Statistica Sinica **32** (2022), 2455-2476
doi:https://doi.org/10.5705/ss.202021.0031

DISTRIBUTED SUFFICIENT DIMENSION REDUCTION FOR HETEROGENEOUS MASSIVE DATA

Kelin Xu, Liping Zhu and Jianqing Fan

*Fudan University, Renmin University of China
and Princeton University*

Abstract: We propose a distributed sufficient dimension reduction to process massive data characterized by high dimensionality, a huge sample size, and heterogeneity (heterogeneity, and huge sample sizes). To address the high dimensionality, we replace the high-dimensional explanatory variables with a small number of linear projections that are sufficient to explain the variabilities of the response variable. We allow for distinctive function maps for data scattered at different locations, thus addressing the problem of heterogeneity. We assume that the dimension reduction subspaces at different local nodes are identical. This allows us to aggregate the local results obtained from each local node to yield a final estimate on a central server. We explicitly examine the sliced inverse regression and cumulative slicing estimation, and investigate the nonasymptotic error bounds of the resulting dimensionality reduction. Our theoretical results are further supported by simulation studies and an application to meta-genome data from the American Gut Project.

Key words and phrases: Cumulative slicing estimation, distributed estimation, heterogeneity, sliced inverse regression, sufficient dimension reduction.

1. Introduction

In the current big data era, massive data are collected at different times and are usually scattered across locations. Such data are often characterized by high dimensionality, heterogeneity, and huge sample sizes (3H). For example, in biological studies, megabytes of genomics data are collected by large research institutes worldwide (Stephens et al. (2015)). Even a single genome sequence archive can consist of hundreds of thousands of samples and hundreds of millions of single nucleotide polymorphisms (SNPs) (BIG Data Center Members (2018)). For example, the American Gut Project consists of 29 batches and involves 11,336 human participants from 45 countries. There are 15,096 samples in total, each with 48,599 unique gene fragments mapped to 215 classes. In addition to the high dimensionality and huge sample sizes, the batch effects, distinctive popula-

Corresponding author: Liping Zhu, Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing, China. E-mail: zhu.liping@ruc.edu.cn.

tion structures, and potential problem of selection bias result in heterogeneity. It is difficult to manage heterogeneous massive data sets for many reasons, including memory and storage requirements, transmission capacity, privacy issues, and ethical concerns. Processing such data on a single computing device, poses significant challenges to conventional statistical analysis.

In the past two decades, many studies have analyzed massive distributed data with high dimensionality and huge sample sizes. In general, distributed algorithms assume there exists a star network architecture. We compute partial results on each local node and send them to a central server that summarizes these results to produce a final solution. In unsupervised learning, distributed algorithms are developed for principal component analyses to perform dimension reduction when massive data are scattered across different locations; see, for example, Kargupta et al. (2001), Qi, Wang and Birdwell (2004), Bai, Chan and Luk (2005) , and Liang et al. (2014). The theoretical properties of distributed algorithms for principal component analyses are studied thoroughly by Fan et al. (2019). In supervised learning, distributed algorithms typically assume that the response variable, denoted by $Y \in \mathbb{R}^1$, depends upon the p-vector of explanatory variables, denoted by $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$, through a parametric or even a linear model. Examples include Chen and Xie (2014), Lee et al. (2017), Battey et al. (2018), Jordan, Lee and Yang (2019), and Fan, Guo and Wang (2021). In these works, the dimension of $\mathbf{x}$ is reduced using the concept of sparsity; that is, only a small subset of $\mathbf{x}$ is truly important for predicting Y. To the best of our knowledge, very few existing distributed algorithms for unsupervised and supervised learning examine the problem of heterogeneity.

This study deals with massive distributed data that are simultaneously characterized by high dimensionality, heterogeneity, and a huge sample size. Suppose there are m local nodes in addition to a central server, and each local node contains n observations. The total sample size is thus $N \overset{\text{def}}{=} nm$. We denote the massive distributed observations by $\{(\mathbf{x}_{i,j}, Y_{i,j}), i = 1, \ldots, n, j = 1, \ldots, m\}$, where $\mathbf{x}_{i,j} \in \mathbb{R}^p$ is a p-vector of explanatory variables and $Y_{i,j}$ denotes the response. The subscript $_{i,j}$ stands for the ith observation scattered at the jth node. We allow very large values for p, n, and m, thus representing high dimensionality and a huge sample size. The goal of supervised learning is to understand how the response depends upon the explanatory variables, that is, to study how the conditional distribution of the response varies with the realizations of the explanatory variables. We assume that all observations are independent, but not necessarily identically distributed. Specifically, let $F(\cdot \mid \cdot)$ be the conditional distribution function. We assume that the observations at each local node are identically distributed, but the dependence structures at different nodes are allowed to be distinct; that is,

$$F_{i,j}(Y_{i,j} \mid \mathbf{x}_{i,j}) = F_j(Y_{i,j} \mid \mathbf{x}_{i,j}), \text{ for } \; i = 1, \ldots, n, \;\; j = 1, \ldots, m. \qquad (1.1)$$

The observations are homogeneous if all F_j are the same, say, $F_j = F_0$. Model (1.1) indicates the existence of heterogeneity across all m locations.

We propose a distributed sufficient dimension reduction for supervised learning to process heterogeneous massive distributed data. The dimension reduction replaces the high-dimensional explanatory variables, $\mathbf{x} \in \mathbb{R}^p$, with a small number of linear combinations, $\mathbf{B}^{\mathrm{T}}\mathbf{x} \in \mathbb{R}^{d_0}$, for $\mathbf{B} \in \mathbb{R}^{p \times d_0}$. We advocate using sufficient dimension reduction for at least two reasons. First, it does not impose parametric assumptions on F_j in Model (1.1), and, more importantly, these distribution functions are allowed to be distinct. Second, a dimension reduction with $\mathbf{x}_{i,j}$ replaced by $(\mathbf{B}^{\mathrm{T}}\mathbf{x}_{i,j})$ does not cause a loss of regression information in the sense of

$$F_j(Y_{i,j} \mid \mathbf{x}_{i,j}) = F_j(Y_{i,j} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}_{i,j}), \text{ for } \;\; i = 1, \ldots, n, \;\; j = 1, \ldots, m. \qquad (1.2)$$

An important implication of (1.2) is that a common basis $\mathbf{B}$ is shared with observations located at all m local nodes. Note that, if the basis matrices are distinct, say, $F_j(Y_{i,j} \mid \mathbf{x}_{i,j}) = F_j(Y_{i,j} \mid \mathbf{B}_j^{\mathrm{T}}\mathbf{x}_{i,j})$, we can define $\mathbf{B}$ to be a basis matrix that spans the column space of $(\mathbf{B}_1, \ldots, \mathbf{B}_m)$. In such a situation, (1.2) remains true. To ease the subsequent discussion, we assume there exists a mutual basis matrix $\mathbf{B}$. This assumption is not essential. However, it does improve the efficiency of estimating $\mathbf{B}$. If (1.2) holds, $\mathbf{x}_{i,j}$ and $(\mathbf{B}^{\mathrm{T}}\mathbf{x}_{i,j})$ are equivalent in terms of predicting $Y_{i,j}$. Note that (1.2) holds trivially if $d_0 = p$ and $\mathbf{B}$ is a full-rank matrix. Given the dimension reduction, $d_0 \ll p$. In many real-world applications, d_0 is very small, say, one, two, or at most three. The goal of sufficient dimension reduction is to seek a basis matrix $\mathbf{B}$ with the smallest column dimension. The column space of $\mathbf{B}$, if it exists, is referred to as the central subspace, and is denoted by $\mathcal{S}_{Y|\mathbf{x}}$ (Cook (1998)).

In this article, we propose distributed algorithms for two classic sufficient dimension reduction methods, the sliced inverse regression (Li (1991)) and cumulative slicing estimation (Zhu, Zhu and Feng (2010)), to process heterogeneous massive distributed data. A cumulative slicing estimation does not require a careful selection of the slice number, and is thus generally regarded as an improvement over the sliced inverse regression. Wang, Yu and Zhu (2021) demonstrated using empirical studies that, in high dimensions, the performance of the sliced inverse regression depends on the slice number. Thus, we advocate using a cumulative slicing estimation for massive data.

Many sufficient dimension reduction methods have been developed. These methods can be roughly classified into three categories. The first consists of inverse regression methods. These include, but are not limited to, the sliced inverse regression (Li (1991)), cumulative slicing estimation (Zhu, Zhu and Feng (2010)), sliced average variance estimation (Cook and Weisberg (1991)), and directional regression (Li and Wang (2007)). The second category mainly includes forward regression methods, such as the minimum average variance estimation (Xia et al. (2002)) and its variations (Xia (2007); Wang and Xia (2008)). The third category contains semiparametric approaches; see, for instance, Ma and Zhu (2012), Ma and Zhu (2013), and Ma and Zhu (2014). See also Cook (1998), Li (2018), and the references therein for comprehensive reviews on recent developments.

The asymptotic behaviors of existing sufficient dimension reduction methods are well understood when p is small relative to n and the observations are homogeneous. See Hsing and Carroll (1992), Zhu and Ng (1995), and Li and Zhu (2007) for the asymptotic results of inverse regression methods when p is fixed. The slicing estimation for a sliced inverse regression has been proven to be consistent when $p = o(n^{1/2})$ (Zhu, Miao and Peng (2006)) and $p = o(n)$ (Lin, Zhao and Liu (2018)). Wang, Yu and Zhu (2021) show that the cumulative slicing estimation is consistent if $p = o(n)$, or, more generally, $s = o(n)$ and $\log(p) = o(n)$, where s is the number of non-vanishing components. Existing asymptotic behaviors are investigated thoroughly for homogeneous observations. However, these cannot be carried over to heterogeneous massive distributed data without a careful adaption.

We focus on heterogeneous massive distributed data drawn independently from (1.1). In addition, we assume that (1.2) holds and aim to develop a distributed algorithm that aggregates the heterogeneous data across m local machines. Specifically, we propose performing a sliced inverse regression and a cumulative slicing estimation on each local node. With a slight abuse of notation, for now, we denote both estimates of $\mathbf{B}$ by $\widehat{\mathbf{B}}_j$, for $j = 1, \ldots, m$. We send all $\widehat{\mathbf{B}}_j$ to the central server to form

$$\widehat{\mathbf{T}} \overset{\text{def}}{=} m^{-1} \sum_{j=1}^{m} \widehat{\mathbf{B}}_j \widehat{\mathbf{B}}_j^{\mathrm{T}}. \tag{1.3}$$

We denote by $\widehat{\mathbf{B}}$ the top d_0 eigenvectors of $\widehat{\mathbf{T}}$, which is the average space in a least squares sense (Fan et al. (2019)). The minimum communication cost of this algorithm is $O(mpd_0)$.

We investigate the nonasymptotic error bound for

$$\mathrm{dist}(\widehat{\mathbf{B}}, \mathbf{B}) \overset{\mathrm{def}}{=} \|\widehat{\mathbf{B}}(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^{\mathrm{T}} - \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\|_F, \tag{1.4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. There are several distributed algorithms for both supervised and unsupervised learning, although few studies examine the theoretical properties of the resulting estimates obtained from the distributed algorithms. We assume that each node has the same sample size to simplify the presentation. When each local node has a different sample size, (1.3) is replaced by a weighted version.

We show from a theoretical perspective that analyzing distributed algorithms for sufficient dimension reduction is fundamentally different from doing so for a principal component analysis. This is because constructing an unbiased estimate for the covariance matrix in a principal component analysis is straightforward. However, how to find an unbiased estimate for the kernel matrices in sufficient dimension reduction remains unknown in the literature. Consequently, the problem of bias at each local node carries over to the aggregation procedure in the central node. Thus, we must quantify how the bias affects the resulting distributed estimation.

2. Distributed Estimation

The following notation is used throughout the paper. Suppose that, at the jth local node, $\mathbf{x}_j \in \mathbb{R}^p$ is a vector of explanatory variables and $Y_j \in \mathbb{R}^1$ is the associated response, for $j = 1, \ldots, m$. At each local node, n observations are collected and denoted by $\{(\mathbf{x}_{i,j}, Y_{i,j}), i = 1, \ldots, n, j = 1, \ldots, m\}$. Define $\mathbf{\Sigma}_j \overset{\mathrm{def}}{=} E\{\mathbf{x}_j - E(\mathbf{x}_j)\}\{\mathbf{x}_j - E(\mathbf{x}_j)\}^{\mathrm{T}}$, $\mathbf{M}_{j,s} \overset{\mathrm{def}}{=} \mathrm{cov}\{E(\mathbf{x}_j \mid Y_j)\}$ for a sliced inverse regression, and $\mathbf{M}_{j,c} \overset{\mathrm{def}}{=} E\{\mathbf{m}_{j,c}(Y_j)\mathbf{m}_{j,c}(Y_j)^{\mathrm{T}}\}$ for a cumulative slicing estimation, where $\mathbf{m}_{j,c}(y) \overset{\mathrm{def}}{=} \mathrm{cov}\{\mathbf{x}_j, I(Y_j \le y)\}$ and $I(A)$ is an indicator function, taking the value one if A is true, and zero otherwise. The subscripts s and c denote a sliced inverse regression and a cumulative slicing estimation, respectively. At the sample level, the usual moment estimates of $E(\mathbf{x}_j)$, $\mathbf{\Sigma}_j$, $\mathbf{m}_{j,c}(y)$, and $\mathbf{M}_{j,c}$ are defined, respectively, by

$$\overline{\mathbf{x}}_j \overset{\mathrm{def}}{=} n^{-1}\sum_{i=1}^{n}\mathbf{x}_{i,j}, \quad \widehat{\mathbf{\Sigma}}_j \overset{\mathrm{def}}{=} n^{-1}\sum_{i=1}^{n}\left(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_j\right)\left(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_j\right)^{\mathrm{T}}, \tag{2.1}$$

$$\widehat{\mathbf{m}}_{j,c}(y) \overset{\mathrm{def}}{=} n^{-1}\sum_{i=1}^{n}\left(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_j\right) I(Y_{i,j} \le y), \quad \widehat{\mathbf{M}}_{j,c} \overset{\mathrm{def}}{=} n^{-1}\sum_{i=1}^{n}\widehat{\mathbf{m}}_{j,c}(Y_{i,j})\widehat{\mathbf{m}}_{j,c}(Y_{i,j})^{\mathrm{T}}.$$

Li (1991) proposed a slicing procedure to estimate $\mathbf{M}_{j,s}$. Specifically, suppose that $q_{0,j}, q_{1,j}, \ldots, q_{H,j}$ is a sequence of cutting points, such that $-\infty = q_{0,j} < q_{1,j} < \cdots < q_{H-1,j} < q_{H,j} = \infty$. Define $I_{h,j} \overset{\text{def}}{=} (q_{h-1,j}, q_{h,j}]$ as the hth slice. Define $p_{h,j} \overset{\text{def}}{=} \text{pr}(Y_j \in I_{h,j})$ and $\mathbf{m}_{h,j,s} \overset{\text{def}}{=} \text{cov}\{\mathbf{x}_j, I(Y_j \in I_{h,j})\}$. Li (1991) suggested approximating $\mathbf{M}_{j,s}$ by

$$\mathbf{M}_{j,a} \overset{\text{def}}{=} \sum_{h=1}^{H} p_{h,j}^{-1} \mathbf{m}_{h,j,s} \mathbf{m}_{h,j,s}^{\mathrm{T}}.$$

The slicing estimates of $p_{h,j}$, $\mathbf{m}_{h,j,s}$, and $\mathbf{M}_{j,a}$ are given, respectively, by

$$\widehat{p}_{h,j} \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} I(Y_{i,j} \in I_{h,j}), \quad \widehat{\mathbf{m}}_{h,j,s} \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_j)\, I(Y_{i,j} \in I_{h,j}), \text{ and}$$

$$\widehat{\mathbf{M}}_{j,a} \overset{\text{def}}{=} \sum_{h=1}^{H} \widehat{p}_{h,j}^{-1} \widehat{\mathbf{m}}_{h,j,s} \widehat{\mathbf{m}}_{h,j,s}^{\mathrm{T}}. \tag{2.2}$$

Li (1991) suggested specifying $q_{h,j}$ as the $(h/H) \times 100\%$th quantile of Y_j. In this case, all $q_{h,j}$ and their corresponding sample counterparts $\widehat{q}_{h,j}$ slice the observations within each local node evenly. This usually facilitates implementing a sliced inverse regression.

The goal of sufficient dimension reduction is to seek a basis matrix $\mathbf{B}$ with a minimal column dimension d_0 such that (1.2) holds. In general, $\mathbf{B}$ is not identifiable. If $\mathbf{B}$ satisfies (1.2), then $\mathbf{BC}$ satisfies (1.2) as well, for an arbitrary nonsingular matrix $\mathbf{C} \in \mathbb{R}^{d_0 \times d_0}$. This allows us to assume that $\mathbf{B}$ is an orthogonal matrix such that $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}_{d_0 \times d_0}$, where $\mathbf{I}_{d_0 \times d_0} \in \mathbb{R}^{d_0 \times d_0}$ denotes an identity matrix. However, the column space of $\mathbf{B}$, defined as the central subspace and denoted by $\mathcal{S}_{Y|\mathbf{x}}$ throughout, is identifiable. At each local node, Li (1991) suggested recovering $\mathcal{S}_{Y|\mathbf{x}}$ using the column space of $\mathbf{\Sigma}_j^{-1}\mathbf{M}_{j,s}\mathbf{\Sigma}_j^{-1}$, denoted by $\text{span}(\mathbf{\Sigma}_j^{-1}\mathbf{M}_{j,s}\mathbf{\Sigma}_j^{-1})$, and Zhu, Zhu and Feng (2010) suggested recovering $\mathcal{S}_{Y|\mathbf{x}}$ using the column space of $\text{span}(\mathbf{\Sigma}_j^{-1}\mathbf{M}_{j,c}\mathbf{\Sigma}_j^{-1})$, for $j = 1\ldots, m$. For their suggestions to be valid, the linearity condition is required. That is, $E(\mathbf{x}_j \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}_j)$ is a linear function of $\mathbf{x}_j$, which is satisfied when $\mathbf{x}$ follows a normal or, more generally, an elliptically contoured distribution. Hall and Li (1993) showed that this linearity condition holds asymptotically, as long as p is sufficiently large and d_0 is relatively small. Therefore, this linearity condition is typically regarded as mild (Li (1991)).

It can be clearly seen from (2.1) and (2.2) that $\widehat{\mathbf{M}}_{j,a}$ and $\widehat{\mathbf{M}}_{j,c}$ are similar.

To facilitate the subsequent discussion, we use $\widehat{\mathbf{M}}_j$ to denote either $\widehat{\mathbf{M}}_{j,a}$ or $\widehat{\mathbf{M}}_{j,c}$ unless stated otherwise. Define $\mathbf{\Omega}_j \overset{\text{def}}{=} \mathbf{\Sigma}_j^{-1}\mathbf{M}_j\mathbf{\Sigma}_j^{-1}$ and $\widehat{\mathbf{\Omega}}_j$ is the estimation of $\mathbf{\Omega}_j$, where $\mathbf{M}_j$ can be $\mathbf{M}_{j,c}$ or $\mathbf{M}_{j,s}$. This allows us to propose distributed algorithms and investigate the nonasymptotic error bounds for both a sliced inverse regression and a cumulative slicing estimation within a unified framework. The following distributed algorithms and theoretical results apply to both sufficient dimension reduction methods.

2.1. Case I: All sample covariance matrices are invertible

Here, we introduce the distributed algorithms. We first assume all the sample covariance matrices, namely, $\widehat{\mathbf{\Sigma}}_j$, are invertible. This implicitly requires that p be much smaller than n at each local node. In this case, by invoking the theory of sufficient dimension reduction, we can simply recover $\mathcal{S}_{Y|\mathbf{x}}$ from the column space of $\widehat{\mathbf{\Omega}}_{a1,j} \overset{\text{def}}{=} \widehat{\mathbf{\Sigma}}_j^{-1}\widehat{\mathbf{M}}_j\widehat{\mathbf{\Sigma}}_j^{-1}$. The subscript $_{a1,j}$ represents the estimation obtained from the jth node by the distributed Algorithm 1, discussed below. Because all local nodes share an identical central subspace $\mathcal{S}_{Y|\mathbf{x}}$, combining the estimates at all local nodes improves the efficiency of estimating $\mathcal{S}_{Y|\mathbf{x}}$. There are two ways to achieve this goal. The first is that we pass all $\widehat{\mathbf{\Omega}}_{a1,j}$ to the central server to form

$$m^{-1}\sum_{j=1}^{m}\widehat{\mathbf{\Omega}}_{a1,j}.$$

We then apply a singular value decomposition to the above average to obtain the top d_0 eigenvectors. The communication cost of this option is of order $O(mp^2)$. The second option is that at each local node, we apply a singular value decomposition to $\widehat{\mathbf{\Omega}}_{a1,j}$ to obtain the top d_0 eigenvectors, which is denoted by $\widehat{\mathbf{B}}_{a1,j}$. Next, we pass $\widehat{\mathbf{B}}_{a1,j}$ to the central server to form $\widehat{\mathbf{T}}_{a1}$, defined in (1.3). We further apply a singular value decomposition to $\widehat{\mathbf{T}}_{a1}$ to obtain the first top d_0 eigenvectors, which are denoted by $\widehat{\mathbf{B}}_{a1}$. The communication cost of this option is of order $O(mpd_0)$, which is smaller than that of the first option. We advocate using the second option because the reduction of the communication cost is substantial if d_0 is much less than p, which benefits from sufficient dimension reduction.

The distributed Algorithm 1 is as follows.

2.2. Case II: Not all sample covariance matrices are invertible

If the sample covariance matrices are not all invertible, we have to avoid using $\mathbf{\Sigma}_j^{-1}$ directly. To address this issue, Tan et al. (2018) introduce a convex formulation to fit a sparse sliced inverse regression, which is solved using a lin-

Algorithm 1.

1. Estimate $\widehat{\boldsymbol{\Omega}}_{a1,j} \overset{\text{def}}{=} \widehat{\boldsymbol{\Sigma}}_j^{-1}\widehat{\mathbf{M}}_j\widehat{\boldsymbol{\Sigma}}_j^{-1}$ at the jth local node. This amounts to estimating $\widehat{\boldsymbol{\Sigma}}_j$, $\widehat{\mathbf{M}}_{j,c}$, and $\widehat{\mathbf{M}}_{j,a}$ using (2.1) and (2.2).
2. Apply a singular value decomposition to $\widehat{\boldsymbol{\Omega}}_{a1,j}$ at the jth local node. The top d_0 eigenvectors are denoted by $\widehat{\mathbf{B}}_{a1,j}$.
3. Pass all $\{\widehat{\mathbf{B}}_{a1,j}\}$ to the central server to form $\widehat{\mathbf{T}}_{a1}$, defined in (1.3), with $\widehat{\mathbf{B}}_j$ replaced by $\widehat{\mathbf{B}}_{a1,j}$.
4. Apply a singular value decomposition to $\widehat{\mathbf{T}}_{a1}$ to obtain the first d_0 top eigenvectors, which are denoted by $\widehat{\mathbf{B}}_{a1}$.

earized alternating direction method of multipliers algorithm. This algorithm is further improved by Tan, Shi and Yu (2020), although it is methodology specific. Motivated by Wang, Jiang and Zhu (2021), at the jth local node, we seek a matrix $\boldsymbol{\Phi}_j \in \mathbb{R}^{p\times p}$ that is the closest to $\boldsymbol{\Omega}_j$, subject to the sparsity constraints or its relaxations. This is a very general methodology. More importantly, it is computationally very efficient. Specifically, we propose approximating $\boldsymbol{\Omega}_j$ under the following criterion:

$$\text{trace}[\{(\boldsymbol{\Phi}_j-\boldsymbol{\Omega}_j)\boldsymbol{\Sigma}_j\}^2] = \text{trace}\{(\boldsymbol{\Phi}_j\boldsymbol{\Sigma}_j)^2\} - \text{trace}(2\boldsymbol{\Phi}_j\mathbf{M}_j) + \text{trace}(\boldsymbol{\Sigma}_j^{-1}\mathbf{M}_j)^2,$$

where trace$(\cdot)$ is the trace of a matrix. Because the last quantity on the right-hand side of the above display is irrelevant to the unknown parameter $\boldsymbol{\Phi}_j$, at the sample level, we consider minimizing

$$\widehat{\boldsymbol{\Omega}}_{a2,j} \overset{\text{def}}{=} \underset{\boldsymbol{\Phi}_j}{\text{argmin}}\left[\text{trace}\{(\boldsymbol{\Phi}_j\widehat{\boldsymbol{\Sigma}}_j)^2\} - \text{trace}(2\boldsymbol{\Phi}_j\widehat{\mathbf{M}}_j) + \lambda_{n,j}\|\boldsymbol{\Phi}_j\|_1\right], \qquad (2.3)$$

where the tuning parameter $\lambda_{n,j}$ is typically decided using ten-fold cross-validation, the subscript $_{a2,j}$ represents the estimation obtained from the jth node by the distributed Algorithm 2, and

$$\|\boldsymbol{\Phi}_j\|_1 \overset{\text{def}}{=} \sum_{k=1}^{p}\sum_{l=1}^{p}|\Phi_{k,l}|, \text{ where } \Phi_{k,l} \text{ denotes the } (k,l)\text{th entry of } \boldsymbol{\Phi}_j.$$

We use the alternating direction method of multipliers (Boyd et al. (2010)) to solve the optimization problem (2.3), which yields $\widehat{\boldsymbol{\Omega}}_{a2,j}$. Interested readers may refer to Wang, Jiang and Zhu (2021) on how to solve the minimization problem (2.3) at the j-local node. In (2.3), we use $\widehat{\boldsymbol{\Sigma}}_j$ instead of its inversion $\widehat{\boldsymbol{\Sigma}}_j^{-1}$.

Algorithm 2.

```
1. Estimate $\boldsymbol{\Omega}_j$ at the $j$th local node using (2.3) to obtain $\widehat{\boldsymbol{\Omega}}_{a2,j}$.
2. Apply a singular value decomposition to $\widehat{\boldsymbol{\Omega}}_{a2,j}$ at the $j$th local node to obtain its top $d_0$ eigenvectors $\widehat{\mathbf{B}}_{a2,j}$.
3. Pass all $\widehat{\mathbf{B}}_{a2,j}$ to the central server to form $\widehat{\mathbf{T}}_{a2}$, defined in (1.3), with $\widehat{\mathbf{B}}_j$ replaced by $\widehat{\mathbf{B}}_{a2,j}$.
4. Apply a singular value decomposition to $\widehat{\mathbf{T}}_{a2}$ to obtain the first $d_0$ top eigenvectors, which are denoted by $\widehat{\mathbf{B}}_{a2}$.
```

Therefore, it can be readily used, even when p is much greater than n. With $\widehat{\boldsymbol{\Omega}}_{a2,j}$ obtained from (2.3), the second through fourth steps of the distributed Algorithm 2 are, in spirit, the same as those described in Section 2.1. We denote by $\widehat{\mathbf{B}}_{a2}$ the final solution of this distributed algorithm.

2.3. The nonasymptotic error bounds

Next, we investigate the nonasymptotic error bounds for the two distributed estimates $\widehat{\mathbf{B}}_{a1}$ and $\widehat{\mathbf{B}}_{a2}$. To this end, we define the projection matrix of $\mathbf{B}$ as $\mathbf{P}(\mathbf{B}) \overset{\text{def}}{=} \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$. Similarly, $\mathbf{P}(\widehat{\mathbf{B}}) \overset{\text{def}}{=} \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^{\mathrm{T}}$, where $\widehat{\mathbf{B}}$ can be either $\widehat{\mathbf{B}}_{a1}$ or $\widehat{\mathbf{B}}_{a2}$. To quantify the accuracy of $\widehat{\mathbf{B}}$, we define

$$\{\operatorname{dist}(\widehat{\mathbf{B}},\mathbf{B})\}^2 \overset{\text{def}}{=} \operatorname{trace}[\{\mathbf{P}(\widehat{\mathbf{B}})-\mathbf{P}(\mathbf{B})\}^2] \ \text{ and } r^2(d_0) \overset{\text{def}}{=} \frac{\operatorname{trace}\{\mathbf{P}(\widehat{\mathbf{B}})\mathbf{P}(\mathbf{B})\}}{d_0}.$$

It is easy to show that $\{\operatorname{dist}(\widehat{\mathbf{B}},\mathbf{B})\}^2 = 2d_0\{1-r^2(d_0)\}$. In other words, these two metrics are equivalent in terms of measuring the accuracy of $\widehat{\mathbf{B}}$. In addition, $r^2(d_0)$ increases from zero to one, as $\operatorname{dist}(\widehat{\mathbf{B}},\mathbf{B})$ decreases from $(2d_0)^{1/2}$ to zero. In particular, if $\widehat{\mathbf{B}}$ is a poor estimate of $\mathbf{B}$, $r^2(d_0)$ is small and $\operatorname{dist}(\widehat{\mathbf{B}},\mathbf{B})$ is large. Note that $\operatorname{dist}(\widehat{\mathbf{B}},\mathbf{B}) = \|\mathbf{P}(\widehat{\mathbf{B}})-\mathbf{P}(\mathbf{B})\|_F$, where $\|\cdot\|_F$ is the Frobenius norm.

Following Vershynin (2018), we define the ψ_1-norm and ψ_2-norm of a random variable X by

$$\|X\|_{\psi_1} \overset{\text{def}}{=} \sup_{k\geq 1}\frac{(E|X|^k)^{1/k}}{k} \text{ and } \|X\|_{\psi_2} \overset{\text{def}}{=} \sup_{k\geq 1}\frac{(E|X|^k)^{1/k}}{k^{1/2}},$$

respectively. A random vector $\mathbf{x}\in\mathbb{R}^p$ is said to be sub-Gaussian if there exists a positive constant C such that $\|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}\|_{\psi_2} \leq C\{E(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})^2\}^{1/2}$, for all $\boldsymbol{\alpha}\in\mathbb{R}^p$. Throughout, C, c, C_1, c_1, $C_2, c_2, \ldots$ denote generic constants that may vary at each appearance. We assume the following conditions.

(C1) The explanatory variables $\{\mathbf{x}_{i,j} : i = 1, \ldots, n; j = 1, \ldots, m\}$ are all sub-Gaussian.

(C2) There exists a positive constant c such that

$$c^{-1} \leq \inf_{j=1,\ldots,m} \{\lambda_{\min}(\mathbf{\Sigma}_j)\} \leq \sup_{j=1,\ldots,m} \{\lambda_{\max}(\mathbf{\Sigma}_j)\} \leq c,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues, respectively, of $\mathbf{\Sigma}_j$, for $j = 1, \ldots, m$.

(C3) The smallest nonzero eigenvalues of $\mathbf{M}_j$, for $j = 1, \ldots, m$, are uniformly bounded away from zero.

Define

$$\mathbf{\Omega}^* \overset{\text{def}}{=} m^{-1} \sum_{j=1}^{m} E\{\mathbf{P}(\widehat{\mathbf{B}}_j)\},$$

which can be $\mathbf{\Omega}^*_{a1}$ if $\widehat{\mathbf{B}}_j = \widehat{\mathbf{B}}_{a1,j}$ and $\mathbf{\Omega}^*_{a2}$ if $\widehat{\mathbf{B}}_j = \widehat{\mathbf{B}}_{a2,j}$. Denote the top d_0 eigenvectors of $\mathbf{\Omega}^*$ by $\mathbf{B}^*$, which can be either $\mathbf{B}^*_{a1}$ or $\mathbf{B}^*_{a2}$. By the triangle inequality, we have

$$\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}) \leq \text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*) + \text{dist}(\mathbf{B}^*, \mathbf{B}).$$

Note that $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ and $\text{dist}(\mathbf{B}^*, \mathbf{B})$ correspond to the variance and the bias of $\widehat{\mathbf{B}}$, respectively, when estimating $\mathbf{B}$. In what follows, we quantify the accuracy of $\widehat{\mathbf{B}}$, which can be either $\widehat{\mathbf{B}}_{a1}$ or $\widehat{\mathbf{B}}_{a2}$, when estimating $\mathbf{B}$ using $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ and $\text{dist}(\mathbf{B}^*, \mathbf{B})$.

We study the nonasymptotic error bound of $\widehat{\mathbf{B}}_{a1}$ first. Lemmas 1 and 2 give the orders of $\text{dist}(\widehat{\mathbf{B}}_{a1}, \mathbf{B}^*_{a1})$ and $\text{dist}(\mathbf{B}^*_{a1}, \mathbf{B})$, respectively.

Lemma 1. *In addition to* (C1)–(C3), *we assume there exists* $C > 0$ *such that* $n \geq 2C^2 d_0 p$. *Then,* $\|dist(\widehat{\mathbf{B}}_{a1}, \mathbf{B}^*_{a1})\|_{\psi_1} \leq C_1 (d_0 p/N)^{1/2}$, *for* $C_1 > 0$.

Lemma 2. *Assume Conditions* (C1)–(C3) *hold. Then, there exists* $C_2 > 0$ *such that* $dist(\mathbf{B}^*_{a1}, \mathbf{B}) \leq C_2 d_0^{1/2} p/n$.

Here, $\widehat{\mathbf{\Sigma}}_j^{-1}$ and $\widehat{\mathbf{M}}_j$ are the respective biased estimates of $\mathbf{\Sigma}_j^{-1}$ and $\mathbf{M}_j$. Consequently, $\widehat{\mathbf{\Omega}}_{a1,j}$ and its eigenvectors are biased, with the magnitude determined by the local sample size n. These biases are carried over to the central server, and do not necessarily diminish, even when the total sample size N diverges to infinity. If the sample size n at the local nodes is sufficiently large, such that $n \geq mp(C_2^2/C_1^2)$, the bias $\text{dist}(\mathbf{B}^*_{a1}, \mathbf{B})$ is smaller than $\|\text{dist}(\widehat{\mathbf{B}}_{a1}, \mathbf{B}^*_{a1})\|_{\psi_1}$, and thus negligible. In other words, the bias issue of this distributed algorithm is not critical if the sample size n is sufficiently large. However, if m is very large, the bias term $\mathrm{dist}(\mathbf{B}_{a1}^*, \mathbf{B})$ plays a dominant role. This also makes the distributed algorithms for a sufficient dimension reduction quite different from those for a principal component analysis. In particular, we can construct an unbiased estimate for the covariance matrix in a principal component analysis. However, how to construct an unbiased estimate for $\boldsymbol{\Omega}_j \overset{\text{def}}{=} \boldsymbol{\Sigma}_j^{-1}\mathbf{M}_j\boldsymbol{\Sigma}_j^{-1}$ in a sufficient dimension reduction remians unknown. Consequently, the bias issue at each local node carries over to the aggregation procedure in the central node.

Theorem 1 is an immediate result of the above two lemmas.

Theorem 1. *Under the conditions of Lemma* 1, *we have*

$$\|\mathit{dist}(\widehat{\mathbf{B}}_{a1}, \mathbf{B})\|_{\psi_1} \le C\left\{\left(\frac{d_0 p}{N}\right)^{1/2} + d_0^{1/2}\left(\frac{p}{n}\right)\right\}.$$

Here, the error bound of $\widehat{\mathbf{B}}_{a1}$ is minimized when $m = O(n/p)$. In other words, if we are able to distribute all N observations to m local nodes, each of size n, an optimal m is of order (n/p). In many real-world applications (e.g., the American Gut Project), the number of local nodes, m, is relatively small, and the sample size at each local node, n, is sufficiently large. In such applications, the bias term, $\mathrm{dist}(\mathbf{B}_{a1}^*, \mathbf{B})$, is usually negligible when compared to $\|\mathrm{dist}(\widehat{\mathbf{B}}_{a1}, \mathbf{B}_{a1}^*)\|_{\psi_1}$.

Next, we study the nonasymptotic error bound of the distributed estimate $\widehat{\mathbf{B}}_{a2}$. To simplify the subsequent discussion, we introduce the following notation. We define $\mathcal{S}_j \overset{\text{def}}{=} \{(k,l) : \text{the } (k,l)\text{th entry of } \boldsymbol{\Omega}_j \text{ is nonzero}\}$. We denote by s_j the cardinality of $\mathcal{S}_j$. Let $\boldsymbol{\Gamma}_j \overset{\text{def}}{=} \boldsymbol{\Sigma}_j \otimes \boldsymbol{\Sigma}_j$. Then, $\boldsymbol{\Gamma}_{\mathcal{S}_j^c,\mathcal{S}_j,j}$ and $\boldsymbol{\Gamma}_{\mathcal{S}_j,\mathcal{S}_j,j}$ are sub-matrices of $\boldsymbol{\Gamma}_j$ indexed by $(\mathcal{S}_j^c, \mathcal{S}_j)$ and $(\mathcal{S}_j, \mathcal{S}_j)$, respectively. For a matrix $\mathbf{A} = (a_{kl})_{p\times p}$, we define

$$\|\mathbf{A}\|_\infty \overset{\text{def}}{=} \max_{1\le k\le p}\sum_{l=1}^p |a_{kl}|,\ D_j \overset{\text{def}}{=} \|\boldsymbol{\Gamma}_{\mathcal{S}_j,\mathcal{S}_j,j}^{-1}\|_\infty \text{ and } \kappa_j \overset{\text{def}}{=} 1 - \|\boldsymbol{\Gamma}_{\mathcal{S}_j^c,\mathcal{S}_j,j}\boldsymbol{\Gamma}_{\mathcal{S}_j,\mathcal{S}_j,j}^{-1}\|_\infty.$$

Lemma 3. *In addition to* (C1)–(C3), *we assume that* $\kappa_j > 0$ *and* $s_j\{\log(p)/n\}^{1/2} \to 0$, *for* $1 \le j \le m$. *Suppose there exist* $C_3 > 0$ *and* $c_2 > 0$ *such that*

$$n \ge c_2 \max_{1\le j\le m} (\kappa_j^{-1}D_j)^2 d_0 p.$$

Then, it follows that

$$\|dist(\widehat{\mathbf{B}}_{a2}, \mathbf{B}^*_{a2})\|_{\psi_1} \le C_3 \max_{1\le j\le m} (\kappa_j^{-1} D_j s_j^{1/2}) \left\{ \frac{d_0 \log(p)}{N} \right\}^{1/2}.$$

Lemma 4. *In addition to* (C1)–(C3), *we assume that* $\kappa_j > 0$ *and* $s_j\{\log(p)/n\}^{1/2} \to 0$, *for* $1 \le j \le m$. *Then, there exists* $C_4 > 0$ *such that*

$$dist(\mathbf{B}^*_{a2}, \mathbf{B}) \le C_4 \max_{1\le j\le m} (\kappa_j^{-1} D_j s_j^{1/2}) \left\{ \frac{\log(p)}{n} \right\}^{1/2}.$$

Lemma 4 indicates that the dimension p has a very small effect on $\mathrm{dist}(\mathbf{B}^*_{a2}, \mathbf{B})$, in the order of $\{\log(p)\}^{1/2}$. In contrast, the number of truly important explanatory variables, s_j, plays a much more important role than p in both $\mathrm{dist}(\widehat{\mathbf{B}}_{a2}, \mathbf{B}^*_{a2})$ and $\mathrm{dist}(\mathbf{B}^*_{a2}, \mathbf{B})$.

The above two lemmas lead to Theorem 2.

Theorem 2. *Under the conditions of Lemmas* 3 *and* 4, *we have*

$$\|dist(\widehat{\mathbf{B}}_{a2}, \mathbf{B})\|_{\psi_1} \le C \Bigg[\max_{1\le j\le m} (\kappa_j^{-1} D_j s_j^{1/2}) \left\{ \frac{d_0 \log(p)}{N} \right\}^{1/2}, \tag{2.4}$$

$$+ \max_{1\le j\le m} (\kappa_j^{-1} D_j s_j^{1/2}) \left\{ \frac{\log(p)}{n} \right\}^{1/2} \Bigg]. \tag{2.5}$$

In the distributed Algorithm 2, the bias term $\mathrm{dist}(\mathbf{B}^*_{a2}, \mathbf{B})$ is magnified to the order of $\{\log(p)/n\}^{1/2}$, which usually dominates $\|\mathrm{dist}(\widehat{\mathbf{B}}_{a2}, \mathbf{B}^*_{a2})\|_{\psi_1}$, because N is usually much larger than n and d_0 is often a small number.

3. Numerical Examples

We illustrate the performance of the distributed estimates using synthetic examples. Throughout, we fix $p = 200$, draw $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma} = (\rho^{|k-l|})_{p\times p}$, and generate the error term ε from a standard normal distribution. Set $\boldsymbol{\beta}_1 = (1, 1, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$ and $\boldsymbol{\beta}_2 = (0, 0, 1, 1, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$. Let $m = \{2^5, 2^6, 2^7, 2^8\}$ and $N = \{2^9 p, 2^{10} p, 2^{11} p, 2^{12} p\}$. All N observations are scattered uniformly across m nodes, each of size n.

We consider three examples.

Example 1. We set $\rho = 0.5$ in $\mathbf{\Sigma} = (\rho^{|k-l|})_{p\times p}$. At each local node, we generate Y randomly from the following models with equal probability 1/3. Thus, the observations at different local nodes are heterogeneous.

$$Y = \frac{1+2\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}}{0.5+(1.5+\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})^2}+\varepsilon, \tag{3.1}$$

$$Y = \sin(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})+\exp(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})+\varepsilon, \tag{3.2}$$

$$Y = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})\exp(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})+\varepsilon, \tag{3.3}$$

We implement Algorithm 1 for the sliced inverse regression and cumulative slicing estimation. We fix the slice number to $H = 10$ in the sliced inverse regression.

Example 2. We set $\rho = 0.8$ in $\boldsymbol{\Sigma} = (\rho^{|k-l|})_{p\times p}$. The response Y is generated in the same way as in Example 1. We implement Algorithm 2 for both the sliced inverse regression and the cumulative slicing estimation as well.

Note that $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathbb{R}^{p\times 2}$ in the above examples.

Example 3. We set $\rho = 0.5$ in $\boldsymbol{\Sigma} = (\rho^{|k-l|})_{p\times p}$. At each local node, we generate Y from the following two models with equal probability 1/2:

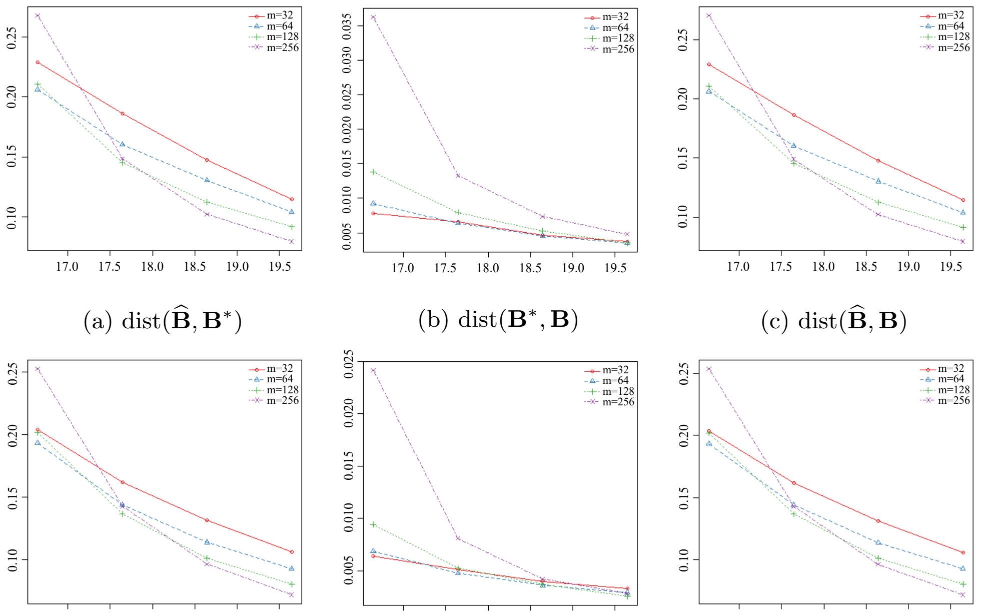
$$Y = \sin\left\{\frac{(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)^{\mathrm{T}}\mathbf{x}}{2}\right\}+\varepsilon, \tag{3.4}$$

$$Y = \sin\left\{\frac{(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2)^{\mathrm{T}}\mathbf{x}}{2}+\frac{(k+3)\pi}{8}\right\}+\varepsilon. \tag{3.5}$$

We vary the value of k from $\{1, 2, 3, 4, 5\}$ to allow for the heterogeneity. In this example, we fix $N = 2^9p$ and $m = 2^4$, and set $\mathbf{B} = (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) \in \mathbb{R}^{p\times 1}$. We implement Algorithms 1 and 2 for both the sliced inverse regression and the cumulative slicing estimation.

We repeat each simulation 1,000 times, and report $\mathrm{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$, $\mathrm{dist}(\mathbf{B}^*, \mathbf{B})$, and $\mathrm{dist}(\widehat{\mathbf{B}}, \mathbf{B})$ to evaluate the performance of the distributed estimates. This requires that we approximate $\mathbf{P}(\mathbf{B}^*) \stackrel{\text{def}}{=} \mathbf{B}^*(\mathbf{B}^{*\mathrm{T}}\mathbf{B}^*)^{-1}\mathbf{B}^{*\mathrm{T}}$. We propose approximating $\mathbf{B}^*$ from the top d_0 eigenvectors of the average of $\mathbf{P}(\widehat{\mathbf{B}})$ obtained from 1,000 replications. The simulation results are summarized in Figures 1–2 for Examples 1–2 and in Table 1 for Example 3.

In subplots (A) and (D) in Figure 1, $\mathrm{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ decreases as the total sample size N increases, for each given m. This is in line with our anticipation that larger sample sizes typically yield better estimates. This phenomenon echoes the theoretical investigations in Lemma 1. The subplots (B) and (E) present the bias term $\mathrm{dist}(\mathbf{B}^*, \mathbf{B})$. As stated in Lemma 2, $\mathrm{dist}(\mathbf{B}^*, \mathbf{B})$ decreases as the local sample size $n = N/m$ increases, for each given m. In these examples, $\mathrm{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ dominates $\mathrm{dist}(\mathbf{B}^*, \mathbf{B})$. This is because n is very large and m is relatively small



(a) $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ (b) $\text{dist}(\mathbf{B}^*, \mathbf{B})$ (c) $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B})$

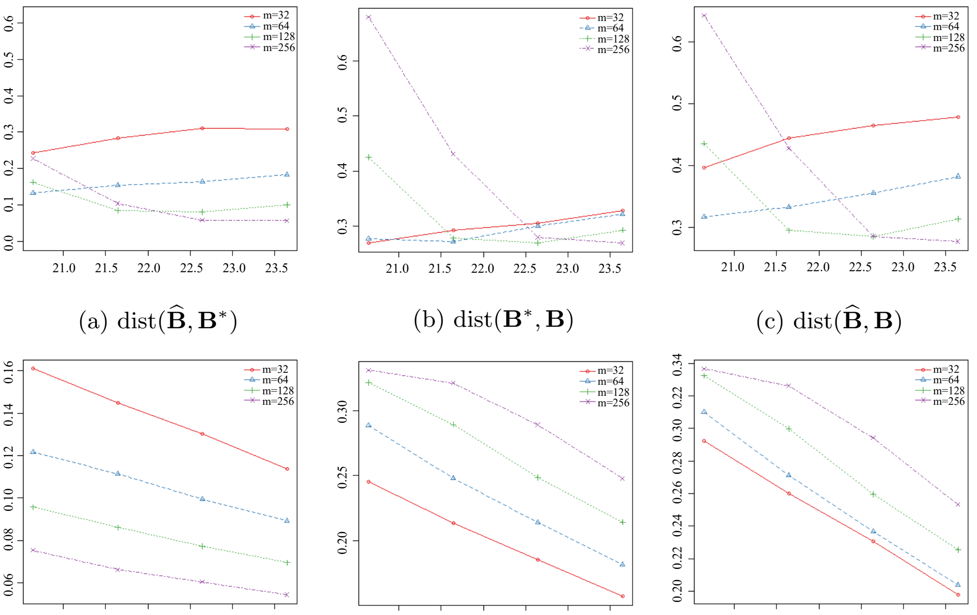
(d) $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ (e) $\text{dist}(\mathbf{B}^*, \mathbf{B})$ (f) $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B})$

Figure 1. The horizontal axis respresents the log(2)-transformed value of the total sample size N, and the vertical axis respresents $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ in (A) and (D), $\text{dist}(\mathbf{B}^*, \mathbf{B})$ in (B) and (E), and $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B})$ in (C) and (F). All distributed estimates of $\mathbf{B}$ are obtained using Algorithm 1. The distributed estimates of the sliced inverse regression are displayed in subplots (A)–(C), and those of the cumulative slicing estimation are displayed in subplots (D)–(F).

in our setting. It is thus not surprising that $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B})$ and $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ exhibit similar patterns.

Subplots (A) to (C) in Figure 2 present the results of the sliced inverse regression using Algorithm 2. Its performance is not very stable, probably because the slice number is fixed. This phenomenon echoes the empirical studies in Wang, Yu and Zhu (2021). Subplots (D) to (F) in Figure 2 show the results of the cumulative slicing estimation using Algorithm 2. For each given m, the distances decrease as N increases. Furthermore, $\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^*)$ does not dominate $\text{dist}(\mathbf{B}^*, \mathbf{B})$ in this example, which is consistent with the theoretical results in Theorem 2.

In Example 3, we compare the following six estimators, of which four are the distributed estimates for the sliced inverse regression and cumulative slicing estimation obtained using either Algorithm 1 or 2, and two pooled estimates are obtained by pooling all heterogeneous observations together. The value of k controls the degree of heterogeneity. Our aim is to compare the performance



(a) dist$(\widehat{\mathbf{B}}, \mathbf{B}^*)$ (b) dist$(\mathbf{B}^*, \mathbf{B})$ (c) dist$(\widehat{\mathbf{B}}, \mathbf{B})$

(d) dist$(\widehat{\mathbf{B}}, \mathbf{B}^*)$ (e) dist$(\mathbf{B}^*, \mathbf{B})$ (f) dist$(\widehat{\mathbf{B}}, \mathbf{B})$

Figure 2. The horizontal axis respresents the log(2)-transformed value of the total sample size N, and the vertical axis respresents dist$(\widehat{\mathbf{B}}, \mathbf{B}^*)$ in (A) and (D), dist$(\mathbf{B}^*, \mathbf{B})$ in (B) and (E), and dist$(\widehat{\mathbf{B}}, \mathbf{B})$ in (C) and (F). All distributed estimates of $\mathbf{B}$ are obtained using Algorithm 2. The distributed estimates of the sliced inverse regression are displayed in subplots (A)–(C), and those of the cumulative slicing estimation are displayed in subplots (D)–(F).

of distributed estimates with that of the pooled estimates. Table 1 summarizes the simulation results for the averages of dist$(\widehat{\mathbf{B}}, \mathbf{B})$ after 1,000 repetitions. Here, when the observations exhibit heterogeneity, the distributed estimates are much better than the pooled estimates, particularly when $k \geq 2$. This example demonstrates the advantages of distributed estimates over pooled estimates in the presence of heterogeneity.

4. American Gut Project Revisited

We revisit the American Gut Project described in Section 1, which is built on open-source and open-access principles. For a detailed description, see `http://humanfoodproject.com/americangut/`. It is known that billions of bacteria in the human gut participate in regulatiing, among others, the digestion function and the immunity system. Human gut microbiota begin to evolve immediately

Table 1. Simulation results for Example 3. We report the averages of $\mathrm{dist}(\widehat{\mathbf{B}}, \mathbf{B})$ after 1,000 repetitions. The six estimates are the two distributed estimates for the sliced inverse regression and cumulative slicing estimation obtained using Algorithm 1, denoted by $\widehat{\mathbf{B}}_{\mathrm{sir},1}$ and $\widehat{\mathbf{B}}_{\mathrm{cume},1}$, respectively, the two distributed estimates for the sliced inverse regression and cumulative slicing estimation obtained using Algorithm 2, denoted by $\widehat{\mathbf{B}}_{\mathrm{sir},2}$ and $\widehat{\mathbf{B}}_{\mathrm{cume},2}$, respectively, and the two pooled estimates obtained by pooling all observations together, denoted by $\widehat{\mathbf{B}}_{\mathrm{sir,pool}}$ and $\widehat{\mathbf{B}}_{\mathrm{cume,pool}}$, respectively.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$\widehat{\mathbf{B}}_{\mathrm{sir},1}$	0.5044	0.4279	0.3043	0.2606	0.2498
$\widehat{\mathbf{B}}_{\mathrm{cume},1}$	0.3656	0.3617	0.3005	0.2615	0.2498
$\widehat{\mathbf{B}}_{\mathrm{sir},2}$	0.4307	0.4103	0.3774	0.3498	0.3434
$\widehat{\mathbf{B}}_{\mathrm{cume},2}$	0.4301	0.4094	0.3765	0.3502	0.3434
$\widehat{\mathbf{B}}_{\mathrm{sir,pool}}$	0.5196	0.8062	1.0247	1.0225	1.0343
$\widehat{\mathbf{B}}_{\mathrm{cume,pool}}$	0.5290	0.8134	1.0232	1.0173	1.0330

when the embryo leaves the mother's body, and experience various stages of evolution at different ages. The immune system is weakest and most unstable during infancy. Therefore, many modern studies have attempted to reveal the underlying human gut microbiota structure at different ages (Yatsunenko et al. (2012); Nagpal et al. (2018); Xu, Zhu and Qiu (2019)). Unfortunately, little information has been revealed thus far on the age-related classes of human gut microbiota. One possible reason for this is that the sample size in a single study is often very limited. The American Gut Project, which aggregates many studies around the world, provides an excellent opportunity to discover the dominant classes of microbiota in growth. To control for the race effect on the gut metagenome construction, we consider only $N = 7470$ Caucasian samples. There are 29 batches of observations in total, two of which have very low expression levels, and are thus removed from the subsequent analysis, leaving $m = 27$ batches of observations. The number of subjects, n, ranges from 106 to 631 at different batches. The age of subjects, Y, ranging from 0 to 26, and the abundance levels for 215 classes predicted from 16S rRNA V4 gene fragments, $\mathbf{x} = (X_1, \ldots, X_{215})^{\mathrm{T}}$, are recorded. In this project, the dimension is large relative to the sample sizes for all batches. It is thus important to reduce the dimension of the explanatory variables prior to a subsequent statistical analysis. The principal components of $\mathbf{x}$ given Y are of independent interest, regardless of the prediction problem.

We first examine the condition number of the sample covariance matrices for all 27 batches of observations, and observe that the minimum of these condition numbers is 1.0666×10^{16}. This indicates that the sample covariance matrices are nearly singular. Thus, we advocate using Algorithm 2, proposed in Section 2, to perform the distributed sufficient dimension reduction. The cumulative slicing estimation is used here because it does not require specifying the number of slices. We apply the maximum eigenvalue ratio criterion (Luo, Wang and Tsai (2009)) to determine the structural dimension d_0 using two steps. In the first step, we apply this criterion to 27 batches individually. This yields 27 estimates of the structural dimensions, denoted as $\widehat{d}_j$, for $j = 1, \ldots, 27$. At each local node, we pass the $\widehat{d}_j$ principal eigenvectors, denoted as $\widehat{\mathbf{B}}_j \in \mathbb{R}^{p \times \widehat{d}_j}$, and the associated eigenvalues, which correspond to the diagonal elements in the diagonal matrix $\widehat{\mathbf{\Lambda}}_j \in \mathbb{R}^{\widehat{d}_j \times \widehat{d}_j}$, to the central node. In the second step, we apply the same criterion to

$$\widehat{\mathbf{T}} \overset{\text{def}}{=} m^{-1} \sum_{j=1}^{m} \widehat{\mathbf{B}}_j \widehat{\mathbf{\Lambda}}_j \widehat{\mathbf{B}}_j^{\mathrm{T}} \tag{4.1}$$

at the central node, which finally yields an estimate of d_0. Our analysis indicates that $\widehat{d}_0 = 2$, which shows how Y depends on $(\mathbf{B}^{\mathrm{T}}\mathbf{x})$, for $\mathbf{B} \in \mathbb{R}^{p \times \widehat{d}_0}$. Figure 3 displays the dependence structures of Y on $(\widehat{\mathbf{B}}_{a2}^{\mathrm{T}}\mathbf{x})$ using the mean functions $E(Y \mid \widehat{\mathbf{B}}_{a2}^{\mathrm{T}}\mathbf{x})$, where $\widehat{\mathbf{B}}_{a2}$ is the distributed estimate of $\mathbf{B}$ obtained from Algorithm 2. This exhibits an obvious heterogeneity issue, which is likely caused by the batch effects, because the underlying structure of gut microbiota correlated with age should, in general, be the same. A bootstrap procedure described in the online Supplementary Material shows that in the presence of heterogeneity, the distributed estimate is much more stable than the pooled estimate. This echoes our observations in Example 5.

A close inspection of the entries of $\widehat{\mathbf{B}}_{a2} = (\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2) \in \mathbb{R}^{p \times 2}$ reveals that there are only six rows with elements larger than 0.01, which are summarized in Table 2. Clearly, the three classes, Bacteroidia, Clostridia, and Gammaproteobacteria, are dominant in that the magnitudes of their coefficients are significantly larger than those of the other classes. This observation is in line with existing knowledge. In particular, Gao et al. (2018) found that the number of microbial interactions involving Bacteroidia increases over time during the first three years of life. Nie et al. (2017) conducted a functional analysis of the gut metagenome from yaks, and observed that Bacteroidia and Clostridia are closely related to energy metabolism and the synthesis of amino acids, which are essential in the early life of animals. An increase of Clostridia is also seen in the gut metagenome of premature neonates during hospitalization (Ferraris et al. (2012)). In addition, Mosbæk et al. (2016) found that Clostridia is actively involved in acetate

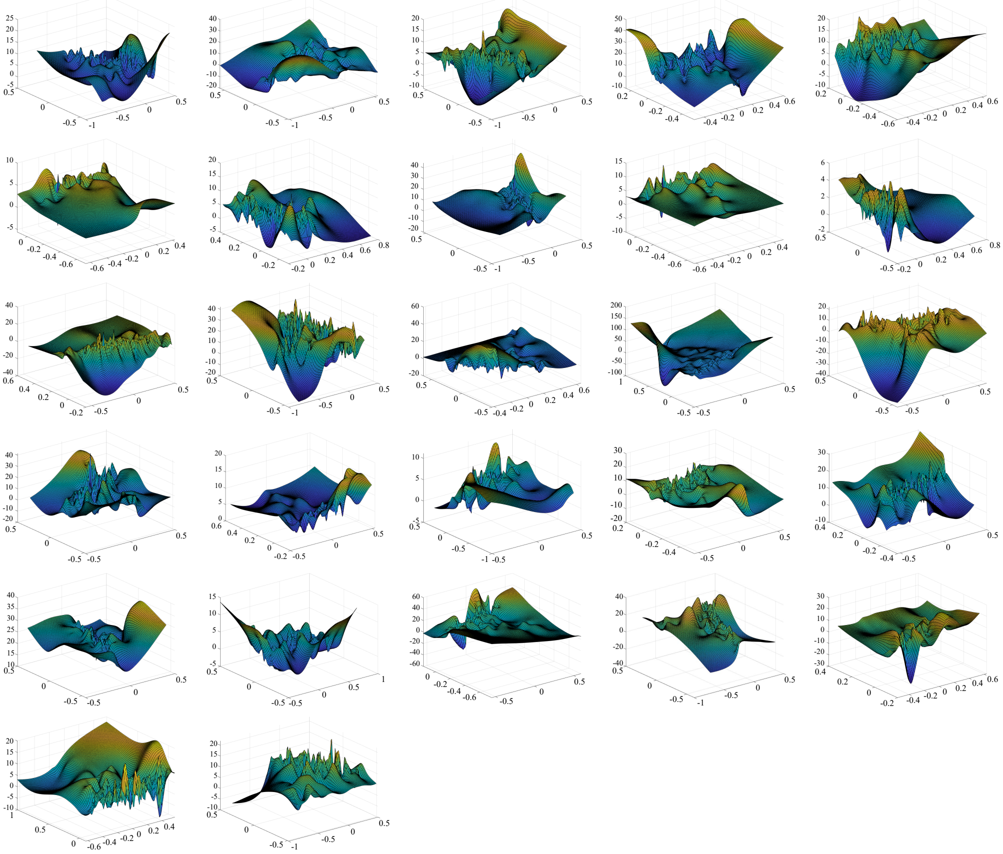


Figure 3. The dependence structures displayed using the mean functions $\widehat{E}(Y \mid \widehat{\mathbf{B}}_{a2}^{\mathrm{T}}\mathbf{x})$ across all 27 batches. They appear to be different, indicating the existence of heterogeneity.

turnover, indicating that it facilitates acetate consumption. Gammaproteobacteria is also found to be the predominant class (Chang et al. (2011)), which is related to fatty acid metabolism (Yao and Rock (2017)). These advances all indicate that the three classes, Bacteroidia, Clostridia and Gammaproteobacteria, play important roles during growth.

5. Conclusion

We propose two distributed algorithms for sufficient dimension reduction, one of which requires that all sample covariance matrices are invertible, and the other does not. These distributed algorithms are communication efficient. In addition, their nonasymptotic error bounds are established in the present context. One

Table 2. The six rows of $\widehat{\mathbf{B}}_{a2} = (\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$ with entries larger than 0.01.

	X_1	X_2	X_3	X_4	X_5	X_6
$\widehat{\boldsymbol{\beta}}_1$	0.8281	-0.5535	-0.0647	-0.0466	-0.0369	-0.0069
$\widehat{\boldsymbol{\beta}}_2$	0.4174	0.6987	-0.5729	-0.0814	-0.0047	-0.0145

X_1: Bacteroidia, X_2: Clostridia, X_3: Gammaproteobacteria, X_4: Bacilli, X_5: Actinobacteria and X_6: Erysipelotrichi

problem with these error bounds is that the biases of both distributed estimates do not vanish, even when the total sample size increases to infinity. A possible solution to this problem is to construct unbiased estimates for the kernel matrices of the sufficient dimension reduction methods; see Zhang and Zhang (2014) and Javanmard and Montanari (2014) for “de-biased” algorithms of regularized M-estimators. In addition, Lin and Li (2019) proposed a bias-correction approach to remove the biases of the least square estimates incurred with the ℓ_1- and ℓ_2-penalties. How to adapt these de-biased algorithms to the context of sufficient dimension reduction deserves further investigation. The second problem is how to decide the structural dimension of the central subspace in a distributed fashion. We adopt the maximum eigenvalue ratio criterion for simplicity. Implementing this criterion requires specifying the upper bound of the structural dimensions, which seems unrealistic in complicated situations because we are usually unaware of the underlying structures. An additional problem is how to study the theoretical properties of distributed estimates for second-order sufficient dimension reduction methods. Both the sliced inverse regression and the cumulative slicing estimation are first-order methods in the context of sufficient dimension reduction. Second-order methods include the sliced average variance estimation and the directional regression. The distributed algorithms can be readily adapted to second-order methods. However, their theoretical properties are much more difficult. These topics are left to future research.

Supplementary Material

The online Supplementary Material contains additional numerical studies and technical details.

Acknowledgments

Liping Zhu is the corresponding author. This research was supported by the Beijing Natural Science Foundation (Z190002), National Natural Science Foundation of China (12171477, 11731011, 11931014 and 71991471), and Shanghai Sailing Program (19YF1403400).

References

Bai, Z., Chan, R. H. and Luk, F. T. (2005). Principal component analysis for distributed data sets with updating. In *International Workshop on Advanced Parallel Processing Technologies* (Edited by J. Cao, W. Nejdl and M. Xu), 471–483. Springer, Berlin.

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

BIG Data Center Members (2018). Database resources of the big data center in 2018. *Nucleic Acids Research* **46**, D14–D20.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* **3**, 1–122.

Chang, J. Y., Shin, S. M., Chun, J., Lee, J.-H. and Seo, J.-K. (2011). Pyrosequencing-based molecular monitoring of the intestinal bacterial colonization in preterm infants. *Journal of Pediatric Gastroenterology and Nutrition* **53**, 512–519.

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

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. lst Edition. Wiley, New York.

Cook, R. D. and Weisberg, S. (1991). Comment on "sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 328–332.

Fan, J., Guo, Y. and Wang, K. (2021). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*. DOI: `10.1080/01621459.2021.1969238`.

Fan, J., Wang, D., Wang, K. and Zhu, Z. (2019). Distributed estimation of principal eigenspaces. *The Annals of Statistics* **47**, 3009–3031.

Ferraris, L., Butel, M. J., Campeotto, F., Vodovar, M., Rozé, J. C. and Aires, J. (2012). Clostridia in premature neonates' gut: Incidence, antibiotic susceptibility, and perinatal determinants influencing colonization. *PLoS ONE* **7**, e30594.

Gao, X., Huynh, B.-T., Guillemot, D., Glaser, P. and Opatowski, L. (2018). Inference of significant microbial interactions from longitudinal metagenomics data. *Frontiers in Microbiology* **9**, 1–10.

Hall, P. and Li, K. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics* **21**, 867–889.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics* **20**, 1040–1061.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15**, 2869–2909.

Jordan, M. I., Lee, J. D. and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* **114**, 668–681.

Kargupta, H., Huang, W., Sivakumar, K. and Johnson, E. (2001). Distributed clustering using collective principal component analysis. *Knowledge and Information Systems* **3**, 422–448.

Lee, J. D., Liu, Q., Sun, Y. and Taylor, J. E. (2017). Communication-efficient sparse regression. *Journal of Machine Learning Research* **18**, 1–30.

Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R.* 1st Edition. Chapman & Hall/CRC Press, Boca Raton.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

Li, Y. and Zhu, L. (2007). Asymptotics for sliced average variance estimation. *The Annals of Statistics* **35**, 41–69.

Liang, Y., Balcan, M.-F. F., Kanchanapally, V. and Woodruff, D. (2014). Improved distributed principal component analysis. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* (Edited by Z. Ghahramani, M. Welling, C. Cortes and N. D. Lawrence), 3113–3121. MIT Press, Cambridge.

Lin, L. and Li, F. (2019). A global bias-correction DC method for biased estimation under memory constraint. *arXiv preprint arXiv: 1904.07477.*

Lin, Q., Zhao, Z. and Liu, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics* **46**, 580–610.

Luo, R., Wang, H. and Tsai, C.-L. (2009). Contour projected dimension reduction. *The Annals of Statistics* **37**, 3743–3778.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107**, 168–179.

Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics* **41**, 250–268.

Ma, Y. and Zhu, L. (2014). On estimation efficiency of the central mean subspace. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 885–901.

Mosbæk, F., Kjeldal, H., Mulat, D. G., Albertsen, M., Ward, A. J., Feilberg, A. et al. (2016). Identification of syntrophic acetate-oxidizing bacteria in anaerobic digesters by combined protein-based stable isotope probing and metagenomics. *The ISME Journal* **10**, 2405–2418.

Nagpal, R., Mainali, R., Ahmadi, S., Wang, S., Singh, R., Kavanagh, K. et al. (2018). Gut microbiome and aging: Physiological and mechanistic insights. *Nutrition and Healthy Aging* **4**, 267–285.

Nie, Y., Zhou, Z., Guan, J., Xia, B., Luo, X., Yang, Y. et al. (2017). Dynamic changes of yak (Bos grunniens) gut microbiota during growth revealed by polymerase chain reaction-denaturing gradient gel electrophoresis and metagenomics. *Asian-Australasian Journal of Animal Sciences* **30**, 957–966.

Qi, H., Wang, T.-W. and Birdwell, J. D. (2004). Global principal component analysis for dimensionality reduction in distributed data mining. In *Statistical Data Mining and Knowledge Discovery* (Edited by H. Bozdogan), 327–342. Chapman and Hall/CRC Press, New York.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J. et al. (2015). Big data: Astronomical or genomical? *PLoS Biology* **13**, e1002195.

Tan, K., Shi, L. and Yu, Z. (2020). Sparse SIR: Optimal rates and adaptive estimation. *The Annals of Statistics* **48**, 64–85.

Tan, K. M., Wang, Z., Zhang, T., Liu, H. and Cook, R. D. (2018). A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika* **105**, 769–782.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Irvine.

Wang, C., Jiang, B. and Zhu, L. (2021). Penalized interaction estimation for ultrahigh dimensional quadratic regression. *Statistica Sinica* **31**, 1549–1570.

Wang, C., Yu, Z. and Zhu, L. (2021). On cumulative slicing estimation for high dimensional data. *Statistica Sinica* **31**, 223–242.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association* **103**, 811–821.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics* **35**, 2654–2690.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 363–410.

Xu, C., Zhu, H. and Qiu, P. (2019). Aging progression of human gut microbiota. *BMC Microbiology* **19**, 236.

Yao, J. and Rock, C. O. (2017). Exogenous fatty acid metabolism in bacteria. *Biochimie* **141**, 30–39.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M. et al. (2012). Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227.

Zhang, C. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 217–242.

Zhu, L., Miao, B. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**, 630–643.

Zhu, L. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica* **5**, 727–736.

Zhu, L., Zhu, L. and Feng, Z. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association* **105**, 1455–1466.

Kelin Xu

School of Public Health, Fudan University, Shanghai, China.

E-mail: xukelin@fudan.edu.cn

Liping Zhu

Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing, China.

E-mail: zhu.liping@ruc.edu.cn

Jianqing Fan

Department of Operations Research and Financial Engineering, Princeton University, USA.

E-mail: jqfan@princeton.edu

(Received January 2021; accepted November 2021)