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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

Communication-Efficient Distributed Linear Discriminant Analysis for Binary Classification

Mengyu Li[1,2] and Junlong Zhao[2*]

Renmin University of China, Beijing, China[1]

Beijing Normal University, Beijing, China[2]

Abstract: Large-scale data are common when the sample size n is large, and these data are often stored on k different local machines. Distributed statistical learning is an efficient way to deal with such data. In this study, we consider the binary classification problem for massive data based on a linear discriminant analysis (LDA) in a distributed learning framework. The classical centralized LDA requires the transmission of some p-by-p summary matrices to the hub, where p is the dimension of the variates under consideration. This can be a burden when p is large or the communication costs between the nodes are expensive. We consider two distributed LDA estimators, two-round and one-shot estimators, which are communication-efficient without transmitting p-by-p matrices. We study the asymptotic relative efficiency of distributed LDA estimators compared to a centralized LDA using 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 as that of the cen-

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

tralized estimator under mild conditions. Moreover, the two-round estimator can relax the restriction on k, allowing $kp/n \to c \in [0, 1)$ 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 inferences often need to deal with massive data. In many cases, the data are too large to be handled conveniently by a single data hub. Moreover, individual agents (e.g., local governments, hospitals, research labs) collect data independently and have communication constraints resulting from costs, privacy, ownership, security, and so on. Consequently, data have to be stored and processed on many local computers connected to a central server, thus forming a distributed system. In this way, researchers break large-scale computation problems into many small pieces, solve them using divide-and-conquer procedures, and then communicate only certain summary statistics. Distributed statistical inference has received considerable attention, covering topics including M-estimation (Chen and Xie, 2014; Rosenblatt and Nadler, 2016; Lee et al., 2017; Battey et al., 2018;

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

Shi, Lu, and Song, 2018; Jordan et al., 2018; Banerjee, Durot, and Sen, 2019; Fan, Guo, and Wang, 2019), hypothesis tests (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), Bayesian methods (Xu et al., 2014; Jordan et al., 2018), quantile regression (Volgushev, Chao, and Cheng, 2019; Chen, Liu, and Zhang, 2019), nonparametric regression (Chang, Lin, and Zhou, 2017; Shang and Cheng, 2017; Han et al., 2018; Szabó and Van Zanten, 2019), and bootstrap inference (Kleiner et al., 2014; Han and Liu, 2016), among others.

Linear discriminant analysis (LDA) is a classical method for classification in statistics, and implementing LDA in distributed systems has begun attracting the attention of researchers. Suppose that $\{(\boldsymbol{X}_i, C_i), 1 \leq i \leq n\}$ are independent and identically distributed ($i.i.d.$) observations, where $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})^\top$ 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; for simplicity, we assume that the size of the subsample for each machine is the same, denoted as $n^{(l)} \equiv n/k$, for $l = 1, \cdots, k$. For a distributed LDA estimator, one can consider its relative efficiency by comparing its classification

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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 a high-dimensional regime, where the dimension p can be much larger than the sample size n. To ensure that their distributed estimator attained 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 study, we focus on a distributed LDA for binary classification, under the setting of $p/n \to 0$, without the sparsity assumption on the parameters. When $p/n \to 0$, we show in Section 3 that a centralized LDA can still be effective under mild conditions. However, a 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, a two-round estimator and a one-shot estimator, according to their communication costs, without transmitting p-by-p summary matrices. We study the relative efficiency of each of these two estimators. It is shown that both estimators achieve the same efficiency

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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 $kp/n \to c \in [0, 1)$. When $c > 0$, the sample covariance matrix constructed from data on a local machine only is not a consistent estimator of the true covariance matrix, which brings challenges for the theoretical analysis. We successfully establish the efficiency of the two-round estimator using 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 their corresponding classification accuracies. Section 3 studies their relative efficiencies and derives their asymptotic properties. Section 4 provides numerical experiments that support the developed theory. In Section 5, we discuss our results, together with potential future directions.

Here, we summarize the notation used throughout the paper. We adopt the common convention of using boldface letters for vectors only, while a regular font is used for both scalars and matrices. For a vector $\boldsymbol{x} = (x_1, \cdots, x_p)^\top \in \mathbb{R}^p$ and $0 < q < \infty$, define the ℓ_q norm by

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

$\|\boldsymbol{x}\|_q = (\sum_{i=1}^p |x_i|^q)^{1/q}$. For a symmetric matrix $M \in \mathbb{R}^{p\times p}$, $\mathrm{tr}(M)$ denotes the trace of M, and $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ represent the maximal and the minimal eigenvalues, respectively. For a matrix $M \in \mathbb{R}^{n\times p}$, the nuclear norm is defined by $\|M\|_* = \mathrm{tr}[(M^\top M)^{1/2}] = \sum_{i=1}^{\min\{n,p\}} \sigma_i(M)$, 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 ith largest singular value. In addition, 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 $|a_n| \le C|b_n|$, for all $n \ge 1$, and write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. For two sequences of random variables $\{X_n\}$ and $\{Y_n\}$ and a random variable X, write $X_n \to_{a.s.} X$ if $\{X_n\}$ converges to X almost surely, and $X_n \to_p X$ 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 LDA

2.1 Centralized LDA in a distributed system

We focus on binary classification problems, assuming that the two classes follow normal distributions with the same covariance matrix, specifically, $N_p(\boldsymbol{\mu}_1, \Sigma)$ for class 1 and $N_p(\boldsymbol{\mu}_2, \Sigma)$ for class 2, where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ 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 $\Theta = \Sigma^{-1}$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

as the precision matrix (i.e., the inverse covariance matrix). For a new observation $\boldsymbol{H} \in \mathbb{R}^p$ with prior probabilities π_1 and π_2 from class 1 and class 2, respectively, Fisher's linear discriminant rule takes the form

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

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

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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

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

In particular, 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, π_1, and π_2 are unknown and can be estimated from the data. Suppose $\{\boldsymbol{X}_i : 1 \le i \le n_1\}$ and $\{\boldsymbol{Y}_i : 1 \le i \le n_2\}$ are *i.i.d.* observations from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$, respectively, where $n_1 + n_2 = n$. We do not impose sparsity assumptions on the parameters. The centralized estimators of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and Θ are

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

respectively, where

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

is the pooled sample covariance matrix, with $n = n_1 + n_2$. Then, π_1 and π_2 can be simply estimated by $\hat{\pi}_1 = n_1/n$ and $\hat{\pi}_2 = n_2/n$, respectively. Plugging these estimators into (2.1) yields the empirical version of $\psi(\boldsymbol{H})$, as follows:

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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

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

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

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

Thus, the classification accuracy of the centralized LDA is given by

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

When the data are stored on k machines, implementing a centralized LDA is still feasible, albeit with considerable communication costs. Let $\mathbb{X}^{(l)}$ be the data from class 1 stored on the lth local machine (i.e., the collection of $\boldsymbol{X}_i$ stored on the lth machine). Similarly, let $\mathbb{Y}^{(l)}$ be the data from class

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

2 stored on the lth machine, for $l = 1, \cdots, k$. For clarity, we denote

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

where $n_{1l} > 0$ and $n_{2l} > 0$ are the cardinalities of $\mathbb{X}^{(l)}$ and $\mathbb{Y}^{(l)}$, respectively. Thus, the lth 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

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

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

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

Because both $B_x^{(l)}$ and $\hat{\boldsymbol{\mu}}_1^{(l)}$ are computed locally from data on the lth machine, we see that B_x can be obtained by transmitting some summary matrices and vectors (i.e., $B_x^{(l)}$ and $\hat{\boldsymbol{\mu}}_1^{(l)}$) to the hub. Computing B_y similarly, one can obtain $\widehat{\Sigma}$, and consequently $\widehat{\Theta}$. Then, $\hat{\boldsymbol{\mu}}_a$ and $\hat{\boldsymbol{\mu}}_d$ can be computed in a similar fashion by transmitting $\hat{\boldsymbol{\mu}}_j^{(l)}$, for $j = 1, 2$. However, the centralized estimator requires the transmission of p-dimensional mean vectors $\hat{\boldsymbol{\mu}}_1^{(l)}$ and $\hat{\boldsymbol{\mu}}_2^{(l)}$, and p-by-p matrices $B_x^{(l)}$ and $B_y^{(l)}$, where $l = 1, \cdots, k$. When p and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

k are large, transmitting these p-by-p matrices to the central hub can be a burden in terms of communication, while transmitting p-dimensional mean vectors is much easier. In the following section, we propose two distributed estimators without transmitting these p-by-p matrices.

2.2 Distributed LDA by averaging

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

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

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 estimators. The first is called the two-round distributed LDA estimator, which estimates the mean vectors using the full data with two rounds of communication. The second is called the one-shot estimator, which esti-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

mates the means based on 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:

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

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

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

It is easy to see that $\hat{\Theta}_{two}^{(l)}$ can be obtained using data on the lth 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 the two-round distributed LDA as

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

As shown in the following Algorithm 1, $\bar{\psi}_{two}(\boldsymbol{H})$ can be computed in a communication-efficient way, with only the p-dimensional mean vectors being transmitted for two rounds. Comparing $\bar{\psi}_{two}(\boldsymbol{H})$ with its centralized counterpart $\hat{\psi}(\boldsymbol{H})$, one can see that the only difference between these two estimators lies in the different estimation of Θ. For the centralized estimator, Θ is estimated by $\widehat{\Theta} = \widehat{\Sigma}^{-1}$, with $\widehat{\Sigma}$ being obtained by transmitting the p-by-p matrices $B_x^{(l)}$ and $B_y^{(l)}$ to the hub.

There are two rounds of communication in Algorithm 1. First, the local estimators $\hat{\boldsymbol{\mu}}_1^{(l)}$ and $\hat{\boldsymbol{\mu}}_2^{(l)}$ are transmitted to the hub to compute $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, $\hat{\boldsymbol{\mu}}_a$,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

Algorithm 1 Two-Round Distributed LDA

Input: Observation $\boldsymbol{H}$ and data matrices $\mathbb{X}^{(l)}, \mathbb{Y}^{(l)}$ on the lth machine, for $l \in \{1, \cdots, k\}$.

1: Compute local sample means $\hat{\boldsymbol{\mu}}_1^{(l)}$ and $\hat{\boldsymbol{\mu}}_2^{(l)}$ on local machines, and then transmit them to the hub.

2: Compute on the hub the estimator $\hat{\boldsymbol{\mu}}_j$ by (2.8), for $j \in \{1, 2\}$, 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 $\widehat{\Theta}_{two}^{(l)}$ by (2.8), and obtain $V_l = \hat{\boldsymbol{\mu}}_a^\top \widehat{\Theta}_{two}^{(l)} \hat{\boldsymbol{\mu}}_d$ and $\boldsymbol{U}_l = \widehat{\Theta}_{two}^{(l)} \hat{\boldsymbol{\mu}}_d$ from the data on the lth machine, for $l \in \{1, \cdots, k\}$.

4: Send V_l and $\boldsymbol{U}_l$ to the hub, and compute their averages $\bar{\boldsymbol{U}} = n^{-1} \sum_{l=1}^k n^{(l)} \boldsymbol{U}_l$ and $\bar{V} = n^{-1} \sum_{l=1}^k n^{(l)} V_l$. Then, define the distributed LDA estimator $\bar{\psi}_{two}(\boldsymbol{H}) = \boldsymbol{H}^\top \bar{\boldsymbol{U}} - \bar{V}$.

5: **return** Classification result $\bar{\psi}_{two}(\boldsymbol{H})$

and $\hat{\boldsymbol{\mu}}_d$, and then the vector $(\hat{\boldsymbol{\mu}}_1^\top, \hat{\boldsymbol{\mu}}_2^\top, \hat{\boldsymbol{\mu}}_a^\top, \hat{\boldsymbol{\mu}}_d^\top)^\top \in \mathbb{R}^{4p}$ is broadcast to each local node. The second round sends $\boldsymbol{U}_l$ and V_l 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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

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

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

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

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

(2) When communication between nodes is prohibitively expensive, we consider a one-shot estimator, where only one round of communication is required. Denote the estimator of Θ from the data on the lth machine as

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

where

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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

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

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

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

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

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

Thus, the corresponding classification accuracy of the one-shot distributed

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

LDA is equal to

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

Define $\hat{R}_{one} = \hat{A}_{one}/\hat{A}_{cen}$ as the relative efficiency of the one-shot estimator. In 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 LDAs. Denote $\gamma_p = p/n$, $\gamma_p^{(l)} = p/n^{(l)}$, for $l = 1, \cdots, k$. For simplicity, we assume the data are evenly distributed to each machine; that is,

$$n_{11} = \cdots = n_{1k} = n_1/k, \quad n_{21} = \cdots = n_{2k} = n_2/k. \tag{3.1}$$

From (3.1), it follows that $n^{(l)} \equiv n/k$ and $\gamma_p^{(l)} \equiv kp/n = k\gamma_p$, for all l. Here, assumption (3.1) is assumed to reduce the complexity of the notation. The

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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

In this paper, we consider the case of $\gamma_p \to 0$, but $\gamma_p^{(l)} \to c \in [0, 1)$. For distributed estimators, when $c \neq 0$, the sample covariance matrix constructed from the data on the lth machine is not a consistent estimator of Σ. Consequently, $\widehat{\Theta}_{two}^{(l)}$ and $\widehat{\Theta}^{(l)}$ are not consistent estimators of Θ, which introduces challenges into 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 use the technique of deterministic equivalents (Couillet and Debbah, 2011, Chap. 6) to obtain the 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

$$\lim_{n\to\infty} |\mathrm{tr}\left[C_n \left(A_n - B_n\right)\right]| = 0,$$

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

Dobriban and Sheng (2018) studied the deterministic equivalent of the sample covariance matrix in elliptical models, which is a consequence of the generalized Marchenko–Pastur theorem (Rubio and Mestre, 2011). For the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

elliptical model, observations take the form $\{\boldsymbol{z}_i = g_i^{1/2}\Sigma^{1/2}\boldsymbol{u}_i, 1 \le i \le m\}$, where $\boldsymbol{u}_i \in \mathbb{R}^p$ is a vector with $i.i.d.$ entries, g_i is a datapoint-specific scale parameter allowing observations to have different scales, and $\Sigma \in \mathbb{R}^{p\times p}$ is the covariance matrix of $\boldsymbol{z}_i$. A special case of the elliptical model is $\boldsymbol{z}_i$ following a normal distribution, where we have $g_i = 1$. Arrange the samples as rows of a matrix Z, which has the form

$$Z = \Gamma^{1/2}U\Sigma^{1/2} \in \mathbb{R}^{m\times p}, \tag{3.2}$$

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

$$\widetilde{\Sigma}_*^{-1} \asymp e_p\Theta. \tag{3.3}$$

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

$$1 = \frac{1}{m}\,\mathrm{tr}\left[e_p\Gamma(I_m + \gamma_p e_p\Gamma)^{-1}\right]. \tag{3.4}$$

To study our problem, define the following pooled sample covariance

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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

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

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

(C1) Assume that (i) $0 < c_1 < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < c_2$, and (ii) $\gamma_p = p/n \to 0$, where c_1 and c_2 are constants independent of p.

(C2) As $n \to \infty$, $k = k_n$ satisfies the following conditions: (i) $n^{(l)} \equiv n/k \to \infty$; and (ii) $\gamma_p^{(l)} \equiv pk/n \to c \in [0, 1)$, for $l \in \{1, \cdots, k\}$.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

Proposition 1. *Under condition (C1), for the sample covariance matrix* $\widetilde{\Sigma}$ *with known* $\boldsymbol{\mu}_1$ *and* $\boldsymbol{\mu}_2$*, we have*

$$\widetilde{\Sigma}^{-1} \asymp \Theta. \tag{3.5}$$

Under condition (i) of (C1) and condition (C2), for the sample covariance matrix $\widetilde{\Sigma}^{(l)}$ *with known* $\boldsymbol{\mu}_1$ *and* $\boldsymbol{\mu}_2$ *on the* l*th machine, we have*

$$(\widetilde{\Sigma}^{(l)})^{-1} \asymp \frac{1}{1-\gamma_p^{(l)}}\Theta, \quad l = 1, \cdots, k. \tag{3.6}$$

In particular, if taking $c = 0$ *in (C2), we have* $(\widetilde{\Sigma}^{(l)})^{-1} \asymp \Theta$*, for* $l = 1, \cdots, k$*.*

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 the distributed LDA compared with that of the centralized case is the ratio of their classification accuracies. Then, the relative efficiency of the two-round distributed LDA is equal to

$$\hat{R}_{two} = \frac{n_1\Phi\left(-\frac{(\hat{\boldsymbol{\mu}}_a-\boldsymbol{\mu}_1)^\top\bar{\Theta}\hat{\boldsymbol{\mu}}_d+\log(n_2/n_1)}{\hat{\Delta}_p}\right) + n_2\Phi\left(\frac{(\hat{\boldsymbol{\mu}}_a-\boldsymbol{\mu}_2)^\top\bar{\Theta}\hat{\boldsymbol{\mu}}_d+\log(n_2/n_1)}{\hat{\Delta}_p}\right)}{n_1\Phi\left(-\frac{(\hat{\boldsymbol{\mu}}_a-\boldsymbol{\mu}_1)^\top\widehat{\Theta}\hat{\boldsymbol{\mu}}_d+\log(n_2/n_1)}{\hat{\Delta}_p}\right) + n_2\Phi\left(\frac{(\hat{\boldsymbol{\mu}}_a-\boldsymbol{\mu}_2)^\top\widehat{\Theta}\hat{\boldsymbol{\mu}}_d+\log(n_2/n_1)}{\hat{\Delta}_p}\right)}, \tag{3.7}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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

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

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

$$A_{two} = \pi_1 \Phi\left(\frac{\delta}{2} - \frac{(1-c)\log(\pi_2/\pi_1)}{\delta}\right) + \pi_2 \Phi\left(\frac{\delta}{2} + \frac{(1-c)\log(\pi_2/\pi_1)}{\delta}\right), \tag{3.8}$$

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

$$\begin{cases} R_{two} = 1, & c = 0; \\ R_{two} = 1, & c \in (0,1), \pi_1 = 1/2; \\ R_{two} < 1, & c \in (0,1), \pi_1 \neq 1/2. \end{cases} \tag{3.9}$$

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

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

According to (3.9), we discuss Theorem 1 in three cases: (1) $c = 0$; (2) $c \in (0, 1)$ and $\pi_1 = 1/2$; and (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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

the same efficiency as the centralized estimator. This coincides with our expectation, because $\widehat{\Sigma}_{two}^{(l)}$ is a good estimator of Σ when k is small. Case (2) is an interesting result, and is contrary to our expectation. When $c \in (0, 1)$, we see that k has the order n/p. By the well-known results of random matrices (Bai and Silverstein, 2010, Chap. 3), the local sample covariance matrix $\widehat{\Sigma}_{two}^{(l)}$ 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 $\widehat{\Sigma}_{two}^{(l)}$ 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 loses 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 $n \to \infty$, both $\hat{A}_{one}$ and $\hat{A}_{cen}$ converge to A_{cen} in probability. Consequently, $\hat{R}_{one}$ converges to one 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 achieves the same efficiency as the centralized one. We briefly discuss the difference between the two-round estimator and the one-shot one. The two-round estimator replaces the local sample means $\hat{\boldsymbol{\mu}}_i^{(l)}$ with global sample means $\hat{\boldsymbol{\mu}}_i$ using an extra round

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

of communication. This relaxes the restriction on k, allowing k to be of the same order as n/p. In particular, when $\pi_1 = 1/2$, the two-round estimator loses no information, even if k is 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 $\gamma_p^{(l)} \to c \in (0, 1)$, $\hat{\boldsymbol{\mu}}_i^{(l)}$ are no longer consistent estimators of $\boldsymbol{\mu}_i$ in terms of the ℓ_2 norm. In this case, from the proof of Theorem 2 (see Section S4 in the Supplementary Material), it is easy to see that $\hat{A}_{one}$ converges to the quantity

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

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

4. Simulations

In this section, we compare the performance of the distributed LDA methods with that of the centralized LDA. To begin with, we introduce the setup in the simulation study. The training data are generated as follows. We first withdraw *i.i.d.* observations of size n from normal distributions $N_p(\boldsymbol{\mu}_1, \Sigma)$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

(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 $N = 1000$, $\boldsymbol{\mu}_1 = (0, \cdots, 0)^\top \in \mathbb{R}^p$, and $\boldsymbol{\mu}_2 = (0.2, \cdots, 0.2)^\top \in \mathbb{R}^p$. 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 $1 \le i, j \le p$.

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

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 $\gamma_p = p/n \to 0$ and $\gamma_p^{(l)} = pk/n \to 0$. Then, we set $n \in \{100 + (i - 1) \times 10^3, i = 1, \cdots, 11\}$.

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 $\gamma_p = p/n \to 0$ and $\gamma_p^{(l)} = pk/n \to c$. Then, we let $n \in \{100 + (i - 1) \times 10^3, i = 1, \cdots, 11\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

Case 2a. Σ is from Example 2. The other settings are the same as those in *Case 1a*.

Case 2b. Σ is from Example 2. The other settings are the same as those in *Case 1b*.

For each case, we perform a distributed LDA and a centralized LDA on the training set to estimate the classification rule, and compute the relative efficiency based on the testing set. Then, we repeat the procedure 100 times to calculate the average relative efficiency. In Figure 1, we report the average values of $\hat{R}_{two}$ for the two-round distributed LDA, and those of $\hat{R}_{one}$ for the 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 quickly to one, coinciding with our 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 one, even when n is small, such as 100, and there is no significant difference between the 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 one for large n, but the performance of $\hat{R}_{one}$ is much worse than that of $\hat{R}_{two}$, especially when $c = 0.6$. This supports our theoretical findings.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

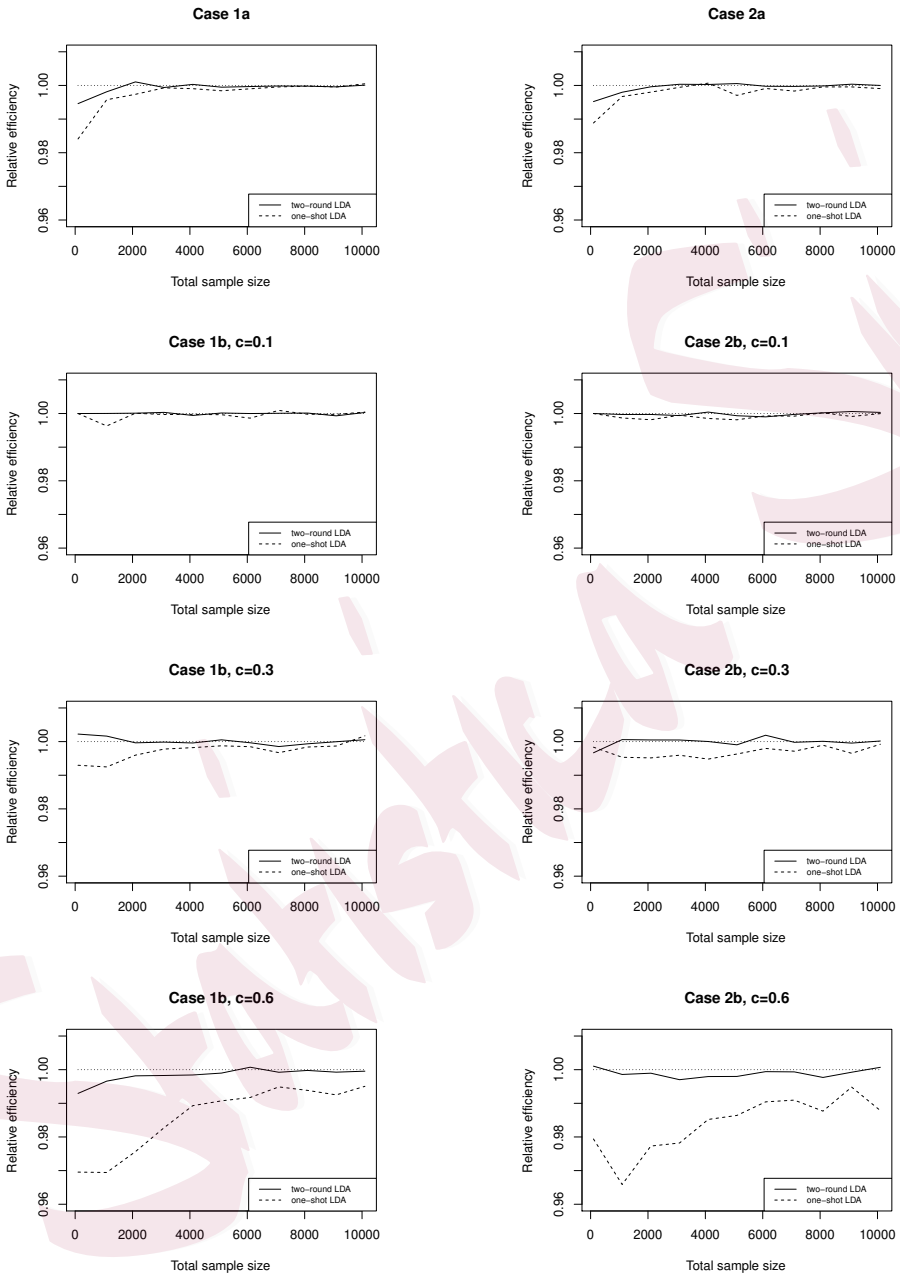

Figure 1: Relative efficiency of the distributed LDA, where the horizontal dotted line is the theoretical limit of the relative efficiency of the distributed LDA.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

5. Discussion

We have examined Fisher's LDA in distributed systems for binary classification, proposing two communication-efficient estimators. The classification accuracy is calculated for the distributed LDA. Using the technique of deterministic equivalents from random matrix theory, we show that the relative efficiency compared with that of centralized LDA can reach one; 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 future research, we will consider using a multi-class LDA to solve more general classification problems. In addition, it is possible to relax the normality assumption on the sample distributions. Cai and Liu (2011) considered the classification accuracy for the elliptical distribution (Fang and Anderson, 1990, Chap. 1). Borrowing from this idea, one can extend our results to the case of the elliptical distribution, which is also left to future work.

Supplementary Material

The online Supplementary Material contains proofs of the theoretical results stated within this paper.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

Acknowledgments

The authors thank the co-editor, associate editor and two referees for their helpful comments and suggestions. Junlong Zhao acknowledges the support of the National Science Foundation of China (No.11871104) and the 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*.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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*. 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. *Advances in Neural Information Processing Systems*, 1795–1803. Barcelona: Neural Information Processing Systems.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

Han, Y., P. Mukherjee, A. Ozgur, and T. Weissman (2018). Distributed statistical estimation of high-dimensional and nonparametric distributions. *2018 IEEE International Symposium on Information Theory*, 506–510. Vail, CO: Institute of Electrical and Electronics Engineers.

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 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. *2014 IEEE International Symposium on Information Theory*, 551–555. Honolulu, HI: Institute of Electrical and Electronics Engineers.

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. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3288–3291. Prague: Institute of Electrical and Electronics Engineers.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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.

Tian, L. and Q. Gu (2017). Communication-efficient distributed sparse linear discriminant analysis. *The 20th International Conference on Artificial Intelligence and Statistics*, 1178–1187. Fort Lauderdale, FL: Society for Artificial Intelligence and Statistics.

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. *Advances in Neural Information Processing Systems*, 3356–3364. Montreal, QC: Neural Information Processing Systems.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0374

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)