Chap. 1). Borrowing this idea, one can extend the results of this paper to the case of elliptical distribution, which will be the work in the future.

Supplementary Materials

The Supplementary Material contains the proofs of theoretical results stated throughout the main manuscript.

Acknowledgements

We appreciate the Co-Editor, the Associate Editor and two referees for their helpful comments and suggestions. Junlong Zhao acknowledges support by the National Science Foundation of China (No.11871104), and the

REFERENCES

Fundamental Research Funds for the Central Universities.

References

Bai, Z. and J. W. Silverstein (2010). *Spectral analysis of large dimensional random matrices*. Springer.

Banerjee, M., C. Durot, and B. Sen (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *Annals of Statistics* **47**(2), 720–757.

Battey, H., J. Fan, H. Liu, J. Lu, and Z. Zhu (2018). Distributed testing and estimation under sparse high dimensional models. *Annals of Statistics* **46**(3), 1352–1382.

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* **106**(496), 1566–1577.

Chang, X., S. Lin, and D. Zhou (2017). Distributed semi-supervised learning with kernel ridge regression. *Journal of Machine Learning Research* **18**(1), 1493–1514.

Chen, X., W. Liu, and Y. Zhang (2018). First-order newton-type estimator for distributed estimation and inference. *ArXiv preprint arXiv:1811.11368*.

Chen, X., W. Liu, and Y. Zhang (2019). Quantile regression under memory constraint. *Annals of Statistics* **47**(6), 3244–3273.

Chen, X. and M. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* **24**, 1655–1684.

Couillet, R. and M. Debbah (2011). *Random matrix methods for wireless communications*.

REFERENCES

- Cambridge University Press.
- Dobriban, E. and Y. Sheng (2018). Distributed linear regression by averaging. *ArXiv preprint arXiv:1810.00412*.
- Fan, J., Y. Guo, and K. Wang (2019). Communication-efficient accurate statistical estimation. *ArXiv preprint arXiv:1906.04870*.
- Fan, J., D. Wang, K. Wang, and Z. Zhu (2019). Distributed estimation of principal eigenspaces. *Annals of Statistics* **47**(6), 3009–3031.
- Fang, K. T. and T. W. Anderson (1990). *Statistical inference in elliptically contoured and related distributions*. Allerton Press.
- Garber, D., O. Shamir, and N. Srebro (2017). Communication-efficient algorithms for distributed stochastic principal component analysis. *ArXiv preprint arXiv:1702.08169*.
- Han, J. and Q. Liu (2016). Bootstrap model aggregation for distributed statistical learning. In *Advances in Neural Information Processing Systems*, pp. 1795–1803.
- Han, Y., P. Mukherjee, A. Ozgur, and T. Weissman (2018). Distributed statistical estimation of high-dimensional and nonparametric distributions. In *2018 IEEE International Symposium on Information Theory*, pp. 506–510. IEEE.
- Jordan, M. I., J. D. Lee, and Y. Yang (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* **114**(526), 668–681.
- Kleiner, A., A. Talwalkar, P. Sarkar, and M. I. Jordan (2014). A scalable bootstrap for massive

REFERENCES

- data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **76**(4), 795–816.
- Lalitha, A., A. Sarwate, and T. Javidi (2014). Social learning and distributed hypothesis testing. In *2014 IEEE International Symposium on Information Theory*, pp. 551–555. IEEE.
- Lee, J. D., Q. Liu, Y. Sun, and J. E. Taylor (2017). Communication-efficient sparse regression. *Journal of Machine Learning Research* **18**(1), 115–144.
- Macua, S. V., P. Belanovic, and S. Zazo (2011). Distributed linear discriminant analysis. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3288–3291. IEEE.
- Rosenblatt, J. D. and B. Nadler (2016). On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA* **5**(4), 379–404.
- Rubio, F. and X. Mestre (2011). Spectral convergence for a general class of random matrices. *Statistics & Probability Letters* **81**(5), 592–602.
- Shang, Z. and G. Cheng (2017). Computational limits of a distributed algorithm for smoothing spline. *Journal of Machine Learning Research* **18**(1), 3809–3845.
- Shi, C., W. Lu, and R. Song (2018). A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association* **113**(524), 1698–1709.
- Szabó, B. and H. Van Zanten (2019). An asymptotic analysis of distributed nonparametric methods. *Journal of Machine Learning Research* **20**, 1–30.

REFERENCES

- Tian, L. and Q. Gu (2017). Communication-efficient distributed sparse linear discriminant analysis. In *Artificial Intelligence and Statistics*, pp. 1178–1187.
- Volgushev, S., S.-K. Chao, and G. Cheng (2019). Distributed inference for quantile regression processes. *Annals of Statistics* **47**(3), 1634–1662.
- Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.
- Wang, X., Z. Yang, X. Chen, and W. Liu (2019). Distributed inference for linear support vector machine. *Journal of Machine Learning Research* **20**(113), 1–41.
- Xu, M., B. Lakshminarayanan, Y. W. Teh, J. Zhu, and B. Zhang (2014). Distributed bayesian posterior sampling via moment sharing. In *Advances in Neural Information Processing Systems*, pp. 3356–3364.
- Zhang, Y., J. Duchi, and M. J. Wainwright (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research* **16**(1), 3299–3340.

Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China

E-mail: (bnulmy@163.com)

School of Statistics, Beijing Normal University, Beijing 100875, China

E-mail: (zhaojunlong928@126.com)