Annals of the Institute of Statistical Mathematics (2022) 74:435–450
https://doi.org/10.1007/s10463-021-00803-5

Robust distributed estimation and variable selection for massive datasets via rank regression

Jiaming Luan[1] · Hongwei Wang[1] · Kangning Wang[1] · Benle Zhang[1]

Received: 23 November 2020 / Revised: 22 March 2021 / Accepted: 6 May 2021 /
Published online: 20 June 2021

Abstract

Rank regression is a robust modeling tool; it is challenging to implement it for the distributed massive data owing to memory constraints. In practice, the massive data may be distributed heterogeneously from machine to machine; how to incorporate the heterogeneity is also an interesting issue. This paper proposes a distributed rank regression (DR^2), which can be implemented in the master machine by solving a weighted least-squares and adaptive when the data are heterogeneous. Theoretically, we prove that the resulting estimator is statistically as efficient as the global rank regression estimator. Furthermore, based on the adaptive LASSO and a newly defined distributed BIC-type tuning parameter selector, we propose a distributed regularized rank regression (DR^3), which can make consistent variable selection and can also be easily implemented by using the LARS algorithm on the master machine. Simulation results and real data analysis are included to validate our method.

Keywords Massive data · Robustness · Communication efficient · Variable selection

K. Wang: The authors are listed in the alphabetical order. The authors would like to thank Dr. Shaomin Li for his valuable suggestions. The authors would like to thank the editor, an associate editor and two anonymous reviewers for their constructive comments that led to a major improvement of this article. The research was supported by NNSF project of China (11901356, 11901149), wealth management project (2019ZBKY047) of Shandong Technology and Business University.

✉ Kangning Wang
wkn1986@126.com

[1] Shandong Technology and Business University, No. 191, Binhai Middle Road, Laishan District, Yantai 264005, China

1 Introduction

With the rapid development of technology, massive data are often encountered in both scientific fields and daily life. Such a large size of data is usually hard to be efficiently stored or dealt by one single computer; thus, they are distributed in many machines over limited memory, and a computer serves as the master, while all the other computers serve as workers. In such setting, due to the limited storage space in primary memory, it fails to deliver efficient estimator by standard algorithms or statistical packages.

In recent years, many methodologies and algorithms toward massive data analysis have been proposed. The first strategy is the one-shot (OS) method, which is one of the most important algorithms to deal with large-scale datasets. The idea of this method is to conduct the estimation on each machine to obtain a local estimate, and the final global estimate computed by the master machine is a simple average of the local ones. For example, Zhang et al., 2013 average the M-estimators obtained by node machines; Battey et al., 2018 average debiased estimators; and Fan et al., 2017 define an average for subspaces and compute it via eigen-decomposition. For more references, one can see (Zhang and Wang, 2007; Lin and Xi, 2011; Chen and Xie, 2014; Zhang et al., 2015; Zhu et al., 2019; Chen and Zhou, 2019) and the references therein. Because only one round of communication between machines is required in OS method, the communication costs are significantly reduced. The other approach includes iterative algorithms, in which multiple iterations are required so that the estimation efficiency can be refined to match the global estimator. For example, Wang et al., 2017 and Jordan et al., 2019 proposed the communication-efficient surrogate likelihood (CSL), Fan et al., 2019 also investigated two communication-efficient accurate statistical estimators. In these methods, optimization problems are solved in only one machine and other machines just evaluate gradients, thus effectively reducing the communication cost of processing large datasets. For more details, one can see Zhang et al. (2013) and Rosenblatt and Nadler (2016).

It should be noted that the aforementioned massive data statistical methods are mainly built on mean regression or likelihood framework. Although these approaches all enjoy the oracle property, i.e., they can estimate the unknown parameters as accurately as all the data were pooled on a single machine. However, as we all know, the mean regression and likelihood methods are not robust and can be adversely influenced by outliers or heavy-tail distributions. This will lead to shortcomings in robustness of such mean regression- or likelihood-based massive data analysis methods.

At first glance, natural alternatives seem to be the least absolute deviation (LAD) estimator (Wang et al., 2007) or quantile regression (Koenker and Bassett, 1978), which can be more robust, when the density function $f(\cdot)$ deviates from the normal or the outliers exist. Also, Chen and Zhou (2019) investigated the related massive data analysis methods. However, LAD and quantile regression all have limitations in terms of efficiency. For example, the efficiency of the LAD compared to the maximum likelihood is proportional to the density at the

median. For the Gaussian error case, the distribution of the greatest interest, this quantity is only 0.637. And, worse still, the efficiency can be arbitrarily small if $f(0)$ is close to zero. To achieve the balance between robustness and efficiency, one can further consider the rank regression (abbreviated to R^2), which is more robust than the mean regression or maximum likelihood estimators and more efficient than the LAD or quantile regression estimators. Many authors have used the R^2 to deal with different kinds of problems; for the recent works, one can see Wang and Li (2009), Wang et al. (2009), Shin (2010), Leng (2010), Feng et al. (2015) and so on.

However, to the best of our knowledge, when the dataset is massive and distributed in many machines, how to implement R^2 is completely unknown. What is more, the existing distributed estimators all assume that the datasets are stored in each machine independently and identically, while the distributed datasets may be heterogeneous from machine to machine in practice. Thus, incorporating the heterogeneity into the estimation procedure appropriately plays an important role in improving the efficiency. To solve these issues, we first propose a distributed rank regression (abbreviated to DR^2) estimator for the massive dataset. By regarding the local R^2 estimators obtained in different machines as an observed dataset and minimizing the approximate likelihood function, we transfer global R^2 into an asymptotically equivalent least-squares problem. Because DR^2 uses the distribution information of the data on each machine sufficiently, it can be adaptive when the datasets are stored heterogeneously. Also, the communication cost is reduced, since only one round of communication is involved and iteration is free. Theoretically, we prove that the resulting estimator is statistically as efficient as the global R^2 estimator. Furthermore, when the adaptive LASSO penalty (Zou 2006) and a new defined distributed BIC-type tuning parameter selector are used, we propose a distributed regularized rank regression (abbreviated to DR^3), which can select the relevant variables and estimate the coefficients simultaneously, and the solution path can also be easily obtained by using the LARS algorithm with minimal computation cost on the master machine. The new DR^2 and DR^3 have superiorities in robustness and efficiency by inheriting the advantage of the R^2.

The rest of this paper is organized as follows: Section 2 introduces the new method and asymptotical properties. Simulation results and real data analysis are reported in Sect. 3. Concluding remarks are discussed in Sect. 4. All the technical proofs and figures of simulation results are provided in the supplementary file.

2 Rank regression for massive datasets

2.1 A brief review on robust rank regression

Let $Y \in \mathbb{R}$ be a response variable, and $\boldsymbol{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ be a p-dimensional covariate vector, $\{\boldsymbol{X}_i, Y_i\}_{i=1}^N$ be N random samples. We consider the following linear regression model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is unknown parameter with true value $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^T$, ϵ_i is independent and identically distributed random error.

The rank regression estimates the unknown parameter $\boldsymbol{\beta}$ via the following optimization problem

$$\hat{\boldsymbol{\beta}}^{R^2} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{N(N-1)} \sum_{1 \leqslant i,j \leqslant N} |Y_{ij} - \boldsymbol{X}_{ij}^T \boldsymbol{\beta}| \right\}, \tag{1}$$

where $Y_{ij} = Y_i - Y_j$ and $\boldsymbol{X}_{ij} = \boldsymbol{X}_i - \boldsymbol{X}_j$. In fact, (1) can be regarded as a LAD regression by taking Y_{ij} and $\boldsymbol{X}_{ij}$ as the observations. Under some regularity conditions, $\hat{\boldsymbol{\beta}}^{R^2}$ admits the following asymptotic rule

$$\sqrt{N}(\hat{\boldsymbol{\beta}}^{R^2} - \boldsymbol{\beta}_0) \to_d N\Big(\mathbf{0}, \frac{1}{12\omega^2} \boldsymbol{\Sigma}^{-1}\Big), \tag{2}$$

where $\omega = \int f^2(t)dt$, $\boldsymbol{\Sigma} = E(\boldsymbol{X}\boldsymbol{X}^T)$ and $f(\cdot)$ is the density function of ϵ_i. The constant ω^2 indicates the height of the density of $Y_1 - Y_2$ at the origin. The asymptotic relative efficiency of $\hat{\boldsymbol{\beta}}^{R^2}$ to mean regression estimator $\hat{\boldsymbol{\beta}}^{MR}$ is $e(\hat{\boldsymbol{\beta}}^{R^2}, \hat{\boldsymbol{\beta}}^{MR}) = 12\text{var}(\epsilon)(\int f^2(t)dt)^2$; Theorem 6.1 of Lehmann (1983) showed that $\inf_{F_S} e(\hat{\boldsymbol{\beta}}^{R^2}, \hat{\boldsymbol{\beta}}^{MR}) = 0.864$ where F_S denotes cumulative distribution functions with finite Fisher information. Note that the maximum likelihood estimator $\hat{\boldsymbol{\beta}}^{ML}$ is asymptotically $N(\mathbf{0}, I_f^{-1}\boldsymbol{\Sigma}^{-1})$, where $I_f = \int (f'(t))^2/f(t)dt$ is the Fisher information. Similarly, we can also calculate the asymptotic relative efficiency of $\hat{\boldsymbol{\beta}}^{R^2}$ to LAD or quantile regression. Generally speaking, $\hat{\boldsymbol{\beta}}^{R^2}$ is almost as efficient as mean regression or maximum likelihood estimators for normal errors but can be more robust for other errors; it is asymptotically much more efficient than LAD or quantile regression for many distributions of interest. For more introduction about its advantages of robustness and estimation efficiency, one can see Leng (2010). Actually, rank-based statistical procedures have played a fundamental role in nonparametric analysis of linear models due to its high efficiency and robustness, and we refer to the review paper of McKean (2004) for many useful references.

2.2 Distributed rank regression

When the sample size N is too large to store the whole data in one machine, we consider the distributed setting, where the samples $\{\boldsymbol{X}_i, Y_i\}_{i=1}^N$ are stored on K machines connected to a central processor, $\{Y_{ki}, \boldsymbol{X}_{ki} = (X_{ki1}, \ldots, X_{kip})^T\}_{i=1}^{n_k}$ denotes the data in the k-th machine, $k = 1, \ldots, K$, and $N = \sum_{k=1}^K n_k$. However, in the distributed setting, there are several issues. First, the above rank regression cannot be implemented directly, because it is infeasible to solve the related optimization problem (1) in one machine based on the full data. Even if the data are stored on one machine, the optimization problem (1) involves $N(N-1)$ pairwise differences, when N is large, it is

also challenging and time-consuming. Second, the existing distributed methods all assume that the datasets in each machine are independent and identically distributed, but data can be collected from various sources in practice, so the heterogeneity from machine to machine cannot be avoided. How to incorporate the heterogeneity also is an interesting issue.

Motivated by the above issue, we will propose a distributed rank regression, which allows the datasets in different machines are heterogeneous but follow the same regression relationship. Now, let us introduce the main idea and construction procedure of the new method. Based on the local dataset $\{Y_{ki}, \boldsymbol{X}_{ki} = (X_{ki1}, \ldots, X_{kip})^T\}_{i=1}^{n_k}$ stored in the k-th machine, we can obtain local R^2 estimator $\hat{\boldsymbol{\beta}}_k^{R^2}$ via

$$\hat{\boldsymbol{\beta}}_k^{R^2} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n_k(n_k-1)} \sum_{1\leqslant i,j\leqslant n_k} |Y_{kij} - \boldsymbol{X}_{kij}^T\boldsymbol{\beta}| \right\}, k = 1, \ldots, K,$$

where $Y_{kij} = Y_{ki} - Y_{kj}$, $\boldsymbol{X}_{kij} = \boldsymbol{X}_{ki} - \boldsymbol{X}_{kj}$. Note that $\hat{\boldsymbol{\beta}}_k^{R^2}$ enjoys the following asymptotic distribution

$$\sqrt{n_k}(\hat{\boldsymbol{\beta}}_k^R - \boldsymbol{\beta}_0) \to_d N\Big(\mathbf{0}, \frac{1}{12\omega^2}\boldsymbol{\Sigma}_k^{-1}\Big), k = 1, \ldots, K,$$

where $\boldsymbol{\Sigma}_k = E(\boldsymbol{X}_{ki}\boldsymbol{X}_{ki}^T)$ includes the covariates distribution information of the k-th machine. Obviously, $\hat{\boldsymbol{\beta}}_k^{R^2}, k = 1, \ldots, K$ are independent mutually, and thus, we can approximately treat $\{\hat{\boldsymbol{\beta}}_1^{R^2}, \ldots, \hat{\boldsymbol{\beta}}_K^{R^2}\}$ as an observed dataset coming from multivariate normal $\{N(\boldsymbol{\beta}_0, \frac{1}{12n_1\omega^2}\boldsymbol{\Sigma}_1^{-1}), \ldots, N(\boldsymbol{\beta}_0, \frac{1}{12n_K\omega^2}\boldsymbol{\Sigma}_K^{-1})\}$. This feature leads us to construct an approximate likelihood function

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K}(2\pi)^{-\frac{p}{2}}\left(\frac{1}{12n_k\omega^2}|\boldsymbol{\Sigma}_k^{-1}|\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta})^T(12n_k\omega^2\boldsymbol{\Sigma}_k)(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta})\right\},$$

which can incorporate $\boldsymbol{\Sigma}_k, k = 1, \ldots, K$ sufficiently. Then, the approximate log-likelihood function for $\boldsymbol{\beta}$ is

$$\log(L(\boldsymbol{\beta})) = C - \frac{1}{2}\sum_{k=1}^{K}(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta})^T(12n_k\omega^2\boldsymbol{\Sigma}_k)(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta}),$$

where C is free of $\boldsymbol{\beta}$. Thus, we can obtain the following least square-type loss function

$$\sum_{k=1}^{K} n_k(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta})^T\boldsymbol{\Sigma}_k(\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta}).$$

However, $\boldsymbol{\Sigma}_k, k = 1, \ldots, K$ are unknown, note that $\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{X}_{ki}\boldsymbol{X}_{ki}^T \to_p \boldsymbol{\Sigma}_k$; these naturally motivate us to consider the following object function

$$Q(\boldsymbol{\beta}) = \sum_{k=1}^{K} (\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta})^T \boldsymbol{X}_k^T \boldsymbol{X}_k (\hat{\boldsymbol{\beta}}_k^{R^2} - \boldsymbol{\beta}),$$

where $\boldsymbol{X}_k = (\boldsymbol{X}_{k1}, \ldots, \boldsymbol{X}_{kn_k})^T$. By minimizing $Q(\boldsymbol{\beta})$, the resulting estimator can be obtained, because it is built upon R^2 in the distributed data setting, we term it as distributed R^2 (abbreviated to DR^2) estimator, i.e.,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{DR^2} &= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q(\boldsymbol{\beta}) \\
&= \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{X}_k \right)^{-1} \left(\sum_{k=1}^{K} \boldsymbol{X}_k^T \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_k^{R^2} \right) \\
&= \left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k} \right)^{-1} \left(\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k} \hat{\boldsymbol{\beta}}_k^{R^2} \right),
\end{aligned}$$

where $w_k = \frac{n_k}{N}$.

Remark 2.1 Obviously, after obtaining the local estimators $\hat{\boldsymbol{\beta}}_k^{R^2}, k = 1, \ldots, K$ and $\frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}, k = 1, \ldots, K$ in different machines, we send them to the master machine; then, the resulting estimator $\hat{\boldsymbol{\beta}}^{DR^2}$ can be easily obtained in the master machine. Because only one round of communication is required, it is highly efficient in terms of communication.

Remark 2.2 A key feature of the new DR^2 is that it uses the covariates distribution information in the covariance matrixes $\boldsymbol{\Sigma}_k, k = 1, \ldots, K$ sufficiently by taking a weighted average of local estimators $\hat{\boldsymbol{\beta}}_k^{R^2}$ using weight matrixes $\frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}, k = 1, \ldots, K$, so our method is adaptive when the data are stored heterogeneously across different machines and achieve the advantage of estimation efficiency. While the CSL estimator (Jordan et al., 2019) only used the weight matrix of the master machine, the weight matrixes of other machines are completely ignored. These will lead to the loss of estimation efficiency when the datasets in different machines are heterogeneous.

Remark 2.3 We further interpret the efficiency advantage of the proposed procedure by comparing it with the "one-shot" method. For a clear illustration, we consider the simple univariate linear regression model $y = \beta x + \epsilon$ with datasets $\{y_{ki}, x_{ki}\}_{i=1}^{n}, k = 1, \ldots, K$, and the data have been centralized. Based on the local R^2 estimator $\hat{\beta}_k^{R^2}, k = 1, \ldots, K$, the "one-shot" estimator can be obtained as $\hat{\beta}^{OS} = \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_k^{R^2}$. Then, the variances of $\hat{\beta}^{OS}$ and $\hat{\beta}^{DR^2}$ can be calculated as

$$\mathrm{Var}(\hat{\beta}^{OS}) = \frac{1}{12\omega^2} \frac{1}{K^2} \left(\sum_{k=1}^{K} \frac{1}{\sum_{i=1}^{n} x_{ki}^2} \right) \text{ and } \mathrm{Var}(\hat{\beta}^{DR^2}) = \frac{1}{12\omega^2} \left(\frac{1}{\sum_{k=1}^{K} \sum_{i=1}^{n} x_{ki}^2} \right).$$

Note that $\frac{1}{K^2}\left(\sum_{k=1}^{K}\frac{1}{\sum_{i=1}^{n}x_{ki}^2}\right) \geqslant \left(\frac{1}{\sum_{k=1}^{K}\sum_{i=1}^{n}x_{ki}^2}\right)$, the equation holds if and only if $\sum_{i=1}^{n}x_{1i}^2 = \ldots = \sum_{i=1}^{n}x_{Ki}^2$. This also confirms the efficiency advantage of our new method, especially when the datasets in different machines are heterogeneous.

In order to investigate the asymptotic properties of the estimator $\hat{\boldsymbol{\beta}}^{DR^2}$, we assume the following regularity conditions.

(A1) For $k = 1, \ldots, K$, $\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{X}_{ki}\boldsymbol{X}_{ki}^T \to_p \boldsymbol{\Sigma}_k$, as $n_k \to \infty$.
(A2) Parameter space $\mathscr{H}$ is a compact subset of R^p, and the true parameter $\boldsymbol{\beta}_0$ is an inner point of $\mathscr{H}$.
(A3) For $k = 1, \ldots, K$, $\max_{i=1,\ldots,n_k}\|\boldsymbol{X}_{ki}\|/\sqrt{n_k} \to 0$.
(A4) The density of ϵ has finite Fisher information, that is $\int (f'(t))^2/f(t)dt < \infty$.

Conditions (A1)–(A4) are mild and routinely made in the rank regression modeling literature.

Theorem 1 *Denote $n = N/K$ as the average sample size for each machine, assume that $n/\sqrt{N} \to \infty$ and all n_k diverge in the same order $O(n)$, i.e., $c_1 \leqslant \min_k n_k/n \leqslant \max_k n_k/n \leqslant c_2$ for some positive constants c_1 and c_2. Under conditions (A1)–(A4), we have that*

$$\boldsymbol{\Xi}^{-\frac{1}{2}}\sqrt{N}(\hat{\boldsymbol{\beta}}^{DR^2} - \boldsymbol{\beta}_0) \to_d N\left(\mathbf{0}, \boldsymbol{I}_p\right),$$

where $\boldsymbol{\Xi} = \frac{1}{12\omega^2}\left(\sum_{k=1}^{K}w_k\boldsymbol{\Sigma}_k\right)^{-1}$, and $\boldsymbol{I}_p$ is a $p \times p$ identity matrix.

In this theorem, asymptotic covariance matrix $\frac{1}{12\omega^2}\left(\sum_{k=1}^{K}w_k\boldsymbol{\Sigma}_k\right)^{-1}$ includes the weighted average of $\boldsymbol{\Sigma}_k, k = 1, \ldots, K$, where weight w_k denotes the ratio of local sample size of the kth machine to global sample size. Obviously, if the data sets in different machines are homogeneous, i.e., $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}$, then $\frac{1}{12\omega^2}\left(\sum_{k=1}^{K}w_k\boldsymbol{\Sigma}_k\right)^{-1}$ becomes to $\frac{1}{12\omega^2}\boldsymbol{\Sigma}^{-1}$. What is more, if we assume that $n_1 = \ldots = n_K = n$, i.e., each machine stores a subsample of n observations, which also is assumed in Jordan et al. (2019); then, $\frac{1}{12\omega^2}\left(\sum_{k=1}^{K}w_k\boldsymbol{\Sigma}_k\right)^{-1}$ becomes to $\frac{1}{12\omega^2}\left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_k\right)^{-1}$.

From Theorem 1, we know that the estimator $\hat{\boldsymbol{\beta}}^{DR^2}$ obtained in the master machine has the same asymptotical distribution as the global rank regression estimator obtained based on the full data, or say, the estimator obtained in the mater machine works as well as all the data were pooled on a single machine. What is more, $n/\sqrt{N} \to \infty$ means that the order of sample size in each machine should be larger than $\sqrt{N}$, or say, the number of machines is smaller than the average local sample size n, which can be satisfied in practice.

2.3 Distributed regularized rank regression

In practice, the number of variables is often large, but only few of them may be really related to the response; thus, variable selection is critically important. However, how to conduct variable selection in the massive data setting has not been sufficiently investigated and the existing methods mainly focus on the mean regression or likelihood framework (Chen and Xie, 2014; Lee et al., 2015; Battey et al., 2018; Wang et al., 2017; Jordan et al., 2019).

Although $\hat{\boldsymbol{\beta}}^{DR^2}$ is robust, it does not enjoy the sparse property and cannot be used to do variable selection. Note that, based on the formula of $\hat{\boldsymbol{\beta}}^{DR^2}$, we can obtain that

$$\begin{aligned}
\frac{1}{N}Q(\boldsymbol{\beta}) &= \sum_{k=1}^{K}(\hat{\boldsymbol{\beta}}_k^{R^2}-\boldsymbol{\beta})^T w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}(\hat{\boldsymbol{\beta}}_k^{R^2}-\boldsymbol{\beta}) \\
&= \sum_{k=1}^{K}\boldsymbol{\beta}^T w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\hat{\boldsymbol{\beta}}_k^{R^2} + \sum_{k=1}^{K}(\hat{\boldsymbol{\beta}}_k^{R^2})^T w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\hat{\boldsymbol{\beta}}_k^{R^2} \\
&= \sum_{k=1}^{K}\boldsymbol{\beta}^T w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\hat{\boldsymbol{\beta}}^{DR^2} + \sum_{k=1}^{K}(\hat{\boldsymbol{\beta}}_k^{R^2})^T w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\hat{\boldsymbol{\beta}}_k^{R^2} \\
&= (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{DR^2})^T\left[\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\right](\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{DR^2}) + C,
\end{aligned}$$

where $C = \sum_{k=1}^{K}(\hat{\boldsymbol{\beta}}_k^{R^2})^T w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\hat{\boldsymbol{\beta}}_k^{R^2} - (\hat{\boldsymbol{\beta}}^{DR^2})^T\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\hat{\boldsymbol{\beta}}^{DR^2}$ is free with $\boldsymbol{\beta}$. Then, we propose the following adaptive LASSO-penalized loss function

$$P_\lambda(\boldsymbol{\beta}) = (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{DR^2})^T\left[\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}\right](\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{DR^2}) + \lambda\sum_{j=1}^{p}\lambda_j|\beta_j|,$$

where λ and $\lambda_j, j = 1,\ldots,p$ are positive tuning parameters. Because it is built upon DR^2 and penalty function, we refer to it as distributed regularized rank regression (abbreviated to DR^3). To simplify computation, we preselect λ_j as $\lambda_j = \frac{1}{|\hat{\beta}_j^{DR^2}|}, j = 1,\ldots,p,$; then, the resulting estimator can be obtained as

$$\hat{\boldsymbol{\beta}}_\lambda^{DR^3} = \left(\hat{\beta}_{\lambda,1}^{DR^3},\ldots,\hat{\beta}_{\lambda,p}^{DR^3}\right)^T = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} P_\lambda(\boldsymbol{\beta}).$$

Obviously, $\hat{\boldsymbol{\beta}}_\lambda^{DR^3}$ can also be obtained in the master machine, because $\hat{\boldsymbol{\beta}}^{DR^2}$ and $\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T\boldsymbol{X}_k}{n_k}$ can be calculated in the master machine.

Without loss of generality, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1 \in R^d$ and $\boldsymbol{\beta}_2 \in R^{p-d}$. We assume that the last $p-d$ components of the true parameter $\boldsymbol{\beta}_0$ are zeros, that is, $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^T, \mathbf{0}^T)^T$, where $\boldsymbol{\beta}_{01} = (\beta_{0,1},\ldots,\beta_{0,d})^T$ contains nonzero components. Furthermore, we define $a_N = \max\{\lambda_j : j = 1,\ldots,d\}$ and $b_N = \min\{\lambda_j : j = d+1,\ldots,p\}$. In fact, a_N controls the largest amount of penalty on the true nonzero parameters,

which should not be too large; otherwise, it will result in a highly biased estimator. In contrast, b_N is responsible for producing sparse solutions of irrelevant covariates, which should be sufficiently large. The asymptotic properties of $\hat{\boldsymbol{\beta}}_{\lambda}^{DR^3}$ are in the following theorem.

Theorem 2 *Assume that* $\sqrt{N}\lambda a_N \to 0$ *and* $\sqrt{N}\lambda b_N \to \infty$. *Then under the conditions in Theorem* 1, *the estimator* $\hat{\boldsymbol{\beta}}_{\lambda}^{DR^3}$ *satisfies*

(a) *Selection consistency*: $P(\hat{\mathcal{A}}_\lambda = \mathcal{A}) \to 1$, *where* $\mathcal{A} = \{1, \ldots, d\}$, $\hat{\mathcal{A}}_\lambda = \{j : \hat{\beta}_{\lambda,j}^{DR^3} \neq 0\}$;

(b) *Asymptotic normality*: $\boldsymbol{\Xi}_{\mathcal{A}}^{-\frac{1}{2}}\sqrt{N}(\hat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}}^{DR^3} - \boldsymbol{\beta}_{01}) \to N\big(\mathbf{0}, \boldsymbol{I}_d\big)$, *where* $\hat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}}^{DR^3} = (\hat{\beta}_{\lambda,1}^{DR^3}, \ldots, \hat{\beta}_{\lambda,d}^{DR^3})^T$, $\boldsymbol{\Xi}_{\mathcal{A}}$ *is the submatrix of* $\boldsymbol{\Xi}$ *whose entries correspond to the variables in* $\mathcal{A}$, *and* $\boldsymbol{I}_d$ *is a* $d \times d$ *identity matrix*.

With regarding the algorithm of minimizing $P_\lambda(\boldsymbol{\beta})$, this is a standard penalized least squares problem, whose solution path can be computed by the fast LARS algorithm at a computational complexity equal to a single least squares fit. Although DR^3 can select the true relevant variables consistently, its performance depends on the appropriate selection of tuning parameter. We propose the following distributed BIC-type criteria to select the tuning parameter λ, which is defined by

$$DBIC(\lambda) = (\hat{\boldsymbol{\beta}}_{\lambda}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2})^T \left[\sum_{k=1}^{K} w_k \frac{\boldsymbol{X}_k^T \boldsymbol{X}_k}{n_k}\right] (\hat{\boldsymbol{\beta}}_{\lambda}^{DR^3} - \hat{\boldsymbol{\beta}}^{DR^2}) + \frac{\log N}{N} df_\lambda, \tag{3}$$

where df_λ denotes the number of nonzero estimated parameters. In fact, the above DBIC is a trade-off between fitting accuracy and model size (i.e., the number of nonzero components in the resulting estimator, df_λ). Then, the optimal tuning parameter is the minimizer of $DBIC(\lambda)$, i.e., $\hat{\lambda} = \arg\min_{\lambda \in \mathbb{R}^+ = [0, +\infty)} DBIC(\lambda)$. Our simulation studies demonstrate that the distributed BIC-type criterion defined in (3) can select the tuning parameter satisfactorily and identify the true model consistently.

Finally, we investigate the theoretical property of the distributed BIC-type criterion. According to the value of $\hat{\mathcal{A}}_\lambda$, we partition $\mathbb{R}^+$ into three mutually exclusive regions, i.e., $\mathbb{R}_-^+ = \big\{\lambda \in \mathbb{R}^+ : \hat{\mathcal{A}}_\lambda \not\supset \mathcal{A}\big\}$, $\mathbb{R}_0^+ = \big\{\lambda \in \mathbb{R}^+ : \hat{\mathcal{A}}_\lambda = \mathcal{A}\big\}$, and $\mathbb{R}_+^+ = \big\{\lambda \in \mathbb{R}^+ : \hat{\mathcal{A}}_\lambda \supset \mathcal{A} \text{ and } \hat{\mathcal{A}}_\lambda \neq \mathcal{A}\big\}$. In other words, $\mathbb{R}_0^+$, $\mathbb{R}_-^+$ and $\mathbb{R}_+^+$ are three subsets of $\mathbb{R}^+$, in which the true-fitted, under-fitted and over-fitted models can be produced.

In order to analyze the efficiency of the distributed BIC-type criterion, we first give a reference tuning parameter, $\lambda_N = \log(N)/N$. By the selection of λ_j and note that $\hat{\boldsymbol{\beta}}^{DR^2}$ is $\sqrt{N}$ consistent, we know that a_N is of $O_p(1)$ and b_N is of $O_p(\sqrt{N}) \to \infty$. Thus, $\sqrt{N}\lambda_N a_N \to 0$ and $\sqrt{N}\lambda_N b_N \to \infty$ as $N \to \infty$, i.e., λ_N satisfies the conditions

about the tuning parameter in Theorem 2. Then, by Theorem 2 and the above discussion, we know $P(\hat{\mathcal{A}}_{\lambda_N} = \mathcal{A}) \to 1$. This implies that the relevant variables identified by λ_N converge to the true ones as the sample size increases.

Theorem 3 *Under the conditions in Theorem 2, we have that*

$$P\left(\inf_{\lambda \in \mathbb{R}_-^+ \bigcup \mathbb{R}_+^+} DBIC(\lambda) > \mathrm{DBIC}(\lambda_N)\right) \to 1.$$

From this theorem, we know that $\hat{\lambda} \notin \mathbb{R}_-^+ \bigcup \mathbb{R}_+^+$ holds asymptotically, because $\lambda_N = \log(N)/N$ that can lead to consistent variable selection is a better choice. This results that the $\hat{\lambda}$ must be one of those in $\mathbb{R}_0^+$ asymptotically. Thus. optimal tuning parameter selected by this distributed BIC-type criterion can identify the true submodel consistently.

3 Simulation studies and applications to real data

3.1 Simulation studies

In this subsection, we illustrate the finite sample performance of our new method by simulation studies. Experiment 1 shows the estimation accuracy of the DR^2 estimator; Experiment 2 demonstrates the variable selection results of the DR^3 method.

Experiment 1. In this experiment, we generate datasets from the linear regression model $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \epsilon_i$, where $\boldsymbol{\beta} = (1, -1, 2, 1)^T$. To illustrate the robust property of our method, for the random error ϵ_i, we considered three distributions, (1) the standard normal, (2) the t distribution with 4 degrees of freedom (denoted by t_4), and (4) contaminated normal (abbreviated to CN) with distribution function $(1-\rho)\Phi(x) + \rho\Phi(\frac{x}{3})$, where $\Phi(x)$ is the distribution function of standard normal, $\rho \in [0, 1]$ is the contamination proportion and we choose $\rho = 0.1$. For the covariates, two different data storage strategies are considered. Specifically,

Case 1 (i.i.d covariates). The covariates in different machines are distributed independently and identically, i.e., the covariates $X_{ij}, i = 1, \ldots, N, j = 1, \ldots, p$ in every machine are all sampled from the standard normal distribution $N(0, 1)$;
Case 2 (heterogeneous covariates). In this case, the covariates may follow different distribution across each machine. Specifically, in the kth machine, $\boldsymbol{X}_{ki}, i = 1, \ldots, n_k$ can be sampled from zero mean multivariate normal or multivariate t_2 distributions with probability 0.6 and 0.4, respectively, and their covariance matrix is a Toeplitz matrix, i.e., $\boldsymbol{\Sigma}_k = (\sigma_{k,j_1 j_2}) = (\rho_k^{|j_1 - j_2|})_{j_1, j_2 = 1}^p$, and ρ_k is sampled from $U[0.3, 0.4]$.

We set the sample size as $N = M \times 10^4$ and consider $M = (1, 2, 10)$, respectively, and the number of machines is set to $K = \{(50, 100), (50, 100), (250, 400, 500)\}$ correspondingly. For a clear comparison, the experiment is repeated 100 times. We compare the proposed DR^2 method with several methods, i.e., (a) the OS estimator $\hat{\boldsymbol{\beta}}^{OS} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\beta}}_k^{R^2}$ (Zhang et al., 2013), (b) the CSL estimator $\hat{\boldsymbol{\beta}}^{CSL}$ (Jordan et al., 2019), c) the global least squares estimator $\hat{\boldsymbol{\beta}}^{GLS} = (\sum_{i=1}^{N}\boldsymbol{X}_i\boldsymbol{X}_i^T)^{-1}(\sum_{i=1}^{N}\boldsymbol{X}_iY_i)$, and (d) the global LAD estimator $\hat{\boldsymbol{\beta}}^{GLAD} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\sum_{i=1}^{N}|Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}|\}$. To measure the estimation efficiency, we calculate the mean square error (MSE) of different estimators based on the 100 replications, i.e., $MSE(\hat{\beta}_j) = \frac{1}{100}\sum_{t=1}^{100}(\hat{\beta}_j - \beta_{0j})^2$, and then, we give the relative estimation efficiency (REE) of our new DR^2 estimator $\hat{\boldsymbol{\beta}}^{DR^2}$ with respect to $\hat{\boldsymbol{\beta}}^{OS}$, $\hat{\boldsymbol{\beta}}^{CSL}$, $\hat{\boldsymbol{\beta}}^{GLS}$ and $\hat{\boldsymbol{\beta}}^{GLAD}$, respectively, e.g., $REE_j(OS) = \frac{MSE(\hat{\beta}_j^{OS})}{MSE(\hat{\beta}_j^{DR^2})}$, $REE_j(CSL)$, $REE_j(GLAD)$ and $REE_j(GLS)$ can be defined similarly, $j = 1, \ldots, p$.

The simulation results are reported in Figs. 1–6, which are given in the supplementary file. From these simulation results, we have the following findings. First, in Case 1, i.e., the datasets are independent and identically distributed, we find that, when the random error follows standard normal distribution, all values of $REE_j(CSL)$ and $REE_j(GLS)$ are approximately equal to 1, which means our DR^2 estimator is asymptotically as efficient as the GLS and CSL estimators, while when errors follow t_4 or contaminated normal distributions, our DR^2 estimator is more efficient than CSL, GLS estimators, because all the values of $REE_j(CSL)$ and $REE_j(GLS)$ are larger than 1, what is more, the $REE_j(GLAD)$ is always larger than 1 for every error distribution in Case 1, which implies that $\hat{\boldsymbol{\beta}}^{DR^2}$ is more efficient than the GLAD estimator. Second, when the datasets are heterogeneous (i.e., Case 2), we can see that our new method performs much better than the two competing distributed estimators, OS and CSL, because the values of $REE_j(OS)$ and $REE_j(CSL)$ are always bigger than 1 for every error distribution and (K, M), where the CSL method behaves worst in this case, because it only used the covariates distribution information of the master machine; the information in other machines is ignored. The comparison results between our DR^2 method and two competing global estimators (GLS and GLAD) in Case 2 are similar to that in Case 1. Also, Table 1 reports the empirical standard

Table 1 The empirical standard deviations (×100) of DR^2 estimators for $N = 1 \times 10^4$

Error	K	Case 1				Case 2			
		β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
$N(0, 1)$	50	0.356	0.375	0.329	0.401	0.375	0.358	0.417	0.392
	100	0.412	0.311	0.308	0.337	0.391	0.423	0.361	0.327
CN	50	0.456	0.439	0.511	0.449	0.456	0.491	0.465	0.462
	100	0.461	0.466	0.471	0.503	0.437	0.477	0.422	0.433
t_4	50	0.438	0.487	0.479	0.459	0.433	0.511	0.497	0.512
	100	0.475	0.505	0.462	0.487	0.419	0.439	0.445	0.476

deviations of DR^2 estimators for $N = 1 \times 10^4$, which further confirm the excellent performance of the proposed method.

Experiment 2. In this experiment, data sets are generated from model $Y_i = \sum_{j=1}^{40} X_{ij}\beta_j + \epsilon_i$, where $\boldsymbol{\beta} = (1, -1, 2, 1, 0, \ldots, 0)^T$, i.e., the first four variables are really relevant and the remaining 36 variables are irrelevant. For the random error ϵ_i, to illustrate the robustness, we also consider the three error distributions that are used in Experiment 1. For the covariates, similar to Experiment 1, we also consider two cases.

Case 3 (i.i.d covariates). In every machine, covariates $X_{ij}, i = 1, \ldots, N, j = 1, \ldots, 4$ are all sampled from the standard normal $N(0, 1)$, while $X_{ij}, i = 1, \ldots, N, j = 5, \ldots, 40$ follow t distribution with 2 degrees of freedom, where $X_{ij}, j = 1, \ldots, 40$ are independent mutually.

Case 4 (heterogeneous covariates). In the kth machine, the covariates $X_{kij}, i = 1, \ldots, n_k, j = 1, \ldots, p$ can be sampled from two strategies. In the first strategy, $X_{kij}, j = 1, \ldots, 4$ follow $N(0, 1)$ and $X_{kij}, j = 5, \ldots, 40$ follow t distribution with 2 degrees of freedom. In the second strategy, $X_{kij}, j = 1, \ldots, 4$ follow t distribution with 2 degrees of freedom and $X_{kij}, j = 5, \ldots, 40$ follow $N(0, 1)$. We select the first strategy and the second strategy randomly with probability 0.6 and 0.4, respectively.

Table 2 Simulation results in experiment 2 for $N(0, 1)$ error

	(K,M)	CF (%)	REE (Oracle)			
			β_1	β_2	β_3	β_4
Case 3	(50,1)	95	0.9673	0.9486	0.9588	1.0113
	(100,1)	96	0.9778	1.0017	1.0103	0.9904
	(50,2)	93	1.0158	0.9593	0.9160	1.0211
	(100,2)	96	0.9894	0.9296	1.0030	0.9865
	(250,10)	93	0.9833	0.9972	1.0278	1.0321
	(400,10)	94	1.0501	1.0309	0.9924	0.9627
	(500,10)	94	0.9833	1.0985	0.9746	0.9812
Case 4	(50,1)	95	0.9912	1.0003	0.9899	0.9883
	(100,1)	93	1.0003	0.9989	0.9719	0.9993
	(50,2)	94	1.0109	1.0008	0.9867	0.9177
	(100,2)	92	0.9453	0.9677	0.9818	0.9376
	(250,10)	95	0.9919	0.9898	0.9162	0.9172
	(400,10)	94	1.0023	0.9387	0.9558	0.9387
	(500,10)	96	0.9177	0.9818	0.9735	0.9198

Table 3 Simulation results in experiment 2 for contaminated normal error

	(K,M)	CF (%)	REE (Oracle)			
			β_1	β_2	β_3	β_4
Case 3	(50,1)	91	0.9207	0.9562	0.9619	0.9856
	(100,1)	94	1.0002	0.9198	0.9618	0.9337
	(50,2)	95	0.9781	1.0217	0.9011	0.9223
	(100,2)	93	0.9513	0.9177	0.9201	1.0006
	(250,10)	95	0.9198	0.9898	0.9162	0.9411
	(400,10)	92	0.9012	0.9535	0.9326	0.9172
	(500,10)	93	0.9977	0.9818	0.9755	0.9003
Case 4	(50,1)	94	0.9478	1.0041	1.0172	1.0177
	(100,1)	95	0.9717	0.9198	0.9431	0.9158
	(50,2)	94	0.9009	1.0021	0.9732	0.9577
	(100,2)	94	0.9327	0.9312	0.9656	0.9346
	(250,10)	96	0.9198	0.9876	0.9165	0.9403
	(400,10)	95	1.0012	1.0005	0.9626	0.9412
	(500,10)	96	0.9977	0.9818	0.9755	0.9179

Table 4 Simulation results in experiment 2 for t_4 error

	(K,M)	CF (%)	REE (Oracle)			
			β_1	β_2	β_3	β_4
Case 3	(50,1)	93	0.9078	0.9405	0.9172	0.9331
	(100,1)	94	1.0012	0.9128	0.9325	0.9119
	(50,2)	96	0.9192	0.9371	0.9077	0.9173
	(100,2)	95	0.8997	0.9005	0.9889	0.9377
	(250,10)	92	0.9198	0.9877	0.9163	0.9612
	(400,10)	96	1.0009	0.9535	0.9321	0.9375
	(500,10)	92	0.9977	0.9818	0.9755	0.9473
Case 4	(50,1)	93	0.9118	0.9077	0.9105	0.9558
	(100,1)	94	0.9367	0.9709	0.9558	0.9319
	(50,2)	92	0.9789	0.9903	0.9659	0.9883
	(100,2)	96	1.0036	1.0135	0.9614	0.9178
	(250,10)	95	0.9188	0.9537	0.9166	0.9272
	(400,10)	94	0.9712	0.9535	0.9323	1.0075
	(500,10)	93	0.9978	0.9188	0.9659	0.9412

To investigate the estimation accuracy of our new DR^3 variable selection procedure, we also calculate the oracle DR^2 estimators, which are obtained by the DR^2 method under the true relevant variables; then, the relative estimation efficiency of the DR^3 estimators with respect to the oracle DR^2 estimators is reported, i.e., REE(Oracle), which is defined as the ratio of MSE of oracle DR^2 estimator to MSE of DR^3 estimator. Furthermore, we also report the percentage of correct

model fitted (CF). Tables 2, 3 and 4 list the simulation results based on 100 replications. First, we can see that, for all the error distributions under consideration, $REE(Oracle) \approx 1$ for every case and (K, M), this means that the DR^3 estimator can estimate the unknown parameters as accurately as the oracle estimator. Second, by the values of CF, we can see that the proposed DR^3 method with the distributed *BIC*-type criterion can select the true relevant variables consistently. All of these conforms to the asymptotic results.

3.2 Applications to real data

In this subsection, we apply the proposed method to the greenhouse gas (GHG) observing network data, which are reported by the UCI machine learning repository. This dataset consists of 954, 840 observations, and the global rank regression cannot work at all. Thus, it can be used to demonstrate the proposed method for massive data. The response variable is GHG concentrations of synthetic observations, and there are 15 predictors. These predictors are GHG concentrations of tracers emitted from 14 distinct spatial regions in California and one outside of California (denoted as Reg1–Reg15). Then, we will predict the GHG concentrations of synthetic and determine the important regions, which play an important role in the GHG concentrations of synthetic.

Table 5 The values of MAPE and coefficients estimation results for DR^2 and DR^3 methods in analysis of the greenhouse gas observing network data set

	$K = 400$		$K = 500$	
	DR^2	DR^3	DR^2	DR^3
Reg1	1.3178	1.5369	1.6382	1.4119
Reg2	0.6722	0.5391	0.5792	0.7033
Reg3	0.8637	1.0117	1.1217	1.0973
Reg4	0.0085	0.0000	0.0026	0.0000
Reg5	0.0534	0.0000	0.0178	0.0000
Reg6	3.9951	3.7782	4.0766	4.1369
Reg7	0.0018	0.0000	0.0011	0.0000
Reg8	0.6452	0.5339	0.7118	0.7091
Reg9	0.5851	0.7732	0.6231	0.5977
Reg10	0.3846	0.2879	0.3742	0.4003
Reg11	0.0846	0.0000	0.0512	0.0000
Reg12	0.6893	0.5017	0.6671	0.5988
Reg13	0.4043	0.3519	0.5138	0.5321
Reg14	1.0765	1.1532	0.9278	1.1763
Reg15	0.6743	0.7739	0.5798	0.5392
MAPE	10.5019	10.3076	10.1749	10.1717

The whole dataset is divided into training dataset and testing dataset; specifically, we use the first 500, 000 observations as training dataset D_{train}, and the rest as the testing dataset D_{test}. Then, we compute the coefficients using the training dataset D_{train} and calculate the mean of the absolute prediction error (MAPE) based on the testing dataset D_{test}, where MAPE is the mean of $\{|\hat{Y}_i - Y_i|, i \in D_{test}\}$. The number of machines K is taken as $\{100, 200, 400, 500\}$, respectively. For different values of K, we first apply the DR^2 method; then, the DR^3 is implemented for variable selection.

From the computation results, we find that the proposed methods are insensitive to the choice of K. Table 5 summarizes the values of MAPE and coefficients estimation results for $K = 400$ and $K = 500$, respectively, which also confirm this finding. We can see that Reg4, Reg5, Reg7, and Reg11 are not selected in the model, which means the monitoring for these regions can be reduced to save the cost of human and material resources. Besides, stations 1, 3, 6 and 14 have the biggest coefficients, so the monitoring of these four regions should be further strengthened.

4 Concluding remarks

This paper proposes a distributed rank regression for the massive data, which can be implemented in the master machine conveniently. Theoretically, the resulting estimator is statistically as efficient as the global rank regression estimator. What is more, combining with the adaptive LASSO, a distributed regularized rank regression variable selection method is constructed, which can make consistent variable selection and can be easily implemented by using the LARS algorithm.

In the existing literature, variable selection is often discussed in the dimensional settings, i.e., the dimension of covariates is larger than the sample size. It is interesting to see how the proposed method can be adapted, when dimension p is much larger than n_k, for $k = 1, \ldots, K$. Furthermore, the asymptotic covariance matrix Ξ involves unknown parameter ω^2, and how to estimate it in the distributed massive data setting is unknown, we will study it in the future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-021-00803-5.

References

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

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

Chen, L., Zhou, Y. (2019). Quantile regression in big data: A divide and conquer based strategy. *Computational Statistics and Data Analysis*. https://doi.org/10.1016/j.csda.2019.106892.

Fan, J., Wang, D., Wang, K., Zhu, Z. (2017). Distributed estimation of principal eigenspaces. arXiv preprint arXiv:1702.06488 .

Fan, J., Guo, Y., Wang, K. (2019). Communication-efficient accurate statistical estimation. arXiv preprint arXiv:1906.04870

Feng, L., Zou, C., Wang, Z., Wei, X., Chen, B. (2015). Robust spline-based variable selection in varying coefficient model. *Metrika, 78,* 85–118.
Jordan, M. I., Lee, J. D., Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association, 14,* 668–681.
Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica, 46,* 33–50.
Lee, J., Sun, Y., Liu, Q., Taylor, J. (2015). Communication-efficient sparse regression: a one-shot approach. arXiv preprint arXiv: 1503.04337.
Lehmann, E. (1983). *Theory of Point Estimation*. New York: Wiley.
Leng, C. (2010). Variable selection and coefficient estimation via regularized rank regression. *Statistica Sinica, 20,* 167–181.
Lin, N., Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and Its Interface, 4,* 73–83.
McKean, J. (2004). Robust analysis of linear models. *Statistical Science, 19,* 562–570.
Rosenblatt, J., Nadler, B. (2016). On the optimality of averaging in distributed statistical learning. *Information and Inference*: *A Journal of the IMA*, *5,* 379–404.
Shin, Y. (2010). Local rank estimation of transformation models with functional coefficients. *Econometric Theory, 26,* 1807–1819.
Wang, H., Li, G., Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business and Economic Statistics, 25,* 347–355.
Wang, J., Kolar, M., Srebro, N., Zhang, T. (2017). Efficient distributed learning with sparsity. In: *International Conference on Machine Learning*, 3636-3645.
Wang, L., Li, R. (2009). Wighted Wilcoxon-type smoothly clipped absolute deviation method. *Biometrics, 65,* 564–571.
Wang, L., Kai, B., Li, R. (2009). Local rank inference for varying coefficient models. *Journal of the American Statistical Association, 488,* 1631–1645.
Zhang, Q., Wang, W. (2007). A fast algorithm for approximate quantiles in high speed data streams. In *Proceedings of the International Conference on Scientific and Statistical Database Management*.
Zhang, Y., Duchi, J., Wainwright, M. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Reaearch, 14,* 3321–3363.
Zhang, Y., Duchi, J., Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research, 16,* 3299–3340.
Zhu, X., Li, F., Wang, H. (2019). Least squares approximation for a distributed system. arXiv preprint arXiv: 1908.04904.
Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association, 101,* 1418–1429.