Journal of Multivariate Analysis 190 (2022) 104991

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Distributed estimation in heterogeneous reduced rank regression: With application to order determination in sufficient dimension reduction

Canyi Chen [a], Wangli Xu [b,*], Liping Zhu [a]

[a] Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China
[b] Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing 100872, China

ARTICLE INFO

Article history:
Received 12 September 2021
Received in revised form 5 March 2022
Accepted 5 March 2022
Available online 15 March 2022

AMS 2020 subject classifications:
68W15
62J99
62B05

Keywords:
Distributed estimation
Oracle rate
Reduced rank regression
Sufficient dimension reduction

ABSTRACT

We are concerned with massive data which are possibly heterogeneous and scattered at different locations. We introduce a communication-efficient distributed algorithm to estimate the rank-deficient loading matrix in reduced rank regressions. The distributed algorithm, which proceeds iteratively, reduces the computational complexity substantially. During each iteration, it yields a closed-form solution and refines the previous estimators gradually. After a finite number of iterations, the final solution estimates the rank consistently, and more importantly, achieves the oracle rate. We recast sufficient dimension reduction methods under the framework of reduced rank regressions, which enables us to recover the central subspace and simultaneously estimate its structural dimension. We demonstrate the efficiency of our proposed distributed algorithm through simulations and an application to the airline on-line performance dataset consisting of 118,914,458 observations.

1. Introduction

Rapid developments of modern information technology allow us to collect unprecedentedly massive datasets in various areas. The observations are perhaps scattered across different locations, due to limitations of memory and storage space, or privacy concerns. Most conventional statistical methods, which require to pool all observations in a single machine to process the whole massive dataset, have limited values in such situations [30]. Distributed algorithms with low communication cost are thus urgently desired. Towards this goal, a plenty of distributed algorithms are developed for classic statistical methods, for example, kernel ridge regression [65], semi-parametric partially linear models [66], sparse regression [31], likelihood-based inference [4,30], quantile regression [12,13], linear support vector machine [54], Newton-type estimate [14], M-estimates [3,51], feature screening [36], principle component analysis [11,20], and bootstrap [61]. These distributed algorithms are able to accomplish large scale tasks and achieve the same convergence rate as their classic counterparts.

We consider a distributed algorithm for reduced rank regression [1,29,48]. To estimate the rank-deficient loading matrix, Bunea et al. [6] introduced an ℓ_0 penalty, and Yuan et al. [63] imposed an ℓ_1 penalty, which penalize the rank and the nuclear norm, respectively, of the rank-deficient loading matrix. Rohde and Tsybakov [49] suggested to impose the Schatten-b quasi-norm penalty. Chen et al. [10] improved Yuan et al. [63] by introducing an adaptive

[*] Corresponding author.
E-mail address: wlxu@ruc.edu.cn (W. Xu).

https://doi.org/10.1016/j.jmva.2022.104991

nuclear norm penalization, which yields a closed-form solution in matrix approximation problems. All these algorithms, however, directly use singular value decomposition in implementation, with computational complexity being at least quadratic in the total number of observations. This disadvantage limits their usefulness in big data analysis. To tackle this issue, we introduce a distributed algorithm, which is communication-efficient and alleviates computational complexity substantially. The distributed algorithm proceeds iteratively. During each iteration, the updated estimator refines its previous ones gradually. We show that, after a finite number of iterations, the resultant distributed estimator converges and achieves the oracle rate. In other words, in an asymptotic sense, it behaves equally well as if all observations were pooled together. This surprise finding is confirmed with extensive simulations. We also remark here that our proposal allows for heterogeneity. That is, we do not require all observations follow identical distributions. The heterogeneity issue is rarely considered in the literature of massive data analysis.

It is important to remark here that the strategy of one-shot divide-and-conquer, which takes the average of all local estimates, is suboptimal to estimate the rank-deficient loading matrix in the context of reduced rank regression. In a distributed setting, the observations are scattered at different locations. Should we estimate the loading matrix at each machine and combine the local estimates together, the final estimate of the loading matrix would likely no longer be low-rank, particularly when the number of local machines is quite many. This phenomenon is analogous to the sparsity. Should we combine many sparse estimates together, the resultant estimate would generally no longer be sparse [67]. Therefore, we advocate using our proposed distributed algorithm in situations with massive data.

Next we connect reduced rank regression with sufficient dimension reduction [15,32,33,44]. We formulate many existing methods under the framework of reduced rank regression. This allows us to implement our proposed distributed algorithm in the context of sufficient dimension reduction. To the best of our knowledge, this is perhaps the first attempt to connect reduced rank regression with sufficient dimension reduction, which allows us to estimate the central subspace and its structural dimension simultaneously. We demonstrate through simulations that, estimating the central subspace by taking advantage of its rank-deficiency property does improve its estimation accuracy.

This paper is organized as follows. We propose a distributed algorithm for reduced rank regressions and study the theoretical properties of the resultant solutions in Section 2. In Section 3, we recast existing sufficient dimension reduction methods under the framework of reduced rank regressions, which enables us to simultaneously estimate the central subspace and its structural dimension. We conduct simulations in Section 4 and a real-world application in Section 5 to demonstrate the efficiency of our proposed distributed algorithm. This paper is concluded in Section 6.

We introduce the following notations which will be used repetitively in subsequent exposition. We use $C, C_0, \ldots$, $c, c_0, \ldots$ to denote generic constants which may vary at each appearance. For a matrix $\mathbf{A} = (a_{i,j}) \in \mathbb{R}^{p\times q}$, we denote its rank and trace by $\mathrm{rank}(\mathbf{A})$ and $\mathrm{tr}(\mathbf{A})$, respectively, where $\mathrm{tr}(\mathbf{A})$ is the summation of all diagonals of $\mathbf{A}$. We denote by $\sigma_i(\mathbf{A})$ the ith largest singular value of $\mathbf{A}$. We further define

$$\|\mathbf{A}\| \overset{\text{def}}{=} \max_{\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2 = \sigma_{\max}(\mathbf{A}),\ \ \|\mathbf{A}\|_F^2 \overset{\text{def}}{=} \sum_{i=1}^{p}\sum_{j=1}^{q} a_{i,j}^2,\ \ \|\mathbf{A}\|_* \overset{\text{def}}{=} \sum_{i=1}^{p\wedge q}\sigma_i(\mathbf{A}).$$

For a vector $\mathbf{v}$, we define $\|\mathbf{v}\|_2$ to be the Euclidean norm.

2. Distributed estimation

We propose a distributed estimation for reduced rank regressions when the observations are heterogeneous and scattered at different locations. We assume there is a star network architecture which consists of m local machines, with the first of which serving as the central machine. We assume that the observations are scattered evenly at these machines. In particular, the observations at the jth machine are denoted as $\mathcal{D}_j \overset{\text{def}}{=} \{(\mathbf{x}_{i,j}, \mathbf{y}_{i,j}), i \in \{1, \ldots, n\}\}$, where $\mathbf{x}_{i,j} = (X_{i,j,1}, \ldots, X_{i,j,p})^\top \in \mathbb{R}^p$ is a p-vector of covariates and $\mathbf{y}_{i,j} = (Y_{i,j,1}, \ldots, Y_{i,j,q})^\top \in \mathbb{R}^q$ is a q-vector of responses. Define $\mathbf{X}_j \overset{\text{def}}{=} (\mathbf{x}_{1,j}, \ldots, \mathbf{x}_{n,j})^\top \in \mathbb{R}^{n\times p}$, and $\mathbf{Y}_j \overset{\text{def}}{=} (\mathbf{y}_{1,j}, \ldots, \mathbf{y}_{n,j})^\top \in \mathbb{R}^{n\times q}$, for $j \in \{1, \ldots, m\}$. We stack all $\mathbf{X}_j$s and $\mathbf{Y}_j$s in rows to form $\mathbf{X} \in \mathbb{R}^{nm\times p}$ and $\mathbf{Y} \in \mathbb{R}^{nm\times q}$, respectively. We assume that all observations, $\mathcal{D} \overset{\text{def}}{=} \{(\mathbf{x}_{i,j}, \mathbf{y}_{i,j}), i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}\}$, are independent; the observations scattered at the jth machine, $\mathcal{D}_j = \{(\mathbf{x}_{i,j}, \mathbf{y}_{i,j}), i \in \{1, \ldots, n\}\}$, are identically distributed; however, the observations scattered at different locations, $\{(\mathbf{X}_j, \mathbf{Y}_j), j \in \{1, \ldots, m\}\}$, are not necessarily identically distributed. In other words, we allow for heterogeneity in massive data. However, we do assume that all $\mathbf{x}_{i,j}$s have identical mean zero and mutual covariance matrix $\Sigma \overset{\text{def}}{=} \mathrm{cov}(\mathbf{x}_{i,j}, \mathbf{x}_{i,j}^\top)$, for all $i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}$. Let $N \overset{\text{def}}{=} nm$, which is extremely large such that distributed algorithms are badly desired.

2.1. A brief review of pooled penalized estimation in reduced rank regression

Let $\mathbf{B} \in \mathbb{R}^{p\times q}$ be a loading matrix. We define the local and global losses as follows,

$$\mathcal{L}_j(\mathbf{B}) \overset{\text{def}}{=} (2n)^{-1}\|\mathbf{Y}_j - \mathbf{X}_j\mathbf{B}\|_F^2,\ \ \mathcal{L}_N(\mathbf{B}) \overset{\text{def}}{=} m^{-1}\sum_{k=1}^{m}\mathcal{L}_k(\mathbf{B}). \tag{1}$$

We assume all $\mathbf{X}_j$s and $\mathbf{Y}_j$s are centralized within each local machine so that no intercept has to be included in the above losses. The underlying true loading matrix $\mathbf{B}_0$ and its rank d_0 are defined, respectively, by

$$\mathbf{B}_0 \overset{\text{def}}{=} \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} E\left\{\mathcal{L}_N(\mathbf{B})\right\}, \; d_0 \overset{\text{def}}{=} \text{rank}(\mathbf{B}_0). \tag{2}$$

At the sample level, we seek for a rank-deficient matrix $\mathbf{B}$ such that

$$\arg\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} \left\{\mathcal{L}_N(\mathbf{B}) + N^{-1}\mathcal{P}_\lambda(\mathbf{B})\right\}, \tag{3}$$

is minimized, where λ is a regularization parameter and $\mathcal{P}_\lambda(\mathbf{B})$ is introduced to encourage a rank-deficient $\mathbf{B}$.

There are two options for $\mathcal{P}_\lambda(\mathbf{B})$. The first is the trace norm penalty, $\mathcal{P}_\lambda(\mathbf{B}) = \lambda\|\mathbf{B}\|_* = \lambda\text{tr}(\mathbf{B})$, which is widely used in the literature. See, for example, Yuan et al. [63], Candès and Recht [9] and Bunea et al. [6]. In general, $(\mathbf{XB})$ is a low-rank matrix if $\mathbf{B}$ is low-rank. In particular, if $\mathbf{X}$ is a full-rank matrix, then $\text{rank}(\mathbf{XB}) = \text{rank}(\mathbf{B})$. Therefore, Chen et al. [10] suggested to modify the above penalty as $\mathcal{P}_\lambda(\mathbf{B}) = \lambda\|\mathbf{XB}\|_* = \lambda\text{tr}(\mathbf{XB})$. The advantage of using this modified penalty is that it allows for a closed-form solution to (3), and accordingly, is computationally more efficient than the first option. We advocate using the modified penalty in (3) for massive data, which boils down to

$$\widehat{\mathbf{B}}_{\text{pool}} \overset{\text{def}}{=} \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} \left\{\mathcal{L}_N(\mathbf{B}) + N^{-1}\lambda\|\mathbf{XB}\|_*\right\}. \tag{4}$$

We apply singular value decomposition to yield $\mathbf{Y} = \mathbf{UDV}^\top$. Define $\mathbf{T}_\lambda(\mathbf{Y}) \overset{\text{def}}{=} \mathbf{UD}_\lambda\mathbf{V}^\top$, where $\mathbf{D}_\lambda$ is a diagonal matrix with its (i, i)-th element being $\{\sigma_i(\mathbf{Y})-\lambda\}_+$ and all off-diagonal elements being identically 0. Throughout we define $x_+ \overset{\text{def}}{=} \max(x, 0)$. By invoking Theorem 2.1 in Cai et al. [8], Chen et al. [10] gave the explicit form by

$$\widehat{\mathbf{B}}_{\text{pool}} = N^{-1}\widehat{\Sigma}^{-1}\mathbf{X}^\top\mathbf{T}_\lambda(\mathbf{Y}), \text{ where } \widehat{\Sigma} \overset{\text{def}}{=} m^{-1}\sum_{j=1}^{m}\widehat{\Sigma}_j, \quad \text{and } \widehat{\Sigma}_j \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n}\mathbf{x}_{i,j}\mathbf{x}_{i,j}^\top. \tag{5}$$

The low-rankness of $\widehat{\mathbf{B}}_{\text{pool}}$ is induced by the soft-thresholding $\mathbf{T}_\lambda(\mathbf{Y})$ and completely controlled by the regularization parameter λ.

Now it remains to specify λ. Analogous to the Lasso problem, imposing an identical penalty intensity λ for different singular values in (4) usually leads to a biased estimate [46,73]. To ameliorate this biasness issue, we consider the adaptive trace norm $\|\mathbf{XB}\|_{*,\mathbf{w}}$ in place of $\|\mathbf{XB}\|_*$ in (4), which is given by

$$\|\mathbf{XB}\|_{*,\mathbf{w}} \overset{\text{def}}{=} \sum_{i=1}^{q} w_i\sigma_i(\mathbf{XB}),$$

where $\mathbf{w} = (w_1, \ldots, w_q)^\top$ is a vector of weights. In particular, for a fixed $\gamma > 0$, we specify $w_i = \{\sigma_i(\mathbf{XB})\}^{-\gamma}$ if $\sigma_i(\mathbf{XB}) > 0$, and $w_i = 1$ if $\sigma_i(\mathbf{XB}) = 0$. In the particular case of $\gamma = 1$, $\|\mathbf{XB}\|_{*,\mathbf{w}} = \text{rank}(\mathbf{XB})$, which indicates that $\|\mathbf{XB}\|_{*,\mathbf{w}}$ is a natural substitution of $\text{rank}(\mathbf{XB})$. In practice, however, $\mathbf{B}$, and accordingly, the "oracle" weights, $w_i = \{\sigma_i(\mathbf{XB})\}^{-\gamma}$, are unknown. Therefore, instead of using $(\mathbf{XB})$, we suggest to use $w_i = \{\sigma_i(\mathbf{Y})\}^{-\gamma}$ if $\sigma_i(\mathbf{Y}) > 0$, and $w_i = 1$ if $\sigma_i(\mathbf{Y}) = 0$, which indeed provides a good approximation to the oracle weights, particularly when $\mathbf{Y}$ is close to $(\mathbf{XB})$. Following Zou [73], we can choose γ through Bayesian information criterion or cross validation. Our limited experience indicates that, the resultant solutions are very robust for a wide range of γ. In Section 4, we shall simply set it as a constant, $\gamma = 10$, unless stated otherwise. We relegate the sensitivity analysis of γ to Section 4.3.

With the adaptive weights $\mathbf{w} = (w_1, \ldots, w_q)^\top$, we define

$$\widehat{\mathbf{B}}_{\text{adp-pool}} \overset{\text{def}}{=} \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} \left\{\mathcal{L}_N(\mathbf{B}) + N^{-1}\lambda\|\mathbf{XB}\|_{*,\mathbf{w}}\right\}. \tag{6}$$

Define $\mathbf{T}_{\lambda\mathbf{w}}(\mathbf{Y}) \overset{\text{def}}{=} \mathbf{UD}_{\lambda\mathbf{w}}\mathbf{V}^\top$, where $\mathbf{D}_{\lambda\mathbf{w}}$ is a diagonal matrix with its (i, i)-th element being $\{\sigma_i(\mathbf{Y}) - \lambda w_i\}_+$. Similar to $\widehat{\mathbf{B}}_{\text{pool}}$ given in (5), Chen et al. [10] derived an explicit form of $\widehat{\mathbf{B}}_{\text{adp-pool}}$ as follows,

$$\widehat{\mathbf{B}}_{\text{adp-pool}} = N^{-1}\widehat{\Sigma}^{-1}\mathbf{X}^\top\mathbf{T}_{\lambda\mathbf{w}}(\mathbf{Y}), \tag{7}$$

To calculate $\widehat{\mathbf{B}}_{\text{pool}}$ in (5) or $\widehat{\mathbf{B}}_{\text{adp-pool}}$ in (7), we are required to perform a singular value decomposition on $\mathbf{Y} \in \mathbb{R}^{N\times q}$, which has the computational complexity of $O(N^2q)$. The total complexities of calculating $\widehat{\mathbf{B}}_{\text{pool}}$ and $\widehat{\mathbf{B}}_{\text{adp-pool}}$ are of the same order $O(N^2q + Npq + p^2q + p^3 + Np^2)$, which is apparently prohibitive when N is extremely large.

2.2. A distributed estimation in heterogeneous reduced rank regression

Next we introduce a distributed algorithm to seek for a low-rank matrix $\mathbf{B}$ which minimizes the discrepancy between $\mathbf{Y}$ and $(\mathbf{XB})$. We first note that, for an arbitrary initial estimate $\widehat{\mathbf{B}}^{(0)}$,

$$\mathcal{L}_N(\mathbf{B}) = \mathcal{L}_N(\widehat{\mathbf{B}}^{(0)}) + \frac{1}{N}\text{tr}\left\{(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}^{(0)})^\top\mathbf{X}(\widehat{\mathbf{B}}^{(0)} - \mathbf{B})\right\} + \frac{1}{2}\text{tr}\left\{(\widehat{\mathbf{B}}^{(0)} - \mathbf{B})^\top\widehat{\Sigma}(\widehat{\mathbf{B}}^{(0)} - \mathbf{B})\right\}.$$

If we work with the observations in the central machine only, we can also have

$$\mathcal{L}_1(\mathbf{B}) = \mathcal{L}_1(\widehat{\mathbf{B}}^{(0)}) + \frac{1}{n}\mathrm{tr}\big\{(\mathbf{Y}_1 - \mathbf{X}_1\widehat{\mathbf{B}}^{(0)})^\top \mathbf{X}_1(\widehat{\mathbf{B}}^{(0)} - \mathbf{B})\big\} + \frac{1}{2}\mathrm{tr}\big\{(\widehat{\mathbf{B}}^{(0)} - \mathbf{B})^\top \widehat{\Sigma}_1(\widehat{\mathbf{B}}^{(0)} - \mathbf{B})\big\}.$$

For notational clarity, we define

$$\mathbf{Z}_j \overset{\mathrm{def}}{=} n^{-1}\sum_{i=1}^{n} \mathbf{x}_{i,j}\mathbf{y}_{i,j}^\top,\ \mathbf{Z}_N \overset{\mathrm{def}}{=} m^{-1}\sum_{j=1}^{m}\mathbf{Z}_j.$$

Both $\widehat{\Sigma}_1$ and $\widehat{\Sigma}$ are consistent estimates of Σ as long as n is large enough. If $\widehat{\Sigma}_1$ and $\widehat{\Sigma}$ are sufficiently "close" such that $\widehat{\Sigma}_1 \approx \widehat{\Sigma}$, it follows that

$$\mathcal{L}_N(\mathbf{B}) \approx \mathcal{L}_N(\widehat{\mathbf{B}}^{(0)}) - \mathcal{L}_1(\widehat{\mathbf{B}}^{(0)}) + \mathrm{tr}\big\{(\widehat{\mathbf{B}}^{(0)})^\top(\mathbf{Z}_N - \mathbf{Z}_1 + \widehat{\Sigma}_1\widehat{\mathbf{B}}^{(0)} - \widehat{\Sigma}\widehat{\mathbf{B}}^{(0)})\big\} + \mathcal{L}_1(\mathbf{B}) - \mathrm{tr}\big\{\mathbf{B}^\top(\mathbf{Z}_N - \mathbf{Z}_1 + \widehat{\Sigma}_1\widehat{\mathbf{B}}^{(0)} - \widehat{\Sigma}\widehat{\mathbf{B}}^{(0)})\big\}.$$

In other words, minimizing $\mathcal{L}_N(\mathbf{B})$ to seek for a rank-deficient matrix $\mathbf{B}$ is approximately equivalent to minimizing the right hand side of the above display. If we further ignore all quantities that are irrelevant to $\mathbf{B}$, then minimizing the right hand side of the above display is indeed equivalent exactly to minimizing

$$\widetilde{\mathcal{L}}(\mathbf{B}) \overset{\mathrm{def}}{=} \frac{1}{2}\left\|\mathbf{M}^{(1)} - \mathbf{X}_1\mathbf{B}\right\|_F^2,$$

where $\mathbf{M}^{(1)} \overset{\mathrm{def}}{=} \mathbf{X}_1\widehat{\Sigma}_1^{-1}\{\mathbf{Z}_N + (\widehat{\Sigma}_1 - \widehat{\Sigma})\widehat{\mathbf{B}}^{(0)}\}$. Similar strategies are also used by Jordan et al. [30]. If the total sample size N is extremely large, we suggest to work with $\widetilde{\mathcal{L}}(\mathbf{B})$ instead of the global loss $\mathcal{L}(\mathbf{B})$. To be precise, once $\mathbf{M}^{(1)}$ is available, it suffices to update $\widehat{\mathbf{B}}^{(0)}$ using the observations scattered on the central machine through

$$\widehat{\mathbf{B}}^{(1)} \overset{\mathrm{def}}{=} \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\arg\min}\big\{\widetilde{\mathcal{L}}(\mathbf{B}) + \lambda^{(1)}\left\|\mathbf{X}_1\mathbf{B}\right\|_{*,\widehat{\mathbf{w}}^{(1)}}\big\}, \tag{8}$$

where $\widehat{\mathbf{w}}^{(1)} = (\widehat{w}_1^{(1)}, \ldots, \widehat{w}_q^{(1)})^\top \in \mathbb{R}^q$, and $\widehat{w}_i^{(1)} = \{\sigma_i(\mathbf{M}^{(1)})\}^{-\gamma}$ if $\sigma_i(\mathbf{M}^{(1)}) > 0$ and $\widehat{w}_i^{(1)} = 1$ if $\sigma_i(\mathbf{M}^{(1)}) = 0$, for a certain constant $\gamma > 0$. Following similar arguments to derive (7), we give the explicit form of $\widehat{\mathbf{B}}^{(1)}$ by

$$\widehat{\mathbf{B}}^{(1)} = n^{-1}\widehat{\Sigma}_1^{-1}\mathbf{X}_1{}^\top\mathbf{T}_{\lambda^{(1)}\widehat{\mathbf{w}}^{(1)}}(\mathbf{M}^{(1)}). \tag{9}$$

To be precise, we apply singular value decomposition to obtain $\mathbf{M}^{(1)} = \mathbf{U}^{(1)}\mathbf{D}(\mathbf{V}^{(1)})^\top$. Define $\mathbf{T}_{\lambda^{(1)}\widehat{\mathbf{w}}^{(1)}}(\mathbf{M}^{(1)}) \overset{\mathrm{def}}{=} \mathbf{U}^{(1)}\mathbf{D}_{\lambda^{(1)}\widehat{\mathbf{w}}^{(1)}}(\mathbf{V}^{(1)})^\top$, where $\mathbf{D}_{\lambda^{(1)}\widehat{\mathbf{w}}^{(1)}}$ is a diagonal matrix with its (i,i)-th element being $\{\sigma_i(\mathbf{M}^{(1)}) - \lambda^{(1)}\widehat{w}_i^{(1)}\}_+$. In particular, if $m = 1$, $\widehat{\mathbf{B}}^{(1)}$, which is updated from an arbitrary $\widehat{\mathbf{B}}^{(0)}$ through (9), coincides with $\widehat{\mathbf{B}}_{\text{adp-pool}}$ defined in (6).

Because all the singular values $\sigma_i(\mathbf{M}^{(1)})$ are sorted in a descending order, we estimate the rank of $\mathbf{B}_0$ through

$$\widehat{d}^{(1)} \overset{\mathrm{def}}{=} \max\big\{i\colon \sigma_i(\mathbf{M}^{(1)}) > \lambda^{(1)}\widehat{\varpi}_i^{(1)}\big\} = \max\big\{i\colon \{\sigma_i(\mathbf{M}^{(1)})\}^{\gamma+1} > \lambda^{(1)}\big\}, \tag{10}$$

for $\lambda^{(1)}$ ranging from 0 to $\{\sigma_1(\mathbf{M}^{(1)})\}^{\gamma+1}$. We define $\widehat{d}^{(1)} = 0$ if $\lambda^{(1)} \geq \{\sigma_1(\mathbf{M}^{(1)})\}^{\gamma+1}$, where the optimal $\lambda^{(1)}$ is decided through a data-driven criterion, say, cross-validation.

We transmit $\mathbf{Z}_j$ and $\widehat{\Sigma}_j\widehat{\mathbf{B}}^{(0)}$ from the jth local machine to the central machine to form $\mathbf{Z}_N$ and $\widehat{\Sigma}\widehat{\mathbf{B}}^{(0)}$ in $\mathbf{M}^{(1)}$ in the first machine, which yields (9) immediately. The communication cost of this distributed algorithm is of order $O(mpq)$, which is the minimal price that we have to pay for. This indicates that our proposed distributed algorithm is communication-efficient.

The estimate in (8) can be iteratively refined. To be specific, let $\widehat{\mathbf{B}}^{(t-1)}$ be the distributed estimate in the $(t-1)$-th iteration, we propose the following estimate,

$$\widehat{\mathbf{B}}^{(t)} \overset{\mathrm{def}}{=} \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\arg\min}\Big\{\frac{1}{2}\left\|\mathbf{X}_1\mathbf{B} - \mathbf{M}^{(t)}\right\|_F^2 + \lambda^{(t)}\left\|\mathbf{X}_1\mathbf{B}\right\|_{*,\widehat{\mathbf{w}}^{(t)}}\Big\}, \tag{11}$$

where $\mathbf{M}^{(t)} \overset{\mathrm{def}}{=} \mathbf{X}_1\widehat{\Sigma}_1^{-1}\{\mathbf{Z}_N + (\widehat{\Sigma}_1 - \widehat{\Sigma})\widehat{\mathbf{B}}^{(t-1)}\}$, $\widehat{\mathbf{w}}^{(t)} = (\widehat{w}_1^{(t)}, \ldots, \widehat{w}_q^{(t)})^\top \in \mathbb{R}^q$, and $\widehat{w}_i^{(t)} = \{\sigma_i(\mathbf{M}^{(t)})\}^{-\gamma}$ if $\sigma_i(\mathbf{M}^{(t)}) > 0$ and $\widehat{w}_i^{(t)} = 1$ if $\sigma_i(\mathbf{M}^{(t)}) = 0$. Singular value decomposition yields $\mathbf{M}^{(t)} = \mathbf{U}^{(t)}\mathbf{D}(\mathbf{V}^{(t)})^\top$. Define $\mathbf{T}_{\lambda^{(t)}\widehat{\mathbf{w}}^{(t)}}(\mathbf{M}^{(t)}) \overset{\mathrm{def}}{=} \mathbf{U}^{(t)}\mathbf{D}_{\lambda^{(t)}\widehat{\mathbf{w}}^{(t)}}(\mathbf{V}^{(t)})^\top$, where $\mathbf{D}_{\lambda^{(t)}\widehat{\mathbf{w}}^{(t)}}$ is a diagonal matrix with its (i,i)-th element being $\{\sigma_i(\mathbf{M}^{(t)}) - \lambda^{(t)}\widehat{w}_i^{(t)}\}_+$. We estimate the rank of $\mathbf{B}_0$ with

$$\widehat{d}^{(t)} \overset{\mathrm{def}}{=} \max\big\{i\colon \sigma_i(\mathbf{M}^{(t)}) > \lambda^{(t)}\widehat{\varpi}_i^{(t)}\big\} = \max\left\{i\colon \{\sigma_i(\mathbf{M}^{(t)})\}^{\gamma+1} > \lambda^{(t)}\right\}. \tag{12}$$

We simply use the observations in the central machine to produce an initial estimate $\widehat{\mathbf{B}}^{(0)}$. To be precise, we define

$$\widehat{\mathbf{B}}^{(0)} \overset{\mathrm{def}}{=} \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\arg\min}\Big\{\frac{1}{2}\|\mathbf{Y}_1 - \mathbf{X}_1\mathbf{B}\|_F^2 + \lambda_0\|\mathbf{X}_1\mathbf{B}\|_{*,\widehat{\mathbf{w}}^{(0)}}\Big\}, \tag{13}$$

where the adaptive weights $\widehat{\mathbf{w}}^{(0)} = (\widehat{w}_1^{(0)}, \ldots, \widehat{w}_q^{(0)})^\top \in \mathbb{R}^q$ are given by $\widehat{w}_i^{(0)} = \{\sigma_i(\mathbf{Y}_1)\}^{-\gamma}$ if $\sigma_i(\mathbf{Y}_1) > 0$ and $\widehat{w}_i^{(0)} = 1$ if $\sigma_i(\mathbf{Y}_1) = 0$ for a pre-specified constant $\gamma > 0$.

In (17), we shall show that, our proposed distributed algorithm converges within a constant number of iterations. Therefore, the computational complexity of our proposed distributed algorithm is $O(n^2q+npq+p^2q+p^3+Np^2)$, which is linear in N as long as $n^2 \le N$, and substantially smaller than $O(N^2q+Npq+p^2q+p^3+Np^2)$. The latter is the computational complexity of the algorithm suggested by Chen et al. [10].

The above distributed algorithm is summarized in Algorithm 1.

Algorithm 1 Distributed Reduced Rank Regression

```
Input: Observations $\{(\mathbf{x}_{i,j}, \mathbf{y}_{i,j}): i \in \{1,\dots,n\}, j \in \{1,\dots,m\}\}$, maximal number of iterations $T$, regularization parameters $\{\lambda^{(t)}: t \in \{1,\dots,T\}\}$, $\lambda^{(0)}$ and $\gamma$.
Compute the initial estimator $\widehat{\mathbf{B}}^{(0)}$ using (13).
for $t \in \{1,\dots,T\}$ do
    Transmit $\widehat{\mathbf{B}}^{(t-1)}$ from the central machine to the local machines labeled with $2,\dots,m$.
    for $j \in \{2,\dots,m\}$ do
        Transmit $\widehat{\Sigma}_j\widehat{\mathbf{B}}^{(t-1)}$ and $\mathbf{Z}_j$ to the central machine, which are obtained from the local machines labeled with $2,\dots,m$.
    end for
    Compute $\widehat{\mathbf{B}}^{(t)}$ on the central machine with (11).
end for
Output: The estimates $\widehat{\mathbf{B}}^{(t)}$and $\widehat{d}^{(t)}$ are generated sequentially on the central machine.
```

2.3. Theoretical properties

Next we study the theoretical properties of $\widehat{\mathbf{B}}^{(t)}$ and $\widehat{d}^{(t)}$. To quantify the accuracy of $\widehat{\mathbf{B}}^{(t)}$, we use the Frobenius norm $\|\widehat{\mathbf{B}}^{(0)} - \mathbf{B}_0\|_F$ and the trace correlation [21] between $\widehat{\mathbf{B}}^{(t)}$ and $\mathbf{B}_0$. The latter has the form of

$$\mathrm{corr}^2(\widehat{\mathbf{B}}^{(t)}, \mathbf{B}_0) \overset{\text{def}}{=} \mathrm{tr}\{\mathbf{P}(\widehat{\mathbf{B}}^{(t)})\mathbf{P}(\mathbf{B}_0)\}/d_0, \tag{14}$$

where $\mathbf{P}(\cdot)$ is a projection operator, namely, $\mathbf{P}(\mathbf{B}) \overset{\text{def}}{=} \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top$. The trace correlation $\mathrm{corr}^2(\widehat{\mathbf{B}}^{(t)}, \mathbf{B}_0)$ ranges from 0 to 1, with larger values indicating better performance. We assume the following conditions.

(C1) Let $\Sigma = \mathrm{cov}(\mathbf{x}_{i,j}, \mathbf{x}_{i,j}^\top)$, for all $i \in \{1,\dots,n\}, j \in \{1,\dots,m\}$. The smallest and largest singular values of Σ, denoted by $\sigma_{\min}(\Sigma)$ and $\sigma_{\max}(\Sigma)$, respectively, satisfy that $c_0^{-1} \le \sigma_{\min}(\Sigma) \le \sigma_{\max}(\Sigma) \le c_0$, for $c_0 > 0$.

(C2) Let $\mathbf{z}_{i,j} \overset{\text{def}}{=} (\mathbf{x}_{i,j}^\top, \mathbf{y}_{i,j}^\top)^\top$. All $\mathbf{z}_{i,j}$s are independent and sub-Gaussian. That is, there exist two positive constants c_1 and C_1 such that

$$\sup_{\|\boldsymbol{\alpha}\|_2=1} E\left[\exp\{c_1(\mathbf{z}_{i,j}^\top\boldsymbol{\alpha})^2\}\right] \le C_1, \text{ holds uniformly for all } i, j.$$

(C3) Let $a_n^2 \overset{\text{def}}{=} O\{d_0(p+q)/n\}$, which is $o(1)$. The initial estimate satisfies $\|\widehat{\mathbf{B}}^{(0)} - \mathbf{B}_0\|_F^2 = O_p(a_n^2)$.

(C4) The local sample size n satisfies $n = N^\alpha$, for $0 < \alpha < 1$, and the dimension p satisfies $d_0p/n = o(1)$.

These conditions are widely used in the literature and typically regarded as mild. See, for example, Chen et al. [12] and Wang et al. [55]. Condition (C3) requires implicitly the initial estimate $\widehat{\mathbf{B}}^{(0)}$ to be consistent, which can be met under Condition (C2). See, e.g., Theorem 4 of Chen et al. [10]. Condition (C4) imposes mild conditions on the local sample size n and the dimension p. We require p be a smaller order of n/d_0, which is easy to be satisfied when the local sample size n grows and the loading matrix B_0 is low-rank with $d_0 = O(1)$.

Define $a_{N,0} \overset{\text{def}}{=} a_n$ and $a_{N,t} \overset{\text{def}}{=} \{d_0(p+q)/N\}^{1/2} + a_{N,0}\,(d_0p/n)^{t/2}$ for $t \ge 1$. Let $b_{N,t} \overset{\text{def}}{=} \{(p+q)/N\}^{1/2} + a_{N,t-1}(p/n)^{1/2}$ and $\lambda^{(t)} \overset{\text{def}}{=} C_0\left(n^{1/2}b_{N,t}\right)^{\gamma+1}$, for $C_0 > 0$. In Theorem 2,we shall show that, $a_{N,t}$ is exactly the convergence rate for tth iteration. By contrast, $\{(p+q)/N\}^{1/2}$ is the convergence rate of (6) with all observations pooled together.

Theorem 1. *In addition to Conditions (C1)–(C4), we assume $\sigma_{d_0}(\mathbf{B}_0) > C_1b_{N,t}$, for a sufficient large constant C_1. Then $\Pr(\widehat{d}^{(t)} = d_0)$ approaches one, as $N \to \infty$.*

Lemma 1 paves the road for proving Theorem 1. Define

$$\mathcal{E}_1 \overset{\text{def}}{=} \{\sigma_s^{1/2}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) > 3/2n^{-1/2}(\lambda^{(t)})^{1/(\gamma+1)}\} \text{ and } \mathcal{E}_2 \overset{\text{def}}{=} \{\sigma_{s+1}^{1/2}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) \le 1/2n^{-1/2}(\lambda^{(t)})^{1/(\gamma+1)}\}.$$

Let $\mathbf{M}_\epsilon^{(t)} \overset{\text{def}}{=} \mathbf{M}^{(t)} - \mathbf{X}_1\mathbf{B}_0 = \mathbf{X}_1\widehat{\Sigma}_1^{-1}\{\mathbf{Z}_N - \widehat{\Sigma}\mathbf{B}_0 + (\widehat{\Sigma}_1 - \widehat{\Sigma})(\widehat{\mathbf{B}}^{(t-1)} - \mathbf{B}_0)\}$, $\mathcal{E}_s^{(t)} \overset{\text{def}}{=} \{\widehat{d}^{(t)} = s\}$ and $\mathcal{E}_\epsilon^{(t)} \overset{\text{def}}{=} \{\sigma_1(\mathbf{M}_\epsilon^{(t)}) \ge 1/2(\lambda^{(t)})^{1/(\gamma+1)}\}$, where $d_0 = \mathrm{rank}(\mathbf{B}_0)$, $\widehat{d}^{(t)}$ is the estimate of d_0, and γ is the parameter in the adaptive weights $\widehat{w}_i^{(t)}$.

Lemma 1. *For each given $t \ge 1$, suppose $\mathcal{E}_1$ and $\mathcal{E}_2$ hold for $s \le d_0$. On the events $\mathcal{E}_1$ and $\mathcal{E}_2$, we have $(\mathcal{E}_\epsilon^{(t)})^C \subseteq \mathcal{E}_s^{(t)}$.*

Proof of Lemma 1. This lemma is in spirit parallel to Lemma 1 in [10]. Invoking the definition of $\widehat{d}^{(t)}$ in (12), $\widehat{d}^{(t)} > s$ holds if and only if $\sigma_{s+1}(\mathbf{M}^{(t)}) > (\lambda^{(t)})^{1/(\gamma+1)}$ and $\widehat{d}^{(t)} < s$ holds if and only if $\sigma_s(\mathbf{M}^{(t)}) \le (\lambda^{(t)})^{1/(\gamma+1)}$. It follows that the event $(\mathcal{E}_s^{(t)})^C = \{\widehat{d}^{(t)} \neq s\}$ is equivalent to $\{\sigma_{s+1}(\mathbf{M}^{(t)}) > (\lambda^{(t)})^{1/(\gamma+1)}$ or $\sigma_s(\mathbf{M}^{(t)}) \le (\lambda^{(t)})^{1/(\gamma+1)}\}$.

By Weyl's inequalities on singular values [22, Section 6.7] and $\mathbf{M}_\epsilon^{(t)} = \mathbf{M}^{(t)} - \mathbf{X}_1\mathbf{B}_0$, we have $\sigma_1(\mathbf{M}_\epsilon^{(t)}) \ge \sigma_{s+1}(\mathbf{M}^{(t)}) - \sigma_{s+1}(\mathbf{X}_1\mathbf{B}_0)$ and $\sigma_1(\mathbf{M}_\epsilon^{(t)}) \ge \sigma_s(\mathbf{X}_1\mathbf{B}_0) - \sigma_s(\mathbf{M}^{(t)})$. Therefore, $\sigma_{s+1}(\mathbf{M}^{(t)}) > (\lambda^{(t)})^{1/(\gamma+1)}$ implies $\sigma_1(\mathbf{M}_\epsilon^{(t)}) \ge (\lambda^{(t)})^{1/(\gamma+1)} - \sigma_{s+1}(\mathbf{X}_1\mathbf{B}_0)$, and $\sigma_s(\mathbf{M}^{(t)}) \le (\lambda^{(t)})^{1/(\gamma+1)}$ implies $\sigma_1(\mathbf{M}_\epsilon^{(t)}) \ge \sigma_s(\mathbf{X}_1\mathbf{B}_0) - (\lambda^{(t)})^{1/(\gamma+1)}$. It ensures that

$$(\mathcal{E}_s^{(t)})^C \subseteq \left[\sigma_1(\mathbf{M}_\epsilon^{(t)}) \ge \min\{(\lambda^{(t)})^{1/(\gamma+1)} - \sigma_{s+1}(\mathbf{X}_1\mathbf{B}_0), \sigma_s(\mathbf{X}_1\mathbf{B}_0) - (\lambda^{(t)})^{1/(\gamma+1)}\}\right].$$

It remains to bound the right hand side of the above display. By definition of singular values, on $\mathcal{E}_1$ and $\mathcal{E}_2$, we have $\sigma_{s+1}(\mathbf{X}_1\mathbf{B}_0) = n^{1/2}\sigma_{s+1}^{1/2}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) \le 1/2(\lambda^{(t)})^{1/(\gamma+1)}$ and $\sigma_s(\mathbf{X}_1\mathbf{B}_0) = n^{1/2}\sigma_s^{1/2}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) \ge 3/2(\lambda^{(t)})^{1/(\gamma+1)}$, which implies that $\min\{(\lambda^{(t)})^{1/(\gamma+1)} - \sigma_{s+1}(\mathbf{X}_1\mathbf{B}_0), \sigma_s(\mathbf{X}_1\mathbf{B}_0) - (\lambda^{(t)})^{1/(\gamma+1)}\} \ge 1/2(\lambda^{(t)})^{1/(\gamma+1)}$. This completes the proof. □

Proof of Theorem 1. We only provide the proof for $t = 1$ because the proofs for $t \ge 2$ are similar. Define $\mathcal{E}_1 \overset{\text{def}}{=} \{\sigma_s^{1/2}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) > 3/2n^{-1/2}(\lambda^{(t)})^{1/(\gamma+1)}\}$ and $\mathcal{E}_2 \overset{\text{def}}{=} \{\sigma_{s+1}^{1/2}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) \le 1/2n^{-1/2}(\lambda^{(t)})^{1/(\gamma+1)}\}$. We first show that events $\mathcal{E}_1$ and $\mathcal{E}_2$ hold with probability approaching one for $s = d_0$.

We show $\Pr(\mathcal{E}_1) \to 1$ as N diverges. Without loss of generality, we assume $p \ge q$. By Theorem 1.3.22 in [26], the nonzero singular values of $\widehat{\Sigma}_1\mathbf{B}_0\mathbf{B}_0{}^\top$ are the same as those of $\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0$. Thus, $\sigma_{d_0}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) = \sigma_{d_0}(\widehat{\Sigma}_1\mathbf{B}_0\mathbf{B}_0{}^\top)$. This, together with the inequality $\sigma_{d_0}(\widehat{\Sigma}_1\mathbf{B}_0\mathbf{B}_0{}^\top) \ge \sigma_{\min}(\widehat{\Sigma}_1)\sigma_{d_0}^2(\mathbf{B}_0)$ [47, Theorem 9] and Condition (C1), immediately implies that $\sigma_{d_0}^{1/2}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) \ge C\sigma_{d_0}(\mathbf{B}_0)$ with probability approaching one. This together with $\sigma_{d_0}(\mathbf{B}_0) > C_1 n^{-1/2}(\lambda^{(1)})^{1/(\gamma+1)}$ yields $\Pr(\mathcal{E}_1) = \Pr\{\sigma_{d_0}^{1/2}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) \ge 3/2n^{-1/2}(\lambda^{(1)})^{1/(\gamma+1)}\} \to 1$, where $C_1 > 0$ is sufficiently large.

Next we show $\Pr(\mathcal{E}_2) \to 1$. The fact that $\operatorname{rank}(\mathbf{B}_0) = d_0$ and $\operatorname{rank}(\widehat{\Sigma}_1) = p$ with probability approaching one imply $\Pr\{\sigma_{d_0+1}^{1/2}(\mathbf{B}_0{}^\top\widehat{\Sigma}_1\mathbf{B}_0) = 0\} \to 1$. Recall that $\mathbf{M}_\epsilon^{(1)} = \mathbf{M}^{(1)} - \mathbf{X}_1\mathbf{B}_0$. Taking $s = d_0$ in Lemma 1 in $\mathcal{E}_1$ and $\mathcal{E}_2$, we have

$$\begin{aligned}\Pr(\widehat{d}^{(1)} = d_0) &\ge 1 - \Pr\{\sigma_1(\mathbf{M}_\epsilon^{(1)}) \ge 1/2(\lambda^{(1)})^{1/(\gamma+1)}\} - o(1)\\ &= 1 - \Pr\left[\|\mathbf{X}_1\widehat{\Sigma}_1^{-1}\{\mathbf{Z}_N - \widehat{\Sigma}\mathbf{B}_0 + (\widehat{\Sigma}_1 - \widehat{\Sigma})(\widehat{\mathbf{B}}^{(0)} - \mathbf{B}_0)\}\| \ge 1/2(\lambda^{(1)})^{1/(\gamma+1)}\right] - o(1).\end{aligned}$$

It remains to bound the second term in the right hand side of above inequality. By the triangular inequality, we have

$$\Pr(\widehat{d}^{(1)} = d_0) \ge 1 - \Pr\left[\|\mathbf{X}_1\widehat{\Sigma}_1^{-1}(\mathbf{Z}_N - \widehat{\Sigma}\mathbf{B}_0)\| + \|\mathbf{X}_1\widehat{\Sigma}_1^{-1}(\widehat{\Sigma}_1 - \widehat{\Sigma})(\widehat{\mathbf{B}}^{(0)} - \mathbf{B}_0)\| \ge 1/2(\lambda^{(1)})^{1/(\gamma+1)}\right] - o(1).$$

Note that $\|\mathbf{X}_1\| = \|\mathbf{X}_1{}^\top\mathbf{X}_1\|^{1/2} = O_p(n^{1/2})$. This, together with Lemma 2, ensures that

$$\|\mathbf{X}_1\widehat{\Sigma}_1^{-1}(\widehat{\Sigma}_1 - \widehat{\Sigma})(\widehat{\mathbf{B}}^{(0)} - \mathbf{B}_0)\| \le \|\mathbf{X}_1\|\|\widehat{\Sigma}_1^{-1}\|\|\widehat{\Sigma}_1 - \widehat{\Sigma}\|\|\widehat{\mathbf{B}}^{(0)} - \mathbf{B}_0\| = O_p\{a_n p^{1/2}\}.$$

By Lemma 3, we have

$$\|\mathbf{X}_1\widehat{\Sigma}_1^{-1}(\mathbf{Z}_N - \widehat{\Sigma}\mathbf{B}_0)\| \le \|\mathbf{X}_1\|\|\widehat{\Sigma}_1^{-1}\|\|\mathbf{Z}_N - \widehat{\Sigma}\mathbf{B}_0\| = O_p\left[\{n(p+q)/N\}^{1/2}\right].$$

The proof for $t = 1$ is now completed by invoking the definition of $\lambda^{(1)}$. □

Theorem 1 implies that $\widehat{d}^{(t)}$ converges in probability to d_0. In other words, the rank of $\mathbf{B}_0$ is estimated consistently, as long as the minimal signal strength is strong enough. We remark here that $b_{N,t}$ shrinks to zero, as t increases, indicating that restrictions on the minimal signal strength become weaker as the distributed algorithm iterates.

Theorem 2. *Under the conditions of Theorem 1, $\|\widehat{\mathbf{B}}^{(t)} - \mathbf{B}_0\|_F^2 = O_p(a_{N,t}^2)$. If we further require $\sigma_{d_0}(\mathbf{B}_0) \ge c_1$ for a constant $c_1 > 0$, then $\operatorname{corr}^2(\widehat{\mathbf{B}}^{(t)}, \mathbf{B}_0) = 1 - O_p(a_{N,t}^2/d_0)$.*

The following three lemmas pave the road for proving Theorem 2.

Lemma 2. *Under Conditions (C1)–(C3), $\|\widehat{\Sigma}_1 - \Sigma\| = O_p\left\{(p/n)^{1/2}\right\}$, $\|\widehat{\Sigma} - \Sigma\| = O_p\left\{(p/N)^{1/2}\right\}$.*

Proof of Lemma 2. This is proved in Vershynin [52, Proposition 2.1]. □

Lemma 3. *Under Conditions (C1)–(C3), $\|\mathbf{Z}_N - \widehat{\Sigma}\mathbf{B}_0\| = O_p\left[\{(p+q)/N\}^{1/2}\right]$.*

Proof of Lemma 3. Let $Q_1 \overset{\text{def}}{=} \|\mathbf{Z}_N - \Sigma\mathbf{B}_0\|$ and $Q_2 \overset{\text{def}}{=} \|(\widehat{\Sigma} - \Sigma)\mathbf{B}_0\|$. By triangular inequality, $\|\mathbf{Z}_N - \widehat{\Sigma}\mathbf{B}_0\| \le Q_1 + Q_2$. It remains to bound Q_1 and Q_2. The Cauchy–Schwarz inequality gives that $\|\mathbf{AB}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$ [24], which implies that Q_2 is dominated by $\|\widehat{\Sigma} - \Sigma\|$. Invoking Lemma 2, we immediately have $Q_2 = O_p\{(p/N)^{1/2}\}$. It thus remains to bound Q_1. We resort to the ε-net argument [53, Theorem 4.6.1]. Recall the definition of $\mathbf{B}_0$ in (2). It follows immediately that

$$\mathbf{B}_0 = \Sigma^{-1}\sum_{j=1}^m E(\mathbf{x}_{i,j}\mathbf{y}_{i,j}{}^\top)/m,$$

which leads to

$$Q_1 = \left\| N^{-1} \sum_{i,j} \{ \mathbf{x}_{i,j}\mathbf{y}_{i,j}^\top - E(\mathbf{x}_{i,j}\mathbf{y}_{i,j}^\top) \} \right\|.$$

We divide the whole proof into three steps.

Step.1 Find a good approximation of Q_1. By Corollary 4.2.13 in [53], there exist 1/4-net $\mathcal{M}$ of the unit sphere $\mathcal{S}^{p-1}$ and 1/4-net $\mathcal{N}$ of the unit sphere $\mathcal{S}^{q-1}$ with cardinalities $|\mathcal{M}| \le 9^p$ and $|\mathcal{N}| \le 9^q$. The exercise 4.4.3 of [53] entails that

$$Q_1 \le 2 \sup_{\boldsymbol{\alpha}\in\mathcal{M},\ \boldsymbol{\beta}\in\mathcal{N}} \left| N^{-1} \sum_{i,j} \{ \boldsymbol{\alpha}^\top \mathbf{x}_{i,j}\mathbf{y}_{i,j}{}^\top \boldsymbol{\beta} - E(\boldsymbol{\alpha}^\top \mathbf{x}_{i,j}\mathbf{y}_{i,j}{}^\top \boldsymbol{\beta}) \} \right|.$$

Let $\mathbf{t}_{i,j}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \overset{\text{def}}{=} \boldsymbol{\alpha}^\top \mathbf{x}_{i,j}\mathbf{y}_{i,j}{}^\top \boldsymbol{\beta} - E(\boldsymbol{\alpha}^\top \mathbf{x}_{i,j}\mathbf{y}_{i,j}^\top \boldsymbol{\beta})$. It suffices to bound

$$\sup_{\boldsymbol{\alpha}\in\mathcal{M},\ \boldsymbol{\beta}\in\mathcal{N}} \left| N^{-1} \sum_{i,j} \mathbf{t}_{i,j}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \right|$$

from above.

Step.2 Derive a probability concentration inequality for each given term $\mathbf{t}_{i,j}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. For any fixed $\boldsymbol{\alpha} \in \mathcal{M}$ and $\boldsymbol{\beta} \in \mathcal{N}$, by Lemma 2.7.6 in [53], $\mathbf{t}_{i,j}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is mean zero sub-exponential random variable. Thus, with Bernstein's inequality [53, Corollary 2.8.3], we obtain that for $\varepsilon \overset{\text{def}}{=} K^2 \max(\delta, \delta^2)$ where $\delta = C[\{(p+q)/N\}^{1/2} + t/N^{1/2}]$ with $t > 0$, K is a generic constant depending on the sub-Gaussian norm of $\mathbf{x}_{i,j}$s and C is a sufficiently large positive constant,

$$\Pr\left(\left| N^{-1} \sum_{i,j} \mathbf{t}_{i,j}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \right| \ge \varepsilon/2 \right) \le 2\exp\{-c_1 \min(\varepsilon^2/K^4, \epsilon/K^2)N\} = 2\exp(-c_1\delta^2 N) \le 2\exp\{-c_1C^2(p+q+t^2)\}.$$

The last inequality follows from the definition of δ and the fact that $(a+b)^2 \ge a^2 + b^2$ for $a, b > 0$.

Step.3 Find a uniform bound to unfix $\boldsymbol{\alpha} \in \mathcal{M}$ and $\boldsymbol{\beta} \in \mathcal{N}$. Recall $|\mathcal{M}| \le 9^p$ and $|\mathcal{N}| \le 9^q$. We have

$$\Pr\left\{ \sup_{\boldsymbol{\alpha}\in\mathcal{M},\ \boldsymbol{\beta}\in\mathcal{N}} \left| N^{-1} \sum_{i,j} \mathbf{t}_{i,j}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \right| \ge \varepsilon/2 \right\} \le 9^{p+q} \exp\{-c_1C^2(p+q+t^2)\}.$$

Set C to be sufficiently large. The right hand side is bounded by $2\exp(-t^2)$. Thus we have $Q_1 = O_p[\{(p+q)/N\}^{1/2}]$. Now we complete the proof of Lemma 3. □

Lemma 4. *Under the conditions in Theorem 1,* $\|\mathbf{X}_1\widehat{\mathbf{B}}^{(t)} - \mathbf{X}_1\mathbf{B}_0\|_F^2 = O_p\{(\lambda^{(t)})^{2/(\gamma+1)}d_0\}$.

Proof of Lemma 4. We provide the proof for $t = 1$ only because the proofs for $t \ge 2$ are similar. Let $r_{\mathbf{A}} \overset{\text{def}}{=} \text{rank}(\mathbf{A})$ for matrix $\mathbf{A}$. Denote $Q(\mathbf{B}) = 1/2\|\mathbf{X}_1\mathbf{B} - \mathbf{X}_1\widehat{\Sigma}_1^{-1}\left\{\mathbf{Z}_N + (\widehat{\Sigma}_1 - \widehat{\Sigma})\widehat{\mathbf{B}}^{(0)}\right\}\|_F^2 + \lambda^{(1)}\|\mathbf{X}_1\mathbf{B}\|_{*,\widehat{\mathbf{w}}^{(1)}}$. By the definition of $\widehat{\mathbf{B}}^{(1)}$ in (8), we have $Q(\widehat{\mathbf{B}}^{(1)}) \le Q(\mathbf{B}_0)$. It immediately follows that

$$\begin{aligned} 1/2\|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)} - \mathbf{X}_1\mathbf{B}_0\|_F^2 &\le \langle \mathbf{M}_\epsilon^{(1)}, \mathbf{X}_1\widehat{\mathbf{B}}^{(1)} - \mathbf{X}_1\mathbf{B}_0 \rangle + \lambda^{(1)}(\|\mathbf{X}_1\mathbf{B}_0\|_{*,\widehat{\mathbf{w}}^{(1)}} - \|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}\|_{*,\widehat{\mathbf{w}}^{(1)}}) \\ &\le \|\mathbf{M}_\epsilon^{(1)}\|\|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)} - \mathbf{X}_1\mathbf{B}_0\|_* + \lambda^{(1)}(\|\mathbf{X}_1\mathbf{B}_0\|_{*,\widehat{\mathbf{w}}^{(1)}} - \|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}\|_{*,\widehat{\mathbf{w}}^{(1)}}), \end{aligned}$$

where $\langle \mathbf{A}, \mathbf{B} \rangle \overset{\text{def}}{=} \text{tr}(\mathbf{A}^\top\mathbf{B})$ for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times q}$. The last step is due to von Neumann's trace duality inequality. Denote $\Delta\mathbf{B}_0 \overset{\text{def}}{=} \mathbf{X}_1\widehat{\mathbf{B}}^{(1)} - \mathbf{X}_1\mathbf{B}_0$. By Cauchy–Schwarz inequality, we have

$$1/2\|\Delta\mathbf{B}_0\|_F^2 \le \|\mathbf{M}_\epsilon^{(1)}\| r_{\Delta\mathbf{B}_0}^{1/2} \|\Delta\mathbf{B}_0\|_F + \lambda^{(1)}(\|\mathbf{X}_1\mathbf{B}_0\|_{*,\widehat{\mathbf{w}}^{(1)}} - \|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}\|_{*,\widehat{\mathbf{w}}^{(1)}}). \tag{15}$$

We bound these two terms in the right hand side of above inequality, respectively. By the proof of Theorem 1, we have $\Pr\{\sigma_1(\mathbf{M}_\epsilon^{(1)}) < 1/2(\lambda^{(1)})^{1/(\gamma+1)}\} \to 1$. This, together with $r_{\Delta\mathbf{B}_0} \le r_{\widehat{\mathbf{B}}^{(1)} - \mathbf{B}_0} \le \widehat{d}^{(1)} + d_0$, ensures that

$$\|\mathbf{M}_\epsilon^{(1)}\| r_{\Delta\mathbf{B}_0}^{1/2} \|\Delta\mathbf{B}_0\|_F \le 1/2(\lambda^{(1)})^{1/(\gamma+1)}(\widehat{d}^{(1)} + d_0)^{1/2}\|\Delta\mathbf{B}_0\|_F. \tag{16}$$

It remains to deal with $\lambda^{(1)}(\|\mathbf{X}_1\mathbf{B}_0\|_{*,\widehat{\mathbf{w}}^{(1)}} - \|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}\|_{*,\widehat{\mathbf{w}}^{(1)}})$. Note that

$$\begin{aligned} &\|\mathbf{X}_1\mathbf{B}_0\|_{*,\widehat{\mathbf{w}}^{(1)}} - \|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}\|_{*,\widehat{\mathbf{w}}^{(1)}} = \sum_{i=1}^{n\wedge q} \widehat{\omega}_i^{(1)}\sigma_i(\mathbf{X}_1\mathbf{B}_0) - \sum_{i=1}^{n\wedge q} \widehat{\omega}_i^{(1)}\sigma_i(\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}) \\ &= \widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} \sum_{i=1}^{\widehat{d}^{(1)}} \sigma_i(\mathbf{X}_1\mathbf{B}_0) - \widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} \sum_{i=1}^{\widehat{d}^{(1)}} \sigma_i(\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}) + \sum_{i=1}^{\widehat{d}^{(1)}} (\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} - \widehat{\omega}_i^{(1)})\sigma_i(\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}) - \sum_{i=1}^{\widehat{d}^{(1)}} (\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} - \widehat{\omega}_i^{(1)})\sigma_i(\mathbf{X}_1\mathbf{B}_0). \end{aligned}$$

Recall that $\widehat{\omega}_i^{(1)} = \sigma_i^{-\gamma}(\mathbf{M}^{(1)})$, so $\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} - \widehat{\omega}_1^{(1)} \geq \cdots \geq \widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} - \widehat{\omega}_{\widehat{d}^{(1)}-1}^{(1)} \geq 0$. Therefore, both $p_1(\cdot) \overset{\text{def}}{=} \sum_{i=1}^{\widehat{d}^{(1)}} \sigma_i(\cdot)$ and $p_2(\cdot) \overset{\text{def}}{=} \sum_{i=1}^{\widehat{d}^{(1)}} (\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} - \widehat{\omega}_i^{(1)})\sigma_i(\cdot)$ satisfy the triangle inequality. The convexity of $p_1(\cdot)$ and $p_2(\cdot)$ has been shown in Theorem 1 of [10]. Therefore, with $p_1(\cdot)$ and $p_2(\cdot)$, we can rewrite (17) as

$$\|\mathbf{X}_1\mathbf{B}_0\|_{*,\widehat{\mathbf{w}}^{(1)}} - \|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}\|_{*,\widehat{\mathbf{w}}^{(1)}} = \widehat{\omega}_{\widehat{d}^{(1)}}^{(1)}\{p_1(\mathbf{X}_1\mathbf{B}_0) - p_1(\mathbf{X}_1\widehat{\mathbf{B}}^{(1)})\} + p_2(\mathbf{X}_1\mathbf{B}_0) - p_2(\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}),$$

which, together with the triangle inequality, leads to $\|\mathbf{X}_1\mathbf{B}_0\|_{*,\widehat{\mathbf{w}}^{(1)}} - \|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}\|_{*,\widehat{\mathbf{w}}^{(1)}} \leq \widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} p_1(\varDelta\mathbf{B}_0) + p_2(\varDelta\mathbf{B}_0)$. Note that $\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} - \widehat{\omega}_i^{(1)} \leq \widehat{\omega}_{\widehat{d}^{(1)}}^{(1)}$. The right hand side of the above inequality can be further bounded above by $2\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} p_1(\varDelta\mathbf{B}_0)$. It remains us to bound $\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)}$ and $p_1(\varDelta\mathbf{B}_0)$, respectively.

We first bound $\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)}$. Weyl's inequalities [22, Section 6.7], together with the fact that $\mathbf{M}^{(1)} = \mathbf{X}_1\mathbf{B}_0 + \mathbf{M}_\epsilon^{(1)}$, imply that $\sigma_{\widehat{d}^{(1)}}(\mathbf{M}^{(1)}) \geq \sigma_{\widehat{d}^{(1)}}(\mathbf{X}_1\mathbf{B}_0) - \sigma_1(\mathbf{M}_\epsilon^{(1)})$. Note that, by the proof of Theorem 1,

$$\Pr\{\sigma_1(\mathbf{M}_\epsilon^{(1)}) < 1/2(\lambda^{(1)})^{1/(\gamma+1)}\} \to 1,\ \Pr\{\sigma_{d_0}^{1/2}(\mathbf{B}_0{}^\top\widehat{\varSigma}_1\mathbf{B}_0) > C_2 n^{-1/2}(\lambda^{(1)})^{1/(\gamma+1)}\} \to 1.$$

Therefore, $\Pr\{\sigma_{\widehat{d}^{(1)}}(\mathbf{M}^{(1)}) \geq C_3(\lambda^{(1)})^{1/(\gamma+1)}\} \to 1$ for a certain constant $C_3 > 0$. Recall that $\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} = \sigma_{\widehat{d}^{(1)}}^{-\gamma}(\mathbf{M}^{(1)})$. It follows immediately that $2\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)} \leq C(\lambda^{(1)})^{-\gamma/(\gamma+1)}$.

As for $p_1(\varDelta\mathbf{B}_0)$, by Cauchy–Schwarz inequality, we have $p_1(\varDelta\mathbf{B}_0) \leq r_{\varDelta\mathbf{B}_0}^{1/2}\|\varDelta\mathbf{B}_0\|_F$. Summarizing the bounds for $\widehat{\omega}_{\widehat{d}^{(1)}}^{(1)}$ and $p_1(\varDelta\mathbf{B}_0)$, we have, $\|\mathbf{X}_1\mathbf{B}_0\|_{*,\widehat{\mathbf{w}}^{(1)}} - \|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)}\|_{*,\widehat{\mathbf{w}}^{(1)}} \leq C(\lambda^{(1)})^{-\gamma/(\gamma+1)} r_{\varDelta\mathbf{B}_0}^{1/2}\|\varDelta\mathbf{B}_0\|_F$. This, together with (15) and (16), ensures $\|\varDelta\mathbf{B}_0\|_F^2 \leq 2\{1/2(\lambda^{(1)})^{1/(\gamma+1)}(\widehat{d}^{(1)} + d_0)^{1/2} + C(\lambda^{(1)})^{1/(\gamma+1)}(\widehat{d}^{(1)} + d_0)^{1/2}\}\|\varDelta\mathbf{B}_0\|_F$.

We complete the proof by invoking Theorem 1. □

Proof of Theorem 2. We only show Theorem 2 for $t = 1$. The proof for $t \geq 2$ follows from similar arguments. We first show Theorem 2(i). Note that $\|\mathbf{X}_1\widehat{\mathbf{B}}^{(1)} - \mathbf{X}_1\mathbf{B}_0\|_F^2 \geq n\sigma_{\min}(\widehat{\varSigma}_1)\|\widehat{\mathbf{B}}^{(1)} - \mathbf{B}_0\|_F^2 \geq c_1 n\sigma_{\min}(\varSigma)\|\widehat{\mathbf{B}}^{(1)} - \mathbf{B}_0\|_F^2$. This, together with Lemma 4, completes the proof of Theorem 2(i).

Turn to Theorem 2 (ii). Let $\widehat{\mathbf{B}} \overset{\text{def}}{=} \widehat{\mathbf{B}}^{(1)}$ for brevity. We first note that $\operatorname{corr}^2\{\widehat{\mathbf{B}}, \mathbf{B}_0\} = 1 - \|\mathbf{P}(\widehat{\mathbf{B}}) - \mathbf{P}(\mathbf{B}_0)\|_F^2/(2d_0)$ with probability approaching one. It suffices to bound $\|\mathbf{P}(\widehat{\mathbf{B}}) - \mathbf{P}(\mathbf{B}_0)\|_F^2$. By Davis–Kahan Theorem [62, Theorem 2], Condition (C1) and the assumption that $\sigma_{d_0}(\mathbf{B}_0) \geq c_1$ for $c_1 > 0$, we can use similar arguments in the proof of Theorem 1 in [55] to obtain that $\|\mathbf{P}(\widehat{\mathbf{B}}) - \mathbf{P}(\mathbf{B}_0)\|_F \leq C_0\|\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top - \mathbf{B}_0\mathbf{B}_0{}^\top\|_F$. Therefore, $\operatorname{corr}^2(\widehat{\mathbf{B}}, \mathbf{B}_0) = 1 - O_p(\|\widehat{\mathbf{B}} - \mathbf{B}_0\|_F^2/d_0)$. This proof is completed by invoking Theorem 2(i). □

Theorem 2 ensures that the estimation accuracy of $\widehat{\mathbf{B}}^{(t)}$ improves $\widehat{\mathbf{B}}^{(t-1)}$ by a proportionality of $(d_0p/n)^{1/2}$, as the distributed algorithm proceeds iteratively. It can be verified that, as long as

$$t \geq c_2 \log(N/n)/\log\{n/(d_0p)\},\ \text{for a certain positive constant } c_2, \tag{17}$$

$\{d_0(p+q)/N\}^{1/2}$ dominates $a_{N,0}\,(d_0p/n)^{t/2}$. In this circumstance, $\|\widehat{\mathbf{B}}^{(t)} - \mathbf{B}_0\|_F^2 = O_p\,\{d_0(p+q)/N\}$ and $\operatorname{corr}(\widehat{\mathbf{B}}^{(t)}, \mathbf{B}_0) = 1 - O_p\{(p+q)/N\}$, both of which are the oracle convergence rate of the pooled estimate $\widehat{\mathbf{B}}_{\text{adp-pool}}$. We also remark here that, under Condition (C4), the number of iterations t is finite. Because the communication complexity of each iteration is of order $O(mpq)$, the total communication cost of our algorithm is still of order $O(mpq)$. In this sense, our proposed algorithm is communication-efficient [30]. Therefore, (17) is satisfied if $t \geq c_2$. An important implication is that, within a constant number of iterations, $\widehat{\mathbf{B}}^{(t)}$ achieves the oracle convergence rate. In other words, our proposed distributed algorithm provides a nearly oracle estimate.

We remark here that, to ease illustration, we assume that the observations are evenly partitioned. The local sample size n is in spirit corresponding to the number of observations in the first machine in Algorithm 1. Once the local sample size n is specified, our algorithm is free of the partitioning of the whole dataset.

3. Order determination

We study how to recover the central subspace and decide its structural dimension in the context of sufficient dimension reduction [15,32,33,44]. Existing literature dealt these two problems separately, whereas we shall recast them as a single problem of estimating a rank-deficient matrix, which allows us to solve these two problems simultaneously. In addition, our proposed distributed algorithm generalizes the usefulness of sufficient dimension reduction in situations with heterogeneous massive data. The issues of heterogeneity and huge volume of big data are rarely simultaneously studied in the literature.

In this section we consider the univariate response case for now. We assume that all N observations are independent and scattered evenly on m local machines. The observations in the jth machine, $\mathcal{D}_j \overset{\text{def}}{=} \{(\mathbf{x}_{i,j}, Y_{i,j}), i \in \{1, \ldots, n\}\}$, follow an identical distribution. However, $\mathcal{D}_1, \ldots, \mathcal{D}_m$ are not necessarily identically distributed. For simplicity, we assume all $\mathbf{x}_{i,j}$s are centralized to have identical mean zero and mutual covariance matrix $\varSigma$. We denote the conditional distribution by $F(\cdot \mid \cdot)$, and assume that $F_{i,j}(Y_{i,j} \mid \mathbf{x}_{i,j}) = F_j(Y_{i,j} \mid \mathbf{x}_{i,j})$, for $j \in \{1, \ldots, m\}$. Sufficient dimension reduction assumes there exists a $p \times d_j$ matrix $\mathbf{A}_j$ with the smallest column dimension such that $F_j(Y_{i,j} \mid \mathbf{x}_{i,j}) = F_j(Y_{i,j} \mid \mathbf{x}_{i,j}^\top\mathbf{A}_j)$. That is, the observations within each local machine are independent and identically distributed. Although $\mathbf{A}_j$ is not unique, its column space is. We denote the column space of $\mathbf{A}_j$ by $\mathcal{S}_j$. We aim to recover the central space $\mathcal{S}_{Y|\mathbf{x}}$, which is defined to be the space spanned by the columns of $\mathbf{A}_1, \dots, \mathbf{A}_m$, and its structural dimension d_0. Apparently,

$$\mathcal{S}_{Y|\mathbf{x}} = \mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_m, \quad \max_{j\in\{1,\dots,m\}} d_j \le d_0 \le \sum_{j=1}^{m} d_j.$$

Suppose $\mathbf{U}_0 \in \mathbb{R}^{p\times d_0}$ is an arbitrary basis matrix of $\mathcal{S}_{Y|\mathbf{x}}$.

We advocate using sufficient dimension reduction because it possesses at least two distinctive advantages. First, it does not assume parametric models for $F_j(\cdot \mid \cdot)$s. In this sense, it is model-free. Second, using $(\mathbf{x}_{i,j}^{\top}\mathbf{A}_j)$ in place of $\mathbf{x}_{i,j}$ to predict $Y_{i,j}$ does not cause any loss of information in the sense that $F_j(Y_{i,j} \mid \mathbf{x}_{i,j}) = F_j(Y_{i,j} \mid \mathbf{x}_{i,j}^{\top}\mathbf{A}_j)$. Therefore, it is "sufficient" to use the dimension-reduced instead of the original high dimensional covariates to perform prediction.

Should all observations be pooled together, which corresponds to the special case of $m = 1$, many methods have been proposed to recover $\mathcal{S}_{Y|\mathbf{x}}$ and estimate d_0. Popular methods to identify $\mathcal{S}_{Y|\mathbf{x}}$ include inverse regressions [18,32,35,37,72], forward regressions [56,57] and semiparametric approaches [42,43,45]. The methods to estimate d_0 can be classified into four categories: the sequential tests which are introduced by Li [32] and developed by Schott [50], Cook and Li [16],Cook and Ni [17] and Bura and Yang [7], the information criteria which include Zhu et al. [68], Luo et al. [40] and Ma and Zhang [41], the bootstrap methods which measure the variations of eigen-spaces [59] or combine the magnitudes of eigenvalues additionally [39], and penalization method [70] which recovers $\mathcal{S}_{Y|\mathbf{x}}$ and estimates d_0 simultaneously.

We consider massive data which are possibly heterogeneous and scattered at m local machines. We first formulate many sufficient dimension reduction methods under the least squares framework. Towards this goal, we consider q transformations of $Y_{i,j}$, denoted as $T_k(Y_{i,j})$. Define $Y_{i,j,k} \overset{\text{def}}{=} T_k(Y_{i,j}) - E\{T_k(Y_{i,j})\}$ and $\mathbf{y}_{i,j} \overset{\text{def}}{=} (Y_{i,j,1}, \dots, Y_{i,j,q})^{\top}$. There are many choices for such transformations. For example, Li [32] considered $T_k(Y_{i,j}) = I(b_{k-1} \le Y_{i,j} < b_k)$, where $-\infty = b_0 < b_1 \cdots < b_{q-1} < b_q = \infty$. Zhu et al. [72] suggested $T_k(Y_{i,j}) = I(Y_{i,j} < b_k)$. Yin and Cook [60] chose the polynomial functions $T_k(Y_{i,j}) = Y_{i,j}^{k}$. Fung et al. [23] and Bi and Qu [5] used the B-spline bases as the transformations. Indeed, Fourier [72] or other transformation functions can also be used.

We consider the minimization problem (2) where $\mathcal{L}_N(\mathbf{B})$ is defined in (1). The minimizer $\mathbf{B}_0$ is a $p \times q$ matrix. We seek for $\mathbf{U}_0$, which is a basis of $\mathcal{S}_{Y|\mathbf{x}}$. A natural question arise: is there any relation between $\mathbf{B}_0$ and $\mathbf{U}_0$? We assume the following linearity condition to answer this question.

(C5) Assume $E(\mathbf{x}_{i,j} \mid \mathbf{x}_{i,j}^{\top}\mathbf{U}_0)$ is linear in $(\mathbf{x}_{i,j}^{\top}\mathbf{U}_0)$.

Proposition 1. *Assume Σ is non-singular and (C5) holds. Then $span(\mathbf{B}_0) \subseteq span(\mathbf{U}_0)$, where $span(\mathbf{A})$ stands for the column space of* $\mathbf{A}$.

Proof of Proposition 1. We follow similar arguments used in Theorem 2.1 of Li and Duan [35]. Let $\mathbf{P}_{\mathbf{U}_0}(\Sigma) \overset{\text{def}}{=} \mathbf{U}_0(\mathbf{U}_0{}^{\top}\Sigma\mathbf{U}_0)^{-1}\mathbf{U}_0{}^{\top}\Sigma$ be the projection onto the linear space of $\mathbf{U}_0$ under the inner product induced by Σ. We have

$$\begin{aligned}\mathbf{B}_0 &= \Sigma^{-1}\sum_{j=1}^{m} E(\mathbf{x}_j\mathbf{y}_j{}^{\top})/m = \Sigma^{-1}\sum_{j=1}^{m} E\{E(\mathbf{x}_j \mid \mathbf{y}_j)\mathbf{y}_j{}^{\top}\}/m = \Sigma^{-1}\sum_{j=1}^{m} E[E\{E(\mathbf{x}_j \mid \mathbf{x}_j{}^{\top}\mathbf{U}_0) \mid \mathbf{y}_j\}\mathbf{y}_j{}^{\top}]/m \\ &= \Sigma^{-1}\mathbf{P}_{\mathbf{U}_0}{}^{\top}(\Sigma)\sum_{j=1}^{m} E\{E(\mathbf{x}_j \mid \mathbf{y}_j)\mathbf{y}_j{}^{\top}\}/m = \mathbf{U}_0(\mathbf{U}_0{}^{\top}\Sigma\mathbf{U}_0)^{-1}\mathbf{U}_0{}^{\top}\sum_{j=1}^{m} E(\mathbf{x}_j\mathbf{y}_j{}^{\top})/m.\end{aligned}$$

The first equality follows from the definition of $\mathbf{B}_0$ in (2), the second owes to the law of total expectation, the third holds because $F_j(Y_j \mid \mathbf{x}_j) = F_j(Y_j \mid \mathbf{x}_j{}^{\top}\mathbf{U}_0)$, for all $j \in \{1, \dots, m\}$, and the fourth is a direct consequence of the linearity List (C5). The last yields the desired result directly. □

We remark here that Proposition 1 holds for arbitrary transformations, which includes Li [32], Fung et al. [23], Yin and Cook [60] Zhu et al. [72] and Bi and Qu [5] as special cases. Proposition 1 also implies that $\mathbf{B}_0$ is a rank-deficient matrix because d_0, by definition, is the structural dimension of $\mathcal{S}_{Y|\mathbf{x}}$. The linearity condition is widely assumed in the context of sufficient dimension reduction and is typically regarded as mild, which is satisfied if the covariates follow normal or elliptically contour distributions [32]. Hall and Li [25] show that this linearity condition always provides a good approximation to reality as long as d_0 is small and p is sufficiently large. Interested readers refer to Diaconis and Freedman [19], Zhu and Zhu [71] and Ma and Zhu [42] for additional discussions.

Proposition 1 ensures that the distributed algorithm introduced in Section 2.2 can be readily used to recover $\mathcal{S}_{Y|\mathbf{x}}$ and estimate d_0 simultaneously, even when the massive data are scattered at different locations. In addition, the theoretical properties derived in Section 2.3 do carry over in the present context of sufficient dimension reduction.

The above methodology can be readily adapted to account for multivariate responses. To be specific, we let $\mathbf{y}_{i,j} = (Y_{i,j,1}, \dots, Y_{i,j,q})^{\top} \in \mathbb{R}^q$ be the q-vector of responses in the multivariate case. We adapt the transformation functions as follows. To generalize the classic slicing transformation [32], we first divide the response space $\mathbb{R}^q$ into K disjoint subspaces, $\{\mathcal{C}_k : k \in \{1, \dots, K\}\}$, which satisfy that $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_K = \mathbb{R}^q$. We then simply define $\mathbf{t}_k(\mathbf{y}_{i,j}) = I(\mathbf{y}_{i,j} \in \mathcal{C}_k)$. To generalize the polynomial transformation [60], we can simply define $\mathbf{t}_k(\mathbf{y}_{i,j}) = \mathbf{y}_{i,j}^{\otimes k}$, where $\otimes$ stands for the Kronecker product. The Fourier transformation [72] can be readily used in the multivariate case. With either of the above transformation functions, we define $\mathbf{y}_{i,j,k} = \mathbf{t}_k(\mathbf{y}_{i,j}) - E\{\mathbf{t}_k(\mathbf{y}_{i,j})\}$. The above methodology can be applied to the multivariate case by simply replacing $Y_{i,j,k}$s with $\mathbf{y}_{i,j,k}$s. More importantly, the desirable properties that we derive for the univariate case carry over to the multivariate one. One can also use an ensemble approach that aggregates the solutions obtained from the univariate cases. However, this ensemble approach requires implicitly that $\mathcal{S}_{\mathbf{y}|\mathbf{x}} = \bigcup \mathcal{S}_{Y_k|\mathbf{x}}$, for $\mathbf{y} = (Y_1, \dots, Y_q)^\top$, which is not always true in practice.

4. Simulations

We evaluate the performance of our distributed algorithm and compare it with classic algorithms which requires to pool all observations together in terms of recovering $\mathcal{S}_{Y|\mathbf{x}}$ and estimating d_0.

Throughout we generate $\mathbf{x}_{i,j}$s from multivariate normal with mean zero and covariance matrix $\Sigma = (\rho^{|k-l|})_{p\times p}$, where $p = 30$, and $\rho \in \{0.2, 0.5, 0.8\}$. Let $\boldsymbol{\beta}_1 = (1, 1, 0, \dots, 0)^\top \in \mathbb{R}^p$ and $\boldsymbol{\beta}_2 = (0, 0, 1, 1, 0, \dots, 0)^\top \in \mathbb{R}^p$. The errors $\varepsilon_{i,j}$s are standard normal. We consider three models:

$$Y_{i,j} = (1 + 2\mathbf{x}_{i,j}{}^\top\boldsymbol{\beta}_1)/\{0.5 + (1.5 + \mathbf{x}_{i,j}{}^\top\boldsymbol{\beta}_2)^2\} + 6\varepsilon_{i,j}, \quad Y_{i,j} = \sin(\mathbf{x}_{i,j}{}^\top\boldsymbol{\beta}_1) + \exp(\mathbf{x}_{i,j}{}^\top\boldsymbol{\beta}_2 + 6\varepsilon_{i,j}),$$
$$Y_{i,j} = (\mathbf{x}_{i,j}{}^\top\boldsymbol{\beta}_1)\exp(\mathbf{x}_{i,j}{}^\top\boldsymbol{\beta}_2) + 6\varepsilon_{i,j}.$$

Within each local machine we generate $Y_{i,j}$s randomly from one of the above models. In other words, the observations are heterogeneous. The central subspace $\mathcal{S}_{Y|\mathbf{x}}$ is spanned by $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and its structural dimension $d_0 = 2$. We vary $N \in \{12000, 24000, 36000\}$, $n \in \{400, 600, 800\}$ at each local machine, and replicate each experiment 1000 times to report the simulated results.

4.1. Estimation of structural dimension d_0

We compare our proposed distributed algorithm introduced in Section 2.2 with its classic counterparts which require to pool all observations in a single computer. The competitors include the classic reduced rank algorithm reviewed in Section 2.1 [10], the sequential test [32] and the ladle estimate [39]. For the purposes of fair comparison, all methods are built upon sliced inverse regression (SIR for short), which is the first and perhaps the most popular method in the context of sufficient dimension reduction. In particular, we choose the transformation functions $T_k(Y_{i,j}) = I(b_{k-1} \le Y_{i,j} < b_k)$, where b_ks are the $(k/q) \times 100\%$ percentile of $Y_{i,j}$s. This corresponds to the classic sliced inverse regression method. Upon request by an anonymous reviewer, we vary the number of slices $q \in \{5, 10, 20\}$ in our simulation studies. Li [32] first observes that the performance of sliced inverse regression is very robust to the number of slices in the empirical studies. This observation is confirmed by thorough theoretical investigations in Hsing and Carroll [27] and Zhu and Ng [69]. Let $Y_{i,j,k} = T_k(Y_{i,j}) - \overline{T}_k(Y_{i,j})$, for $k \in \{1, \dots, q\}$, and $\mathbf{y}_{i,j} = (Y_{i,j,1}, \dots, Y_{i,j,q})^\top$. We refer to our proposal as SIR (distributed), and the aforementioned competitors as SIR (pooled), SIR (sequential) and SIR (ladle), respectively. Following Chen et al. [10], we fix $\gamma = 10$ in SIR (distributed) and SIR (pooled), and choose λ through cross-validation. In Table 1 we report the percentages of correctly estimating the structural dimension d_0.

It is not surprising to see from Table 1 that, all methods are very insensitive to the number of slices. The SIR (pooled) performs the best, followed by the SIR (distributed). Their performances are comparable, which echoes our theoretical investigations. Both are superior to other competitors. The SIR (ladle) is very sensitive to ρ. Its performance deteriorates sharply when ρ is increased from 0.5 to 0.8.

4.2. Recovery of the central subspace $\mathcal{S}_{Y|\mathbf{x}}$

Next we evaluate the performance of our proposal in terms of recovering $\mathcal{S}_{Y|\mathbf{x}}$. We compared the SIR (distributed) and the SIR (pooled) with the following competitors: the classic SIR [32], the SIR method implemented with a divide and conquer procedure [38,58] and the distributed principal component analysis [20], which are referred to as the SIR (classic), the SIR (div-conq) and the PCA (distributed), respectively. Only the SIR (distributed) and the SIR (pooled) can estimate $\mathcal{S}_{Y|\mathbf{x}}$ and d_0 simultaneously, while other methods estimate $\mathcal{S}_{Y|\mathbf{x}}$ and d_0 separately. We use the trace correlation [21] to assess their performances. The averages of trace correlations are reported in Table 2 based on 1000 repetitions.

The simulated results once again indicate that the estimates obtained with sliced inverse regression is very robust to the number of slices. Therefore, we simply choose $q = 5$ in subsequent simulations. All supervised learners improves as N increases, or ρ decreases. The performances of the SIR (distributed) and the SIR (pooled) are comparable, particularly when n is relatively large, which again confirms our theoretical results in Section 2. The SIR (classic) is much worse than the SIR (distributed) and the SIR (pooled), because the SIR (classic) does not use the low-rank structure in estimating the central subspace. The SIR (div-conq) performs the worst among the supervised learners, particularly because $\Sigma = \mathrm{cov}(\mathbf{x}_{i,j}, \mathbf{x}_{i,j}{}^\top)$ is nearly singular when $\rho = 0.8$, indicating that the divide-and-conquer procedure is perhaps suboptimal in such situations.

It is not surprising that the methods built upon sliced inverse regression are all superior to the distributed principal component analysis. The sliced inverse regression is a supervised learner, which performs dimension reduction with

Table 1
For each method, the number in each cell is the percentage of correctly estimating the structural dimension d_0 out of 1000 replicates. We vary the sample size $N \in \{12000, 24000, 36000\}$, the local sample size at each machine $n \in \{400, 600, 800\}$ and the number of slices $q \in \{5, 10, 20\}$ for different correlation parameters ρ in the covariance matrix $\Sigma = (\rho^{|k-l|})_{p\times p}$.

		N								
		12000			24000			36000		
		n			n			n		
		400	600	800	400	600	800	400	600	800
	q					$\rho = 0.2$				
SIR(sequential)	5	75.0	74.8	72.4	78.4	75.8	79.7	75.4	78.0	79.4
	10	77.8	76.5	73.7	79.0	78.2	80.8	79.4	78.4	81.1
	20	74.6	76.0	73.1	77.3	76.1	78.2	75.7	76.7	80.6
SIR(ladle)	5	57.2	56.0	52.3	58.5	55.4	56.4	58.2	56.6	60.7
	10	56.7	55.2	51.9	58.0	55.7	56.1	57.8	56.5	60.7
	20	57.4	56.4	53.5	58.7	56.4	57.7	58.8	57.5	62.1
SIR(pooled)	5	99.3	97.0	93.9	100.0	99.8	98.7	100.0	99.9	99.7
	10	98.8	96.7	94.0	99.8	99.5	98.7	100.0	100.0	99.7
	20	97.4	94.0	92.2	99.8	99.4	98.4	100.0	100.0	99.7
SIR(distributed)	5	99.3	96.9	93.7	99.9	99.6	98.8	100.0	99.9	99.7
	10	98.1	96.1	93.7	99.8	99.5	98.7	100.0	100.0	99.7
	20	95.5	92.1	91.1	99.8	99.3	98.2	100.0	100.0	99.5
	q					$\rho = 0.5$				
SIR(sequential)	5	71.6	71.9	68.8	74.1	72.3	75.1	72.4	72.0	74.7
	10	71.7	72.8	71.5	75.8	75.5	75.3	75.3	74.3	77.0
	20	70.7	72.5	68.8	74.0	73.0	74.3	73.3	73.9	76.2
SIR(ladle)	5	54.0	52.5	48.2	50.2	48.7	50.5	46.7	47.0	49.1
	10	55.1	51.7	48.2	50.6	48.8	50.4	46.5	47.3	49.1
	20	54.7	54.2	50.4	51.2	50.7	52.0	48.7	49.0	50.5
SIR(pooled)	5	99.7	97.7	95.6	100.0	99.9	99.7	100.0	100.0	100.0
	10	99.2	97.9	95.8	100.0	100.0	99.7	100.0	100.0	100.0
	20	98.6	95.9	93.7	100.0	100.0	99.5	100.0	100.0	99.9
SIR(distributed)	5	99.4	97.4	95.1	99.9	99.8	99.6	100.0	100.0	99.9
	10	98.7	97.1	95.2	99.9	100.0	99.6	100.0	100.0	99.9
	20	96.2	94.4	93.0	100.0	99.9	99.4	100.0	100.0	99.9
	q					$\rho = 0.8$				
SIR(sequential)	5	59.4	58.6	59.1	59.7	59.9	61.5	59.3	58.3	62.1
	10	61.2	61.7	61.7	63.6	61.6	63.9	62.1	64.0	65.6
	20	58.4	57.0	57.5	57.7	59.0	60.5	61.5	60.3	61.6
SIR(ladle)	5	18.9	16.6	17.4	17.3	17.8	20.2	14.6	17.3	18.6
	10	18.6	16.4	17.3	17.1	17.4	20.2	15.5	16.8	19.2
	20	20.6	17.8	18.2	19.3	18.9	22.0	15.7	18.7	19.5
SIR(pooled)	5	93.0	89.1	84.6	98.5	96.1	93.3	99.5	97.9	95.8
	10	95.3	91.1	88.0	99.9	98.4	97.0	100.0	99.8	99.2
	20	94.7	89.2	84.3	99.9	98.3	96.5	100.0	99.8	98.9
SIR(distributed)	5	90.9	88.1	84.1	94.2	94.5	92.8	96.4	95.9	95.2
	10	94.2	90.1	86.9	99.7	98.1	96.8	99.9	99.6	99.0
	20	93.1	88.5	82.7	99.8	98.0	96.3	100.0	99.7	98.8

the aid of $Y_{i,j}$s. The distributed principal component analysis is an unsupervised learning method, which completely ignores the information of $Y_{i,j}$s. In the particular simulated examples, it is indeed not sufficient to capture the majority of variations in the covariates with merely two principal components. To elaborate this issue, we plot the Pareto chart of the eigenvalues of Σ in Fig.1 for $\rho \in \{0.2, 0.5, 0.8\}$, respectively. It can be clearly seen from Fig.1 that, the larger ρ is, the more concentrated the eigenvalues of Σ are. This partly explains the PCA (distributed) with $\rho = 0.8$ performs slightly better than that with $\rho = 0.2$ or 0.5. Even with $\rho = 0.8$, the PCA (distributed) requires more than 10 principle components to capture over 80% variability in $\mathbf{x}_{i,j}$s. The first two principal components only capture at most 50% variability in $\mathbf{x}_{i,j}$s. By contrast, sliced inverse regression targets to find the most predictive linear combinations of $\mathbf{x}_{i,j}$s. Bearing this target in mind, it suggests to use two linear combinations to predict $Y_{i,j}$.

4.3. Sensitivity analysis of the regularization parameter γ

We conduct a sensitivity analysis of the regularization parameter γ for the SIR (distributed). For simplicity, we fix $n = 600$, vary $N \in \{12000, 24000, 36000\}$ and $\gamma \in \{6, \dots, 13\}$. We report the averages of trace correlations and those of the estimated dimensions obtained with the SIR (distributed) in Table 3. It can be clearly seen that, the SIR (distributed)

Table 2
For each method, the number in each cell is the average of trace correlations out of 1000 replicates. We vary the sample size $N \in \{12,000, 24,000, 36,000\}$, the local sample size at each machine $n \in \{400, 600, 800\}$ and the number of slices $q \in \{5, 10, 20\}$ for different correlation parameters ρ in the covariance matrix $\Sigma = (\rho^{|k-l|})_{p\times p}$. All numbers reported are multiplied by 100.

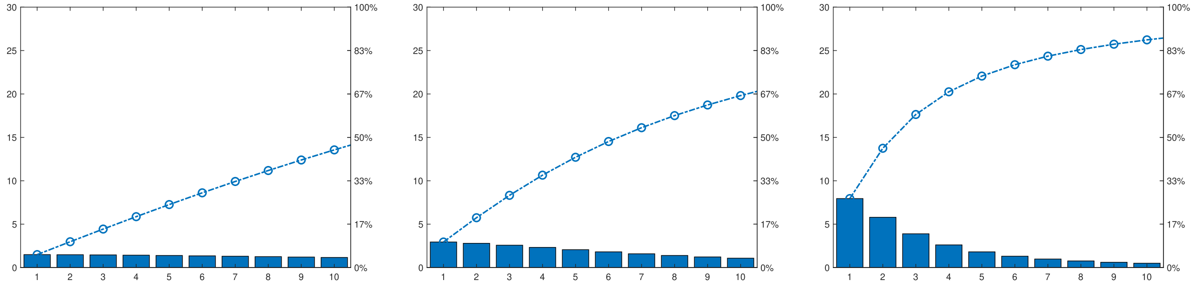
		N								
		12000			24000			36000		
		n			n			n		
		400	600	800	400	600	800	400	600	800
	q					$\rho = 0.2$				
SIR(distributed)	5	89.61	88.64	87.55	94.67	94.63	94.16	96.33	96.48	96.37
	10	91.99	91.17	90.08	96.06	96.13	95.68	97.27	97.54	97.41
	20	92.25	90.89	89.90	96.51	96.54	95.94	97.63	97.89	97.66
SIR(pooled)	5	90.00	88.86	87.68	95.06	94.85	94.24	96.64	96.59	96.46
	10	92.62	91.62	90.32	96.38	96.26	95.79	97.61	97.64	97.50
	20	93.01	91.49	90.61	96.91	96.75	96.16	97.98	98.01	97.86
SIR(classic)	5	73.22	72.71	71.51	78.38	77.59	78.89	79.62	80.01	81.50
	10	78.28	77.95	76.65	82.86	81.73	83.18	83.44	83.80	85.51
	20	78.98	78.98	77.31	83.43	82.62	83.80	84.18	84.49	86.41
SIR(div-conq)	5	89.23	90.62	90.65	94.42	95.20	95.39	96.24	96.78	96.97
	10	90.69	92.10	91.96	95.35	96.08	96.24	96.88	97.39	97.54
	20	90.19	91.71	91.80	95.15	96.07	96.29	96.80	97.42	97.60
PCA(distributed)	–	7.43	7.57	7.36	6.67	6.97	6.89	6.06	6.16	6.22
	q					$\rho = 0.5$				
SIR(distributed)	5	86.41	85.62	84.62	92.70	92.90	92.66	94.85	95.23	95.18
	10	89.56	88.91	88.00	94.46	94.86	94.60	96.01	96.48	96.46
	20	90.17	89.29	88.52	95.10	95.42	95.10	96.47	96.94	96.89
SIR(pooled)	5	86.94	85.91	84.97	93.23	93.06	92.81	95.34	95.35	95.31
	10	90.10	89.49	88.38	94.94	94.98	94.75	96.57	96.58	96.60
	20	91.06	89.69	88.82	95.61	95.63	95.27	97.06	97.05	96.99
SIR(classic)	5	67.14	66.40	65.45	72.45	71.63	72.77	73.62	73.74	75.55
	10	72.75	72.08	70.99	77.51	76.64	77.99	78.42	78.45	80.07
	20	73.40	73.08	71.98	78.38	77.88	78.80	79.53	79.52	81.34
SIR(div-conq)	5	56.55	64.79	66.54	73.52	82.90	85.21	82.42	89.29	90.93
	10	58.61	68.36	69.48	76.74	85.46	87.11	84.95	91.13	92.43
	20	49.48	62.50	66.24	67.71	81.37	84.30	78.01	88.88	91.33
PCA(distributed)	–	6.12	6.20	6.04	6.13	6.13	6.13	6.13	6.06	6.07
	q					$\rho = 0.8$				
SIR(distributed)	5	68.98	68.02	66.57	81.03	81.52	80.78	86.26	86.79	86.51
	10	74.29	73.15	71.86	86.45	86.11	85.56	90.23	90.75	90.59
	20	75.24	73.94	71.54	87.51	87.15	86.36	91.19	91.72	91.30
SIR(pooled)	5	70.35	68.78	67.16	83.67	82.56	81.29	88.73	87.98	87.07
	10	75.40	73.96	72.74	87.17	86.44	85.83	91.06	90.95	90.77
	20	76.59	74.64	72.68	88.27	87.51	86.65	92.01	91.88	91.46
SIR(classic)	5	49.22	48.62	48.48	55.38	54.78	56.24	57.41	57.90	59.20
	10	54.83	54.08	54.08	61.08	60.74	62.51	63.37	63.84	65.18
	20	55.11	54.72	54.82	61.90	61.46	62.95	64.24	64.48	66.09
SIR(div-conq)	5	5.69	9.41	12.79	6.46	12.03	16.81	6.44	13.76	21.26
	10	5.09	8.95	12.68	5.25	10.81	16.72	5.14	12.07	21.27
	20	3.75	6.26	9.44	3.48	6.97	11.68	3.23	7.47	14.04
PCA(distributed)	–	10.07	10.08	10.07	10.08	10.08	10.08	10.11	10.09	10.08

has a very stable performance for a wide range of γ. The averages of the estimated dimensions, $\widehat{d}$, match the underlying true dimension d_0 with high probability, indicating that the SIR (distributed) is robust for a wide range of γ.

Next we elaborate this sensitivity analysis under the Bayesian information criterion. To be precise, we vary the regularization parameter $\gamma \in \{0, \ldots, 16\}$, and plot the Bayesian information criterion versus γ in Fig. 2. It can be clearly seen once again that, the Bayesian information criterion is relatively stable for $\gamma \in \{2, \ldots, 10\}$.

4.4. Comparison of computational efficiency

We compare the computational efficiency of different proposals. In Table 4, we vary $n \in \{400, 600, 800\}$ and $N \in \{12000, 24000, 36000\}$, and report the CPU time of the SIR (distributed), the SIR (pooled), the SIR (classic), the SIR (div-conq) and the PCA (distributed) based on 1000 replications. All methods are implemented with Matlab (Version R2020a) and conducted on a server with 64-core Intel(R) Xeon(R) Gold 6148 CPU (2.40 GHz) and 252 GB RAM. We implement the three distributed algorithms in a fully synchronized distributed setting.



(A): ρ = 0.2

(B): ρ = 0.5

(C): ρ = 0.8

Fig. 1. The eigenvalues of Σ are sorted in a descending order. The horizontal axis stands for the indexes of the eigenvalues, the left vertical axis represents the magnitude of the eigenvalues and the right vertical axis is the cumulative distribution of the eigenvalues.

Table 3
The averages of trace correlations (corr^2) and those of the estimated dimensions obtained with the SIR (distributed) out of 1000 replicates. We vary the sample size $N \in \{12000, 24000, 36000\}$ and the regularization parameter $\gamma \in \{6, \dots, 13\}$ for different correlation parameters ρ in the covariance matrix $\Sigma = (\rho^{|k-l|})_{p\times p}$.

ρ	γ	6		7		8		9		10		11		12		13	
	N	corr^2	$\widehat{d}$	corr^2	$\widehat{d}$	corr^2	$\widehat{d}$	corr^2	$\widehat{d}$	corr^2	$\widehat{d}$	corr^2	$\widehat{d}$	corr^2	$\widehat{d}$	corr^2	$\widehat{d}$
	12000	0.89	2.19	0.89	2.12	0.89	2.02	0.89	1.98	0.89	1.98	0.89	1.97	0.88	1.95	0.87	1.92
0.2	24000	0.95	2.12	0.95	2.00	0.95	2.00	0.95	2.00	0.94	1.99	0.94	1.99	0.94	1.98	0.93	1.96
	36000	0.96	2.01	0.96	2.00	0.96	2.00	0.96	2.00	0.96	2.00	0.96	2.00	0.96	1.99	0.95	1.97
	12000	0.86	2.16	0.86	2.12	0.86	2.01	0.86	1.99	0.86	1.98	0.86	1.97	0.86	1.97	0.85	1.95
0.5	24000	0.93	2.07	0.93	2.00	0.93	2.00	0.93	2.00	0.93	2.00	0.93	2.00	0.93	1.99	0.92	1.98
	36000	0.95	2.00	0.95	2.00	0.95	2.00	0.95	2.00	0.95	2.00	0.95	2.00	0.95	2.00	0.95	1.99
	12,000	0.70	2.01	0.70	1.96	0.70	1.93	0.70	1.92	0.68	1.88	0.65	1.77	0.57	1.53	0.50	1.32
0.8	24000	0.83	2.00	0.83	1.98	0.83	1.98	0.83	1.98	0.81	1.94	0.74	1.76	0.62	1.46	0.51	1.18
	36000	0.89	2.00	0.89	2.00	0.89	2.00	0.88	1.99	0.87	1.96	0.78	1.77	0.62	1.41	0.51	1.14

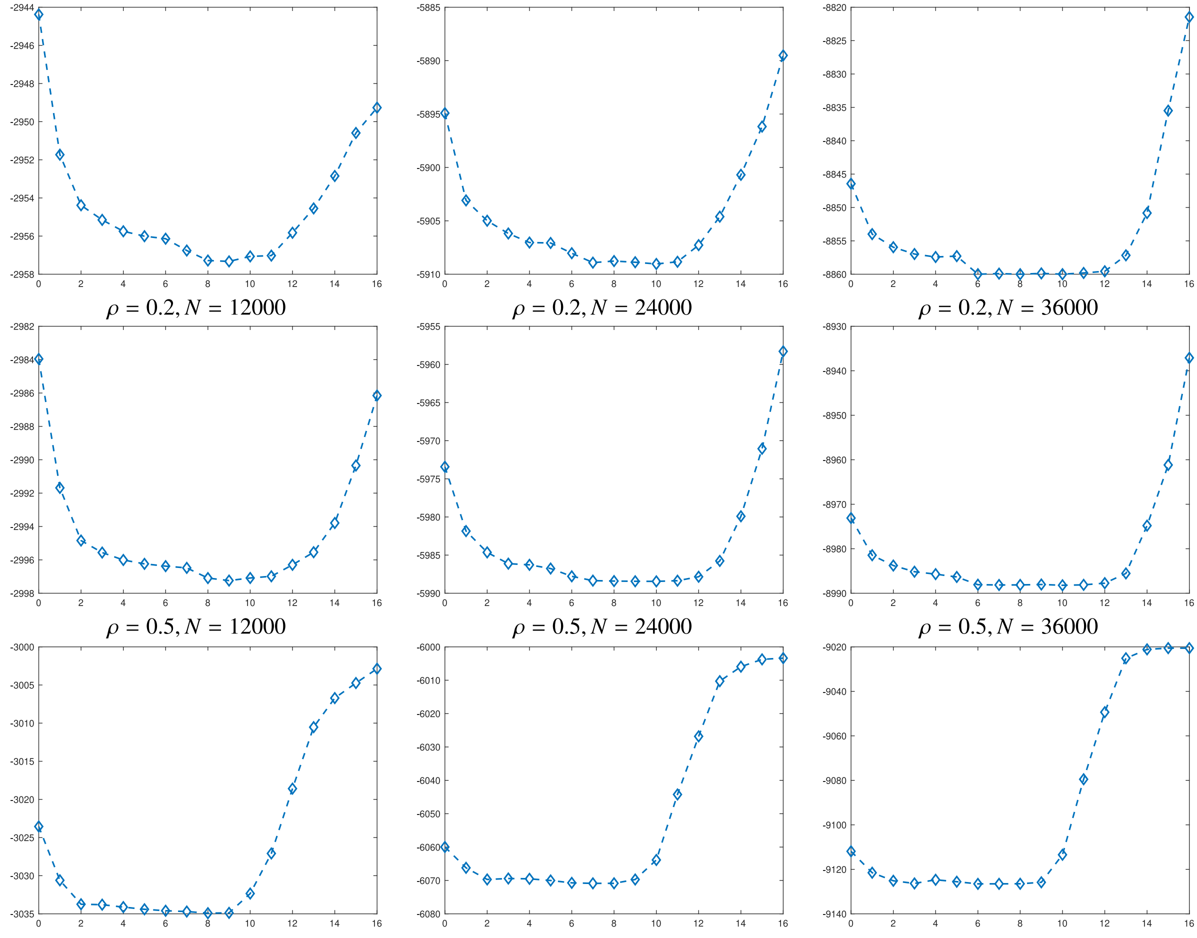
Table 4
For each method, the number in each cell is the average of the CPU time (in seconds) out of 1000 replicates. We vary the sample size $N \in \{12000, 24000, 36000\}$ and the local sample size at each machine $n \in \{400, 600, 800\}$. All numbers reported below are multiplied by 100.

	N								
	12000			24000			36000		
	n			n			n		
	400	600	800	400	600	800	400	600	800
SIR(distributed)	0.16	0.13	0.13	0.23	0.15	0.14	0.20	0.15	0.16
SIR(pooled)	0.36	0.41	0.43	1.01	0.92	0.91	1.40	1.47	1.47
SIR(classic)	1.39	1.51	1.56	3.63	3.36	3.33	4.57	4.71	4.76
SIR(div-conq)	3.11	2.88	2.67	8.54	6.58	6.11	10.97	9.83	9.37
PCA(distributed)	0.87	0.81	0.70	2.29	1.73	1.48	3.04	2.65	2.32

It can be clearly seen from Table 4 that the SIR (distributed) is computationally much faster than the SIR (pooled), although both have comparable performance in terms of estimating the central subspace. This echoes our analysis of the computational complexity in Section 2.3.

5. An application

We apply our proposed distributed algorithm to the airline on-time performance dataset that is provided by the 2009 ASA Data Expo and available at https://doi.org/10.7910/DVN/HG7NV7. This dataset consists of the arrival and departure information of 118,914,458 flights in the USA, during the period of October 1987 to April 2008. It occupies 12 GB of storage space in a hard drive. We remove all observations with missing values, leaving N = 79,674,837 observations. The response variable is the magnitude of arrival delays (ArrDelay, in minutes). There are p = 11 covariates: the year of the flight (Year), the scheduled departure time (CRSDepTime, in hhmm), the actual elapsed time (ActualElapsedTime, in minutes), the distance from origin to destination (Distance, in miles), the departure delay (DepDelay, in minutes), the scheduled elapsed time (CRSElapsedTime, in minutes), the actual departure time (DepTime, in hhmm), the actual arrival time (ArrTime, in hhmm), the scheduled arrival time (CRSArrTime, in hhmm), the taxi in time (TaxiIn, in minutes) and the taxi out time (TaxiOut, in minutes). All the covariates have been standardized.



$\rho = 0.2, N = 12000$ $\rho = 0.2, N = 24000$ $\rho = 0.2, N = 36000$

$\rho = 0.5, N = 12000$ $\rho = 0.5, N = 24000$ $\rho = 0.5, N = 36000$

$\rho = 0.8, N = 12000$ $\rho = 0.8, N = 24000$ $\rho = 0.8, N = 36000$

Fig. 2. The horizontal axis stands for the regularization parameter γ and the vertical axis represents the Bayesian information criterion of our distributed estimate for different combinations of the correlation parameter ρ and the sample size N.

We use the same transformation functions as in Section 4 with the number of slices $q = 5$. We fix $\gamma = 10$ in the SIR (distributed) and the SIR (pooled) methods, and choose λ through cross-validation. We randomly distribute the observations over $m \in \{1, 10, 100, 500, 1000\}$ machines, and apply the SIR (distributed) method to this dataset.

To implement our proposal it is required that all $\mathbf{x}_{i,j}$s have identical mean and mutual covariance matrix. To test if these requirements are satisfied, we denote $E(\mathbf{x}_{i,j}) = \mathbf{u}_j$ and $\mathrm{cov}(\mathbf{x}_{i,j}) = \Sigma_j$, for $j \in \{1, \ldots, m\}$. It is required to test

$$H_0 : \mathbf{u}_1 = \cdots = \mathbf{u}_m, \;\; \Sigma_1 = \cdots = \Sigma_m \text{ versus } \; H_1 : \text{otherwise}.$$

This is a classic hypothesis test in multivariate statistical analysis [2,28,34,64]. We simply apply the test procedure proposed by [28] to the airline on-time performance dataset. The resulting p-values are 0.5106, 0.7905, 0.6304 and 0.6361, respectively, for $m \in \{10, 100, 500, 1000\}$, indicating that we have no strong evidence to reject the null hypothesis. In other words, the requirements are likely satisfied in this study.

Distinctive m values yield the same estimate $\widehat{d} = 1$, which complies with Liquet and Saracco [38] where the structural dimension is determined based on a random subsample by the ad hoc "elbow rule". We remark here that the SIR (distributed) method with $m = 1$ reduces to the SIR (pooled) method. However, the computational efficiency of these methods are quite different. The SIR (pooled) method takes around 6134.02 s. By contrast, the SIR (distributed) method with $m \in \{10, 100, 500, 1000\}$ takes 1603.18, 443.40, 329.28 and 307.13 s, respectively. Their trace correlations with SIR (pooled) are all larger than 0.9999. This indicates the SIR (distributed) method reduces the computational complexity greatly.

The SIR (distributed) method with $m = 500$ identifies ActualElapsedTime and DepDelay as two most influential factors to ArrDelay, whose coefficients are the largest among all. We plot the responses versus the dimension-reduced covariates for the SIR (distributed) with $m = 500$ in Fig. 3. The plot exhibits an obvious pattern of heterogeneity: The observations

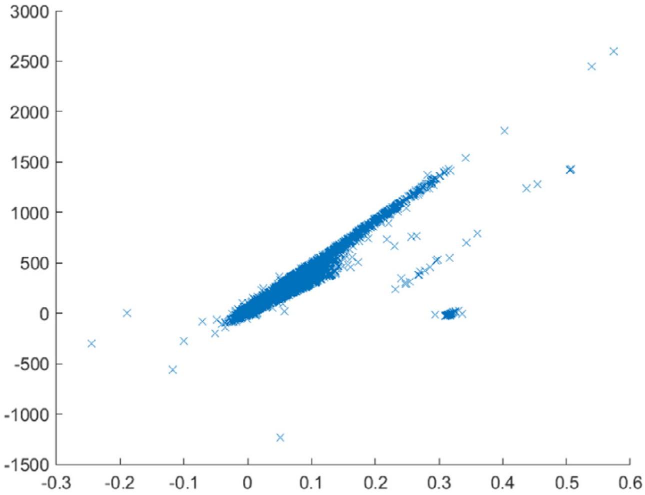


Fig. 3. The scatter plot of the reduced covariate versus the delay at arrival on the test set based on SIR (distributed) with $m = 500$.

are scattered along two parallel lines with different intercepts. Had we ignored this heterogeneity and fitted a single straight line, it would lead to a poor prediction.

6. Concluding remarks

In the present context of sufficient dimension reduction, we assume the transformation functions are known. This excludes the case where the transformations must be recovered in a data-driven fashion. Should we estimate the transformations parametrically or non-parametrically at each local machine, substantial errors would be induced, particular when the number of transformations, q, is extremely large. How these errors affect the resultant distributed estimates remains unknown. This question is open yet deserves thorough investigations in our future studies.

CRediT authorship contribution statement

Canyi Chen: Methodology, Software, Investigation. **Wangli Xu:** Methodology, Writing – original draft. **Liping Zhu:** Conceptualization, Supervision, Writing – review & editing.

Acknowledgments

This research is supported by National Natural Science Foundation of China (12171477, 11731011, 11931014, 11971478). The authors contribute equally to this paper and their names are listed in an alphabetical order.

References

[1] T.W. Anderson, Estimating linear restrictions on regression coefficients for multivariate normal distributions, Ann. Math. Stat. 22 (3) (1951) 327–351.
[2] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Third Edition, New York: Wiley, 2003.
[3] Moulinath Banerjee, Cécile Durot, Bodhisattva Sen, Divide and conquer in nonstandard problems and the super-efficiency phenomenon, Ann. Statist. 47 (2) (2019) 720–757.
[4] Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, Ziwei Zhu, Distributed testing and estimation under sparse high dimensional models, Ann. Statist. 46 (3) (2018) 1352–1382.
[5] Xuan Bi, Annie Qu, Sufficient dimension reduction for longitudinal data, Statist. Sinica 25 (2) (2015) 787–807.
[6] Florentina Bunea, Yiyuan She, Marten H. Wegkamp, Optimal selection of reduced rank estimators of high-dimensional matrices, Ann. Statist. 39 (2) (2011) 1282–1309.
[7] E. Bura, J. Yang, Dimension estimation in sufficient dimension reduction: A unifying approach, J. Multivariate Anal. 102 (1) (2011) 130–142.
[8] Jian-Feng Cai, Emmanuel J. Candès, Zuowei Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (4) (2010) 1956–1982.
[9] Emmanuel J. Candès, Benjamin Recht, Exact matrix completion via convex optimization, Found. Comput. Math. 9 (6) (2009) 717.
[10] K. Chen, H. Dong, K.-S. Chan, Reduced rank regression via adaptive nuclear norm penalization, Biometrika 100 (4) (2013) 901–920.
[11] Xi Chen, Jason D. Lee, He Li, Yun Yang, Distributed estimation for principal component analysis: an enlarged eigenspace analysis, 2021, cs,stat.
[12] Xi Chen, Weidong Liu, Xiaojun Mao, Zhuoyi Yang, Distributed high-dimensional regression under a quantile loss function, J. Mach. Learn. Res. 21 (2020) 1–43.
[13] Xi Chen, Weidong Liu, Yichen Zhang, Quantile regression under memory constraint, Ann. Statist. 47 (6) (2019) 3244–3273.
[14] Xi Chen, Weidong Liu, Yichen Zhang, First-order newton-type estimator for distributed estimation and inference, 2021, cs,stat.

[15] R. Dennis Cook, Regression Graphics: Ideas for Studying Regressions Through Graphics, First, in: Wiley Series in Probability and Statistics. Probability and Statistics Section, Wiley, New York and Chichester, 1998.
[16] R. Dennis Cook, Bing Li, Dimension reduction for conditional mean in regression, Ann. Statist. 30 (2) (2002) 455–474.
[17] R. Dennis Cook, Liqiang Ni, Sufficient dimension reduction via inverse regression: A minimum discrepancy approach, J. Am. Stat. Assoc. 100 (470) (2005) 410–428.
[18] R. Dennis Cook, Sanford Weisberg, Sliced inverse regression for dimension reduction: comment, J. Am. Stat. Assoc. 86 (414) (1991) 328–332.
[19] Persi Diaconis, David Freedman, Asymptotics of graphical projection pursuit, Ann. Statist. 12 (3) (1984) 793–815.
[20] Jianqing Fan, Dong Wang, Kaizheng Wang, Ziwei Zhu, Distributed estimation of principal eigenspaces, Ann. Statist. 47 (6) (2019) 3009–3031.
[21] Louis Ferré, Determining the dimension in sliced inverse regression and related methods, J. Am. Stat. Assoc. 93 (441) (1998) 132–140.
[22] Joel N. Franklin, Matrix Theory, Dover Publications, Mineola, N.Y, 2000.
[23] Wing Kam Fung, Xuming He, Li Liu, Peide Shi, Dimension reduction based on canonical correlation, Statist. Sinica 12 (4) (2002) 1093–1113.
[24] Gene H. Golub, Charles F. Van Loan, Matrix Computations, Johns Hopkins Series in the Mathematical Sciences, (3) Johns Hopkins University Press, Baltimore, 1983.
[25] Peter Hall, Ker-Chau Li, On almost linearity of low dimensional projections from high dimensional data, Ann. Statist. 21 (2) (1993) 867–889.
[26] Roger A. Horn, Charles R. Johnson, Matrix analysis, 2nd ed, Cambridge University Press, Cambridge; New York, 2012.
[27] Tailen Hsing, Raymond J. Carroll, An asymptotic theory for sliced inverse regression, Ann. Statist. 20 (2) (1992) 1040–1061.
[28] Masashi Hyodo, Takahiro Nishiyama, Simultaneous testing of the mean vector and covariance matrix among k populations for high-dimensional data, Commun. Stat. - Theory Methods 50 (3) (2021) 663–684.
[29] Alan Julian Izenman, Reduced-rank regression for the multivariate linear model, J. Multivariate Anal. 5 (2) (1975) 248–264.
[30] Michael I. Jordan, Jason D. Lee, Yun Yang, Communication-efficient distributed statistical inference, J. Am. Stat. Assoc. 114 (526) (2019) 668–681.
[31] Jason D. Lee, Qiang Liu, Yuekai Sun, Jonathan E. Taylor, Communication-efficient sparse regression, J. Mach. Learn. Res. 18 (5) (2017) 1–30.
[32] Ker-Chau Li, Sliced inverse regression for dimension reduction, J. Am. Stat. Assoc. 86 (414) (1991) 316–327.
[33] Bing Li, Sufficient Dimension Reduction: Methods and Application with R, CRC Press, Taylor & Francis Group, Boca Raton, 2018.
[34] Jun Li, Song Xi Chen, Two sample tests for high-dimensional covariance matrices, Ann. Statist. 40 (2) (2012).
[35] Ker-Chau Li, Naihua Duan, Regression analysis under link violation, Ann. Statist. 17 (3) (1989) 1009–1052.
[36] Xingxiang Li, Runze Li, Zhiming Xia, Chen Xu, Distributed feature screening via componentwise debiasing, J. Mach. Learn. Res. 21 (2020) 32.
[37] Yingxing Li, Li-Xing Zhu, Asymptotics for sliced average variance estimation, Ann. Statist. 35 (1) (2007) 41–69.
[38] Benoit Liquet, Jerome Saracco, BIG-SIR: A sliced inverse regression approach for massive data, Stat. Interface 9 (4) (2016) 509–520.
[39] Wei Luo, Bing Li, Combining eigenvalues and variation of eigenvectors for order determination, Biometrika 103 (4) (2016) 875–887.
[40] Ronghua Luo, Hansheng Wang, Chih-Ling Tsai, Contour projected dimension reduction, Ann. Statist. 37 (6B) (2009) 3743–3778.
[41] Yanyuan Ma, Xinyu Zhang, A validated information criterion to determine the structural dimension in dimension reduction models, Biometrika 102 (2) (2015) 409–420.
[42] Yanyuan Ma, Liping Zhu, A semiparametric approach to dimension reduction, J. Am. Stat. Assoc. 107 (497) (2012) 168–179.
[43] Yanyuan Ma, Liping Zhu, Efficient estimation in sufficient dimension reduction, Ann. Statist. 41 (1) (2013) 250–268.
[44] Yanyuan Ma, Liping Zhu, A review on dimension reduction: *A review on dimension reduction*, Int. Stat. Rev. 81 (1) (2013) 134–150.
[45] Yanyuan Ma, Liping Zhu, On estimation efficiency of the central mean subspace, J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (5) (2014) 885–901.
[46] Rahul Mazumder, Trevor J. Hastie, Robert Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, J. Mach. Learn. Res. 11 (2010) 2287–2322.
[47] Jorma K. Merikoski, Ravinder Kumar, Inequalities for spreads of matrix sums and products., Appl. Math. E-Notes [electronic only] 4 (2004) 150–159.
[48] Gregory C. Reinsel, Rajabather Palani Velu, Multivariate Reduced-Rank Regression: Theory and Applications, in: Lecture Notes in Statistics, (136) Springer, New York, 1998.
[49] Angelika Rohde, Alexandre B. Tsybakov, Estimation of high-dimensional low-rank matrices, Ann. Statist. 39 (2) (2011) 887–930.
[50] James R. Schott, Determining the dimensionality in sliced inverse regression, J. Am. Stat. Assoc. 89 (425) (1994) 141–148.
[51] Chengchun Shi, Wenbin Lu, Rui Song, A massive data framework for M-estimators with cubic-rate, J. Am. Stat. Assoc. 113 (524) (2018) 1698–1709.
[52] Roman Vershynin, How close is the sample covariance matrix to the actual covariance matrix?, J. Theoret. Probab. 25 (3) (2012) 655–686.
[53] Roman Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, in: Cambridge Series in Statistical and Probabilistic Mathematics, vol. 47, Cambridge University Press, Cambridge, 2018.
[54] Xiaozhou Wang, Zhuoyi Yang, Xi Chen, Weidong Liu, Distributed inference for linear support vector machine, J. Mach. Learn. Res. 20 (2019) 113:1–113:41.
[55] Cheng Wang, Zhou Yu, Liping Zhu, On cumulative slicing estimation for high dimensional data, Statist. Sinica 31 (1) (2021) 223–242.
[56] Yingcun Xia, A constructive approach to the estimation of dimension reduction directions, Ann. Statist. 35 (6) (2007) 2654–2690.
[57] Yingcun Xia, Howell Tong, W.K. Li, Li-Xing Zhu, An adaptive estimation of dimension reduction space, J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (3) (2002) 363–410.
[58] Kelin Xu, Liping Zhu, Jianqing Fan, Distributed sufficient dimension reduction for heterogeneous massive data, 2021, p. 31, submitted.
[59] Zhishen Ye, Robert E. Weiss, Using the bootstrap to select one of a new class of dimension reduction methods, J. Am. Stat. Assoc. 98 (464) (2003) 968–979.
[60] Xiangrong Yin, R. Dennis Cook, Dimension reduction for the conditional kth moment in regression, J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (2) (2002) 159–175.
[61] Yang Yu, Shih-Kang Chao, Guang Cheng, Simultaneous inference for massive data: distributed bootstrap, in: International Conference on Machine Learning, PMLR, 2020, pp. 10892–10901.
[62] Y. Yu, T. Wang, R.J. Samworth, A useful variant of the davis–kahan theorem for statisticians, Biometrika 102 (2) (2015) 315–323.
[63] Ming Yuan, Ali Ekici, Zhaosong Lu, Renato Monteiro, Dimension reduction and coefficient estimation in multivariate linear regression, J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (3) (2007) 329–346.
[64] Chao Zhang, Zhidong Bai, Jiang Hu, Chen Wang, Multi-sample test for high-dimensional covariance matrices, Commun. Stat. - Theory Methods 47 (13) (2018) 3161–3177.
[65] Yuchen Zhang, Lin Xiao, Communication-efficient distributed optimization of self-concordant empirical loss, 2015, arXiv preprint cs,math,stat.
[66] Tianqi Zhao, Guang Cheng, Han Liu, A partially linear framework for massive heterogeneous data, Ann. Statist. 44 (4) (2016) 1400–1437.
[67] Tianqi Zhao, Mladen Kolar, Han Liu, A general framework for robust testing and confidence regions in high-dimensional quantile regression, 2015, stat.
[68] Lixing Zhu, Baiqi Miao, Heng Peng, On sliced inverse regression with high-dimensional covariates, J. Am. Stat. Assoc. 101 (474) (2006) 630–643.
[69] Li-Xing Zhu, Kai W. Ng, Asymptotics of sliced inverse regression, Statist. Sinica 5 (2) (1995) 727–736.
[70] Li-Ping Zhu, Zhou Yu, Li-Xing Zhu, A sparse eigen-decomposition estimation in semiparametric regression, Comput. Stat. Data Anal. 54 (4) (2010) 976–986.
[71] Li-Ping Zhu, Li-Xing Zhu, On distribution-weighted partial least squares with diverging number of highly correlated predictors, J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2) (2009) 525–548.
[72] Li-Ping Zhu, Li-Xing Zhu, Zheng-Hui Feng, Dimension reduction in regressions through cumulative slicing estimation, J. Am. Stat. Assoc. 105 (492) (2010) 1455–1466.
[73] Hui Zou, The adaptive lasso and its oracle properties, J. Am. Stat. Assoc. 101 (476) (2006) 1418–1429.