Distributed Mallows C_L Averaging for Ridge Regressions

Haili Zhang [1], Alan T.K. Wan [2] *, Guohua Zou [3], and Kang You [4]

[1] *Institute of Applied Mathematics, Shenzhen Polytechnic, Shenzhen, 518055, China. E-mail: zhanghl@szpt.edu.cn*
[2] *Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong. E-mail: Alan.Wan@cityu.edu.hk, * Corresponding author.*
[3] *School of Mathematical Sciences, Capital Normal University, Beijing, 100048, China. E-mail: ghzou@amss.ac.cn,*
[4] *School of Mathematical Sciences, Capital Normal University, Beijing, 100048, China. E-mail:*

Abstract

Ridge regression is an effective tool for multicollinearity in regressions. It is also an essential type of shrinkage and regularization method for high dimensional data, and is widely used in big data and distributed data applications. The divide and conquer trick in distributed data, combining the estimator in each subset with equal weight, is commonly applied. In order to overcome multicollinearity and improve the estimation accuracy in the presence of distributed data, we propose a Mallows-type model averaging method for ridge type regressions, which combines the estimators from all subsets. The asymptotic optimality and consistency of the proposed method are derived. A simulation study and a real data analysis illustrate that the proposed model averaging methods often perform better than the commonly used model selection and model averaging methods in distributed data cases.

Keywords: Asymptotic optimality, consistency, divide and conquer, Mallows model averaging, ridge regression.

1. Introduction

In recent years, rapid advances in big data have had a profound impact on all aspects of human life. The availability of granular data sets allows hidden patterns, unknown correlations and trends, and other insights from the data to be uncovered. Researchers can now ask questions that were impossible two or three decades ago when traditional data repository could hold only limited amounts of aggregated data. The analysis of big data poses significant challenges to statistical analysis. Important statistical advances including the refinement of existing tools and the development of new tools have been made to face the new phenomena emerged under the big data regime. This paper considers ridge regression (RR) (Hoerl and Kennard, 1970) and devise a fast and stable algorithm for its execution in a distributed computing environment for big data. Ridge regression is a well-known statistical regularisation technique, being particularly useful for overcoming the multicollinearity problem in linear regression. RR alters the usual least squares objective function by adding a penalty equivalent to the square of the magnitude of the coefficients. This penalty factor penalises large values of the coefficient estimates, introducing bias but shrinking the estimates towards zero, thereby reducing the mean square error (MSE). Ridge regression has the objective of trading a small amount of bias for a large reduction in the variance of the estimator. The RR technique has found applications in many diverse fields including economic forecasting (Mirakyan et al., 2017), face recognition (Xue et al., 2009), genetic analysis (Shen et al., 2013; Zhan and Xu, 2012), transfer learning and neural networks (Deng et al., 2014), machine learning (Dobriban and Sheng, 2020) and others. The popular dropout method in deep learning may also be viewed as a ridge-type regression method (Srivastava et al., 2014). One important aspect of ridge regression is the choice of tuning parameter that controls the strength of the penalty term. Common methods include ridge mapping (Hoerl and Kennard, 1970), Bayesian method (Dempster et al., 1977), and cross-validation (Golub et al., 1979; Lee, 1987). See Kibria (2003, 2012) for a comparison of these methods.

When confronted with massive data sets, the limited memory storage of any single machine makes it infeasible to do everything on one machine. Distributed computing, which divides a computing task into smaller tasks and assigns these tasks to multiple computers, is becoming a popular approach for tackling big data. One common strategy within the distributed computing paradigm is the divide-and-conquer strategy, whereby the data are split into independent subsets, each being analysed independently and the results from the partitioned subsets are aggregated to give the final solution (Chen and Xie, 2014; Wang et al., 2016, 2018). While the original divide-and-conquer method is based on simple averaging, the method has been continually refined. For example, Xi et al. (2009) developed a compression and aggregation scheme for logistic regression in data cubes that shares the spirit of the divide-and-conquer strategy but uses a weighted average aggregation formula. Lin and Xi (2011) generalised the method to estimating equations. Zhang et al. (2013) developed a resampling method that reduces the bias and improves the MSE performance of the divide-and-conquer scheme, and Fan et al. (2021) developed a divide-and-conquer strategy based on a multi-round algorithm for distributed inference. Numerous other statistical techniques covering the divide-and-conquer strategy have been developed within the framework of sparse regression (Lee et al., 2017; Tang et al., 2020), quantile regression (Chen and Zhou, 2019; Volgushev et al., 2019), M-estimation (Wang et al., 2017; Shi et al., 2018; Jordan et al., 2019), partially linear model (Zhao et al., 2016), principal component analysis (Garber et al., 2017; Fan et al., 2019), support vector machine (Lian and Fan, 2017; Wang et al., 2019), and others.

Ridge regression in the context of distributed computing has received some attention in the recent body of literature. Zhang and Yang (2017) developed a RR method based on sufficient statistics obtained through scanning. These sufficient statistics then form the basis of subsequent computation without having to access the raw data. Zhang et al. (2015) developed a divide-and-conquer algorithm which, when implemented with kernel RR, yields an estimator that achieves the minimax asymptotic convergence rate. Chang et al. (2017) studied a distributed kernel RR that uses the ratio of the size of data subset to the total data size as the weight in the conquer step of the divide-and-conquer strategy.

In this paper, we pursue a divide-and-conquer strategy for RR that utilises the popular Mallow's criterion in model averaging to derive the weights that determines the contribution of the analysis from the individual machines to the weighted ensemble. In regression, model averaging provides an alternative to model selection; instead of choosing and drawing inference from the best-fitting model as in model selection, model averaging combines the models and there is ample evidence that model averaging yields lower prediction errors than the individual contributing models. Ullah et al. (2017) demonstrated an algebraic relationship between the model averaging and the generalised RR estimator in regression. A major part of the model averaging literature is concerned with ways of weighting models. Among the methods that have been proposed, Hansen's (2007) Mallow's averaging (MMA) is arguably the most popular as it does not only possess optimal asymptotic and finite sample properties but is easy to implement. Other model averaging strategies include those based on information scores (Hjort and Claeskens, 2003), adaptive regression by mixing Yang (2001), cross-validation (Hansen and Racine, 2012), and others. Some results on model averaging where the candidate set contains ridge regression estimators with different ridge parameters have been obtained by Schomaker (2012) and Zhao et al. (2020).

The primary object of this paper is to develop a Mallow-type divide-and-conquer algorithm for ridge regression within a distributed computing environment in the face of massive data. Our method aggregates the RRs obtained from individual data subsets by a weighted average with weights determined by the Mallow's criterion. Our proposed method is easy to implement and can be readily applied to homogeneous and heterogeneous data. **The following is unclear , I thought about it and decided to delete it, maybe we can find other example for here or not.** We prove the asymptotic optimality and consistency of the proposed algorithm.

The paper is organized as follows. Section 2 introduces ridge regression model estimation averaging method. The optimality and consistency of the Mallows divide-and-conquer algorithm is studied in Section 3. Simulation studies for different model selection and averaging methods are presented in Section 4. Section 5 is the actual data analysis. Section 6 concludes. Proofs of theoretical results are provided in the Appendix.

2. Ridge Regression Estimation and Averaging

2.1. Model framework and estimation

Consider the following linear model

$$Y = X\beta + \varepsilon, \tag{1}$$

where $Y = (y_1, y_2, \ldots, y_N)^T$ is an $N \times 1$ dimensional response variable, $X = (x_1, x_2, \ldots, x_N)^T$ is an $N \times p$ dimensional covariate, and the error term $\varepsilon = (\varepsilon_1, \varepsilon_1, \ldots, \varepsilon_N)^T$ satisfies the assumptions of $\mathbf{E}(\varepsilon|X) = 0$ and $\mathbf{Var}(\varepsilon|X) = \Omega$ with Ω being an N-dimensional diagonal matrix.

We assume that the data of size N are partitioned into K subsets. Here, we relax the conventional but restrictive assumption in the literature that all partitioned subsets have identical sample sizes and assume that the size of the kth subset is N_k such that $N = \sum_{k=1}^{K} N_k$. Denote the data of the kth subset as $\{Y_k, X_k\}$, where $Y_k = (y_{k,1}, y_{k,2}, \ldots, y_{k,N_k})^T$ and $X_k = (x_{k,1}, x_{k,2}, \ldots, x_{k,N_k})^T$. For simplicity, we write $Y = \left(Y_1^T, Y_2^T, \ldots, Y_K^T\right)^T$, $X = \left(X_1^T, X_2^T, \ldots, X_K^T\right)^T$ and $\Omega = diag(\Omega_1, \Omega_2, \ldots, \Omega_K)$. The RR estimator obtained based on the kth subset of data is given by

$$\hat{\beta}_k(\lambda_k) = \left(X_k^T X_k + \lambda_k n_k I_p\right)^{-1} X_k^T Y_k, \tag{2}$$

where λ_k is a tuning parameter that controls the amount of shrinkage. A range of data driven methods are available for choosing λ_k including Bayesian method, Mallows'C_p criterion and cross-validation methods (Craven and Wahba, 1978; Xiang and Wahba, 1996). Here, we assume that λ_k is chosen by the Mallow's C_p criterion. The following model averaging Ridge Regression (MARR) estimator is a weighted linear combination of the RR estimators obtained from the different data subsets under a distributed computing environment:

$$\hat{\beta}(W) = \sum_{k=1}^{K} w_k \hat{\beta}_k(\lambda_k), \tag{3}$$

where $W = (w_1, w_2, \ldots, w_K)^T$ is a vector from the simplex

$$H = \left\{ W \in [0,1]^K : \sum_{k=1}^{K} w_k = 1 \right\}$$

in the real space $\mathbb{R}^K$.

Our proposed weight choice criterion borrows the idea of the Mallows C_p criterion (Mallows, 1973), to be described below.

2.2. *Global Mallows $C_L$criterion for ridge regressions*

Our weight choice criterion, to be referred to as the distributed global Mallows criterion, is given as follows:

$$HC_g(W) = \left\|Y - X\hat{\beta}(W)\right\|^2 + 2\sum_{k=1}^{K} w_k F_k, \tag{4}$$

where $F_k = tr(P_k \Omega_k)$, $P_k = X_k \left(X_k^T X_k + n_k \lambda_k I_p\right)^{-1} X_k^T$, and $tr(\cdot)$ denotes the trace of a matrix. If each of Ω_k's is known, the optimal weight vector is obtained by minimising $HC_g(W)$, yielding

$$\hat{W}^* = (\hat{w}_1^*, \hat{w}_2^*, \ldots, \hat{w}_K^*)^T = \arg\min_{w \in H} HC_g(W).$$

When Ω_k is unknown and replaced by an estimator, the global Mallows criterion also changes accordingly.

We consider the following error structures for estimating Ω.

2.2.1. *Heteroscedastic distributed subjects*

Let the covariance matrix Ω be

$$\Omega = \begin{pmatrix} \sigma_1^2 I_{N_1} & 0 & \cdots & 0 \\ 0 & \sigma_2^2 I_{N_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_K^2 I_{N_K} \end{pmatrix}, \tag{5}$$

where the errors within a given data subset are homoscedastic but they are heteroscedastic across the different data subsets. For this scenario, $HC_g(W)$ becomes

$$hC_g(W) = \left\|Y - X\hat{\beta}(W)\right\|^2 + 2\sum_{k=1}^{K} w_k \sigma_k^2 tr(P_k).$$

One may estimate σ_k^2 by the maximum Likelihood estimator

$$\hat{\sigma}_k^2 = N_k^{-1} \sum_{i=1}^{N_k} \hat{\sigma}_{k,i}^2. \tag{6}$$

Denote the criterion upon substituting $\hat{\sigma}_k^2$ for σ_k^2 in hC_g as $\widehat{hC}(W)$. We obtain the optimal weight by minimising $\widehat{hC}(W)$ on the simplex H.

2.2.2. Homoscedastic distributed subjects

Let the errors across the data subsets be homoscedastic, i.e.,

$$\Omega = \sigma^2 I_N. \tag{7}$$

The global Mallows criterion is modified to be

$$C_g(W) = \left\| Y - X\hat{\beta}(W) \right\|^2 + 2\sigma^2 \sum_{k=1}^{K} w_k tr(P_k).$$

As σ^2 is unknown, we estimate it by

$$\hat{\sigma}^2 = \left(\sum_{k=1}^{K} N_k \right)^{-1} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \hat{\sigma}_{k,i}^2. \tag{8}$$

We denote the resultant criterion after substituting $\hat{\sigma}^2$ for σ^2 in C_g as $\widehat{C}_g(W)$. The optimal weight is obtained by minimising $\widehat{C}_g(W)$ on the simplex H.

2.2.3. Algorithms for ridge regressions

Table 1: Distributed Algorithm

```
Algorithm 1: Distributed Global Mallows $C_L$ Averaging for Ridge Regressions (DGR)

Input: Distributed datasets as $\{Y_k, X_k\}$.
Output: Optimal weight vector $\hat{W}$ or $\hat{W}'$.
Initialization: In each subset, obtain $\hat{\beta}_k(\lambda_k)$ by (2).
Step 1: Obtain the predictions of $\mu_l, l = 1, 2, \dots, K$.
  1.1: Transmit $\hat{\beta}_k(\lambda_k)$ to the $l$th subset, $l \neq k$.
  1.2: Obtain the predictions of $Y_l$ based on different $\hat{\beta}_k(\lambda_k)$ as $\hat{\mu}_{l,1}, \hat{\mu}_{l,2}, \dots \hat{\mu}_{l,K}$.
  1.3: Calculate the errors of these predicted value, i.e., $\hat{e}_{l,1}, \hat{e}_{l,2}, \dots \hat{e}_{l,K}$ with $\hat{e}_{l,k} = (\hat{e}_{l,k,(1)}, \hat{e}_{l,k,(2)}, \dots, \hat{e}_{l,k,(N_l)})^T$.
       Denote $\hat{e}_l = (\hat{e}_{l,1}, \hat{e}_{l,2}, \dots \hat{e}_{l,K})$.
Step 2: Estimate $\Omega_k, k = 1, 2, \dots, K$.
  2.1: Denote $\Omega_k = diag(\sigma^2_{k,1}, \sigma^2_{k,2}, \dots, \sigma^2_{k,N_k})$. Let $\hat{\sigma}^2_{k,i} = \hat{e}^2_{k,k,(i)}$.
  2.2: Compute $tr(P_k)$. $F_k = \sum_{i=1}^{N_k} P_{k,ii}$,
  2.3: Compute $\hat{\sigma}^2_k, k = 1, 2, \dots, K$ by (6), or $\hat{\sigma}^2$ by (8).
Step 3: Select weight vector $\hat{W}$ or $\hat{W}'$.
  3.1: Compute the inner product matrix $\widehat{E_l} = \hat{e}_l^T \hat{e}_l$ with $K \times K$ dimensions, sum these matrices as $\widehat{E} = \sum_{k=1}^K \widehat{E_k}$
  3.2: The final optimization problem is solved by
         $\hat{W} = \arg\min_W W^T \widehat{E} W + 2W^T F'$,
         with constraints $\sum_{k=1}^K w_k = 1, 0 \leq w_k \leq 1, k = 1, 2, 3, \cdots, K$,
       where $F' = \left(F_1\hat{\sigma}^2_1, F_2\hat{\sigma}^2_2, \dots, F_K\hat{\sigma}^2_K\right)^T$, or
         $\hat{W}' = \arg\min_W W^T \widehat{E} W + 2\hat{\sigma}^2 W^T F$,
         with constraints $\sum_{k=1}^K w_k = 1, 0 \leq w_k \leq 1, k = 1, 2, 3, \cdots, K$,
       where $F = (F_1, F_2, \dots, F_K)^T$.
End
```

3. Theoretical Results

In this section, we will show the optimality and the consistency of our proposed model averaging estimator.

3.1. Global optimality for distributed data

Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ be the maximum and minimum eigenvalue of a matrix respectively. Write

$$D_k = X_k^T X_k + n_k \lambda_k I_p,$$

$$A_{l,k} = X_l \left(X_k^T X_k + n_k \lambda_k I_p\right)^{-1} X_k^T = X_l D_k^{-1} X_k^T,$$

$$A_l = \begin{pmatrix} A_{l,1} & A_{l,2} & \cdots & A_{l,K} \end{pmatrix},$$

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,K} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ A_{K,1} & A_{K,2} & \cdots & A_{K,K} \end{pmatrix} \quad \text{and} \quad \overline{W} = \begin{pmatrix} w_1 I_{n_1} & 0 & \cdots & 0 \\ 0 & w_2 I_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_K I_{n_K} \end{pmatrix}.$$

Define the global squared error loss function

$$L(W) = \|\mu - \hat{\mu}(W)\|^2, \tag{9}$$

and the corresponding global risk function

$$R(W) = \mathbf{E}[L(W)|X],$$

where $\mu = \mathbf{E}(Y|X)$ is the conditional expectation of Y given X, and $\hat{\mu}(W) = X\hat{\beta}(W)$ is the model average estimator of μ. In the following, we will show that $HC_g(W)$ is an unbiased estimator of $R(W) + tr(\Omega)$, and the weight vectors $\hat{W}^*$ and $\hat{W}$ are asymptotically equivalent to the infeasible optimal weight vector under heteroscedastic and homoscedastic errors respectively.

We can write the model average estimator of μ as

$$\hat{\mu}(W) = \left(X_l \sum_{k=1}^{K} w_k \hat{\beta}_k(\lambda_k) \right)_{l=1,2,\ldots,K} = A\overline{W}Y.$$

Hence the loss and risk functions based on the full data are

$$L(W) = \mu^T \left(A\overline{W} - I\right)^T \left(A\overline{W} - I\right)\mu + \varepsilon^T \overline{W} A^T A \overline{W} \varepsilon + 2\varepsilon^T \overline{W} A^T \left(A\overline{W} - I\right)\mu \tag{10}$$

and

$$R(W) = \mu^T \left(A\overline{W} - I\right)^T \left(A\overline{W} - I\right)\mu + tr(\overline{W} A^T A \overline{W} \Omega). \tag{11}$$

respectively. The Global Mallows C_L criterion $HC_g(W|X)$ is an unbiased estimator of $R(W) + N\sigma^2$, that is

$$\begin{aligned} HC_g(W) &= \left\|Y - X\hat{\beta}(W)\right\|^2 + 2tr\left(A\overline{W}\Omega\right) \\ &= L(W) + \varepsilon^T\varepsilon - 2\varepsilon^T\left(A\overline{W} - I\right)\mu - 2\varepsilon^T A\overline{W}\varepsilon + 2tr\left(A\overline{W}\Omega\right). \end{aligned} \tag{12}$$

Lemma 1.

$$\mathbf{E}\left\{HC_g(W)|X\right\} = R(W) + tr(\Omega) \tag{13}$$

Denote $\xi_N = \inf_{W \in H} R(W)$. Consider the following conditions:

Condition 1 $\max_{1 \leqslant i \leqslant N} \mathbf{E}\left(|\varepsilon_i|^{4G}\right) \leq \kappa < \infty$, with $1 \leq G < \infty$ being a fixed integer.

Condition 2 $\frac{1}{K}\sum_{l=1}^{K} \lambda_{\max}\left(D_l^{-1} X_l^T X_l D_l^{-1} \frac{1}{K} \sum_{k=1}^{K} X_k^T X_k\right)$ has a uniform upper bound.

Condition 3 $K^2 \xi_N^{-2G} \sum_{k=1}^{K} R^G(W_k^0) \to 0$, where W_k^0 represents a unit vector of $K \times 1$ with the kth element being 1 and otherwise 0.

Some discussions of the above conditions are in order. Condition 1 is a common condition for the error terms (Li, 1987; Hansen, 2007; Wan et al., 2010). Condition 2 is a constraint on the original data, selected tuning parameters, and partition strategies. Some similarities among subsets of data are implicit in this condition. An analog condition to our Condition 2 is Condition [C1] of Xu et al. (2019). If $K = 1$, we have $\frac{1}{K}\sum_{l=1}^{K} \lambda_{\max}\left(D_l^{-1} X_l^T X_l D_l^{-1} \frac{1}{K} \sum_{k=1}^{K} X_k^T X_k\right) \leq 1$. When the subsets of data are similar, we can guess that $\frac{1}{K}\sum_{l=1}^{K} \lambda_{\max}\left(D_l^{-1} X_l^T X_l D_l^{-1} \frac{1}{K} \sum_{k=1}^{K} X_k^T X_k\right) \leq 1$, and thus Condition 2 is valid. When the dimension of covariates $p \leq N_k, k = 1, 2, \ldots, K$, we shall assume that $X_k, k = 1, 2 \ldots, K$, are diagonal, i.e., $\gamma_{k,i}$ be the ith diagonal element of X_k. This is universal, seeing Section 2 of Li (1986) for example. Then Condition 2 is simply that

$$\frac{1}{K}\sum_{l=1}^{K} \lambda_{\max}\left(D_l^{-1} X_l^T X_l D_l^{-1} \frac{1}{K} \sum_{k=1}^{K} X_k^T X_k\right) = \frac{1}{K}\sum_{l=1}^{K} \max_{i \in \{1,2\ldots,n\}} \left\{ \frac{\gamma_{l,i}^2}{\left(\gamma_{l,i}^2 + \lambda_l\right)^2} \frac{1}{K} \sum_{k=1}^{K} \gamma_{k,i}^2 \right\}.$$

If $\frac{1}{K}\sum_{k=1}^{K}\gamma_{k,i}^{2} \leq \gamma_{l,i}^{2}+\lambda_l$ for all l, we can expect that $\frac{1}{K}\sum_{l=1}^{K}\lambda_{\max}\left(D_l^{-1}X_l^TX_lD_l^{-1}\frac{1}{K}\sum_{k=1}^{K}X_k^TX_k\right) < 1$, where Condition 2 holds.

Condition 3 simply means there is no subset data can give a ridge estimator, which is consistent for the true model or whose bias is zero, and similar conditions are found in the model averaging literature, for example, (8) of Wan et al. (2010) and Condition [C3] of Xu et al. (2019). It is obvious to see that Condition 3 is easy to hold for fixed K. Condition 3 also restricts the growth rate of K, the number of subsets, which should not be too fast. **(I can't see how this condition can guarantee the level of estimation accuracy, answer: maybe K being small means n being large for fixed N, I delete it since it is not clear).** Especially, Condition 3 tentatively indicates that K should not increase faster too more than that of each of sample size N_k. See for example. If we set the sample size of each subset is a fixed n, when K is increasing, then $\sum_{k=1}^{K}R^G(W_k^0)$ is increasing but ξ_N is decreasing, therefore Condition 3 is getting harder and harder to hold. In addition, when the true model is infinite-dimensional, all working models with a finite number of covarites are misspecified, and Condition 3 holds automatically.

Theorem 1. *Assume that Conditions 1-3 hold. Then*

$$\frac{L\left(\hat{W}^*\right)}{\inf_{W\in H}L(W)} \xrightarrow{p} 1. \tag{14}$$

Proof. See the Appendix. □

Condition 4 **(i)** $\tilde{p} = \max\limits_{1\leqslant k\leqslant K}\max\limits_{1\leqslant i\leqslant N_k}P_{k,ii} = o_p(1)$.

(ii) $\tilde{p}\sum_{k=1}^{K}\mu_k^T(I-P_k)(I-P_k)\mu_k/\xi_N = o_p(1)$.

(iii) $s = \sqrt{\max\limits_{1\leqslant k\leqslant K}\lambda_{\max}\left(\{I_{N_k}-P_k\}^2\right)}$ and $\xi_N^{-2G}(s\tilde{p})^{2G}\sum_{k=1}^{K}\left\{\mu_k^T(I-P_k)(I-P_k)\mu_k\right\}^G = o_p(1)$.

(iv) $\xi_N^{-1}\tilde{p}\sum_{k=1}^{K}\left(\varepsilon_k^TP_kP_k\varepsilon_k\right) = o_p(1)$, $\xi_N^{-1}\tilde{p}\sum_{k=1}^{K}tr(P_k\Omega_k) = o_p(1)$ and $\xi_N^{-2G}\tilde{p}^{2G}\left[tr(P_k^2)\right]^G = o_p(1)$.

Condition 4 guarantees that the estimator for Ω must be consistent. It requires us to choose the appropriate tuning parameters $\lambda_k, k = 1, 2, \ldots, K$ to ensure that Condition 4 holds. Similar conditions are given in Assumptions 2.4-2.7 of Liu and Okui (2013).

Theorem 2. *Assume that* (5)*,*

$$\max_k N_k\left\{\min_k N_k\right\}^{-1} = O(1), \tag{15}$$

and Conditions 1-4 hold, then

$$\frac{L\left(\hat{W}\right)}{\inf_{W\in H}L(W)} \xrightarrow{p} 1, \tag{16}$$

where

$$\hat{W} = (\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_K)^T = \arg\min_{w\in H}\widehat{hC}_g(W).$$

Proof. See the Appendix. □

Theorem 3. *Let Conditions 1-4 hold, and assume that* (7)*,*

$$\xi_N^{-2G}\tilde{p}^{2G}\max_k N_k^{2G}N^{-G} = o_p(1). \tag{17}$$

Then we have

$$\frac{L\left(\hat{W}'\right)}{\inf_{W\in H}L(W)} \xrightarrow{p} 1, \tag{18}$$

where

$$\hat{W}' = \left(\hat{w}'_1, \hat{w}'_2, \ldots, \hat{w}'_K\right)^T = \arg\min_{w\in H}\widehat{C}_g(W).$$

Proof. See the Appendix. □

3.2. Weight consistency

In this subsection, we investigate the consistency of the estimated weight vector $\widehat{W}$. We first give some notations. Let $L_k(\beta) = ||Y_k - X_k\boldsymbol{\beta}||^2 + \lambda_k||\boldsymbol{\beta}||^2$, then the parameter estimator $\widehat{\boldsymbol{\beta}}_k(\lambda_k)$ can be defined as the solution which minimizes the objective function $L_k(\boldsymbol{\beta})$. For the case of heteroscedastic distributed subjects, we define the pseudo-true parameter

$$\boldsymbol{\beta}_k^*(\lambda_k) = \arg\min_{\boldsymbol{\beta}} \mathrm{E}\{L_k(\boldsymbol{\beta})\}.$$

Let $\boldsymbol{\varepsilon}_k = (\varepsilon_{k,1}, ..., \varepsilon_{k,N_k})^T$, $\widehat{\boldsymbol{\mathcal{P}}}_k = (\widehat{\boldsymbol{\mu}}_{k,1}, ..., \widehat{\boldsymbol{\mu}}_{k,N_k})^T$ with $\widehat{\boldsymbol{\mu}}_{k,i} = \left(\boldsymbol{x}_{k,i}^T\widehat{\boldsymbol{\beta}}_1(\lambda_1), ..., \boldsymbol{x}_{k,i}^T\widehat{\boldsymbol{\beta}}_K(\lambda_K)\right)^T$, and $\boldsymbol{\mathcal{P}}_k = (\mu_{k,1}, ..., \mu_{k,N_k})$. Hence, $HC_g(W)$ can be rewritten as

$$\begin{aligned}
HC_g(W) &= \sum_{k=1}^K ||Y_k - X_k\widehat{\boldsymbol{\beta}}(W)||^2 + 2\sum_{k=1}^K w_k\sigma_k^2 tr(P_k)\\
&= \sum_{k=1}^K \left\| Y_k - \sum_{m=1}^K w_m X_k \widehat{\boldsymbol{\beta}}_m(\lambda_m)\right\|^2 + 2\sum_{k=1}^K w_k\sigma_k^2 tr(P_k)\\
&= \sum_{k=1}^K ||Y_k - \widehat{\boldsymbol{\mathcal{P}}}_k W||^2 + 2\sum_{k=1}^K w_k\sigma_k^2 tr(P_k).
\end{aligned}$$

Let $R(W) = E\{L(W)\} = \mathrm{E}\left(\left\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(W)\right\|^2\right) = \mathrm{E}\left\{\sum_{k=1}^K ||\boldsymbol{\mathcal{P}}_k - \widehat{\boldsymbol{\mathcal{P}}}_k W||^2\right\}$ and define the optimal weight vector as $W^0 = \arg\min_{W\in\mathcal{W}} R(W)$. Denote $\psi_N = \min_{W\in\mathcal{W}} R(W)$. Assume that W^0 is the interior point of $\mathcal{W}$.

To this end, we need the following regularity conditions with c and C being two generic constants.

Condition 1. $\mathrm{P}\left\{\lambda_{\min}\left(N_k^{-1}\widehat{\boldsymbol{\mathcal{P}}}_k^T\widehat{\boldsymbol{\mathcal{P}}}_k\right) > c\right\}$ *tends to one, and* $\lambda_{\max}\left(N_k^{-1}\widehat{\boldsymbol{\mathcal{P}}}_k^T\widehat{\boldsymbol{\mathcal{P}}}_k\right) = O_p(p)$ *uniformly in k* ($1 \le k \le K$).

Condition 2. $\max_{1\le k\le K} ||\boldsymbol{\beta}_k^*(\lambda_k)|| = O(p^{1/2})$ *and* $\max_{i,j}|x_{ij}| \le C$ *almost surely.*

Condition 3. $\lambda_{\max}\left(N_m^{-1}X_m^T X_m\right) \le Cp$ *and* $\lambda_{\min}\left(N_m^{-1}X_m^T X_m\right) \ge c$ *almost surely uniformly in m* ($1 \le m \le K$).

Condition 4. $\max_{1\le k\le K} \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k/N_k = O(1)$ *almost surely and* $\max_{1\le k\le K} \sigma_k^2 \le C$.

Condition 5. $\max_{1\le k\le K} p_k^* N_K = O(1)$ *almost surely, where* $p_k^* = \max_{1\le i\le N_k} P_{ii}^k$ *and* P_{ii}^k *is the* i^{th} *diagonal element of* P_k

Condition 6. $\psi_N^{1/2}/(N^\delta K p^{3/2}) = o(1)$ *for some* $0 < \delta < 1/2$.

Since $N_k^{-1}\widehat{\boldsymbol{\mathcal{P}}}_k^T\widehat{\boldsymbol{\mathcal{P}}}_k$ is usually a positive definite matrix, the first part of Assumption 1 is easily satisfied. Note that $\max_{1\le i\le N}|\mu_i|$ is often assumed to be bounded (see, for example, Chen et al. (2018)), then as an estimator of $\mu_{k,i}$, $\max_{1\le i\le N}|\boldsymbol{x}_{k,i}^T\widehat{\boldsymbol{\beta}}_m(\lambda_m)| = O_p(1)$ is mild for $k, m = 1, ..., K$. Hence, the second part of Assumption 1 can be reasonably satisfied. As $\boldsymbol{\beta}_k^*(\lambda_k)$ is a p-dimensional vector, the first part of Assumption 2 is reasonable. The second part of Assumption 2 is widely used condition imposed on the covariates x_{ij} (see also Wang et al. (2009) and Wang et al. (2012)). Assumption 3 is mild and commonly used in the linear model literature (Li et al. (2011)). The first part of Assumption 4 requires the average of $\mu_{k,i}$ to be bounded which is quite common and often used in the model averaging literature such as Wan et al. (2010) and Li et al. (2021). The second part of Assumption 4 is a wild bounded condition imposed on the variances in the case of heteroscedastic distributed subjects. In fact, Assumption 5 requires that P_{ii}^k be asymptotically negligible which excludes the extremely unbalanced case that a single observation remains relevant asymptotically. So Assumption 5 is easily satisfied and can be found in Zhao et al. (2020). Assumption 6 imposes restriction on $\{\psi_N, N, K, p\}$, which permits ψ_N to increase at the rate slower than $N^{2\delta}K^2p^3$.

Theorem 4. *Suppose that Assumptions 1-6 hold. Then, there exists a local minimizer* $\widehat{W}$ *of* $HC_g(W)$ *such that*

$$||\widehat{W} - W^0|| = O_p(N^{-\frac{1}{2}+\delta}Kp^2),$$

where δ *is a positive constant given in Assumption 6.*

Proof. See the Appendix. □

4. A simulation study

The purpose of this section is to examine the performance of the MARR estimator via a simulation study. We consider a similar setup to that of Zhao et al. (2020). We let the true model be $y_i = x_i^T\beta + \varepsilon_i$, where $\beta = (0.1, -0.2, 0.3, 0.5, 0.1, -0.3)^T$, and $x_i = (x_{i1}, x_{i2}, \cdots, x_{i6})^T$ is a six-dimensional normal random vector with mean 0 and covariance matrix Σ, with all entries on the diagonal set to 1 and other entries set to $\rho = 0.2, 0.5, 0.8$, and ε_i's are i.i.d. normal errors with mean 0 and variance 1.

The candidate models are set by the variables not only x_i, but also x_{i7}, where $x_{i7} = x_{i2} + x_{i3} + t \times u_i$ with u_i being from the standard normal distribution, which is independent of x_{i2} and x_{i3}, and t being 0.05, 0.2, and 0.5.

Let there be five data subsets, and the sample size of each subset be (Case 1) 25, 50, 75, 50, and 100, or (Case 2) 250, 500, 750, 500, and 100. We estimate λ_k by the method of Hoerl and Kennard (1970).

For the divide-and-conquer algorithms, we compare four methods. Two are homoscedasticity distributed subject methods: Mallows model averaging (MMA) method, and divide-and-conquer (DC) method, which computes the ridge regression estimator on each subset data and combines the local estimator by taking the sample averaging. Two are heteroscedastic distributed subject methods: heteroscedastic Mallows model averaging (HMMA), and variance divide-and-conquer (VDC) algorithm, which computes the ridge regression estimator on each subset data and aggregates the local estimator by a weighted averaging method. The weight vector is determined by $\hat{W}_{vdc} = (\hat{w}_{vdc,1}, \hat{w}_{vdc,2}, \cdots, \hat{w}_{vdc,K})^T$ with $\hat{w}_{vdc,k} = \frac{1/\hat{\sigma}_k^2}{\sum_{l=1}^K 1/\hat{\sigma}_l^2}, k = 1, 2, \ldots, K$. We also consider two method with using all the data in one go: least square method (LS), and feasible generalized least square method (GLS). For GLS, the estimator of σ_k is by (6). The details can be seen in Fomby et al. (1984)

We repeat the experiment 500 times. For convenience, the averaging risk was multiplied by 1000. The results are shown in the tables 2 and 3.

Table 2: Results for Case 1

	t	0.05	0.2	0.5
ρ =0.2	MMA	26.707	25.409	26.856
	DC	38.850	36.100	37.986
	HMMA	28.218	26.797	28.022
	VDC	43.182	39.818	41.811
	LS	27.294	25.644	26.591
	GLS	28.399	26.552	27.623
ρ =0.5	MMA	26.453	26.033	27.011
	DC	37.806	35.778	36.372
	HMMA	27.918	27.178	28.064
	VDC	42.468	39.813	39.836
	LS	26.918	26.815	26.808
	GLS	27.951	27.879	27.435
ρ =0.8	MMA	26.339	23.690	25.477
	DC	35.893	31.660	33.270
	HMMA	27.542	24.998	26.863
	VDC	39.985	34.994	36.672
	LS	27.107	25.151	26.582
	GLS	27.834	26.389	27.535

Experimental results show that our proposed method has some advantages over other methods. Similar results were found for local data.

Table 3: Results for Case 2

	t	0.02	0.2	0.5
ρ =0.2	MMA	8.137	8.193	8.231
	DC	15.276	15.387	15.725
	HMMA	8.279	8.365	8.402
	VDC	15.935	16.068	16.427
	LS	8.536	8.564	8.362
	GLS	8.597	8.713	8.490
ρ =0.5	MMA	8.083	7.832	8.143
	DC	15.563	14.326	14.882
	HMMA	8.244	7.992	8.310
	VDC	16.147	14.965	15.587
	LS	8.556	8.206	8.279
	GLS	8.685	8.297	8.379
ρ =0.8	MMA	7.755	8.125	8.457
	DC	14.529	13.786	14.876
	HMMA	7.892	8.260	8.615
	VDC	15.084	14.513	15.496
	LS	8.288	8.662	8.615
	GLS	8.371	8.755	8.710

As can be seen from the results, our distributed approach performs best in most cases.

5. Real Data analysis

We apply our method to the rental information data set, which contains about 20,000 pieces of data from Fangtianxia, 58.com and Ganji.com about the rents in Beijing, Shanghai and Shenzhen cities from the end of 2020 to the beginning of 2021. In addition, this data also integrates the information of the house like longitude, latitude and surrounding facilities information obtained by the Application Programming Interface form Baidu Map Corporation.

Our aim is to predict the rents for each house based on the information provided by the dataset. Hence we consider the following basic information variables for the rent house: numbers of rooms and bathrooms (RoomNumber and BathroomNumber, respectively), living and dining room total number (LivingDiningRoomNumber), area of the rental house (Area), toward direction (Toward), the number of floors the house is on (Floor), the total number of floors on which the house is to be rented (TotalFloor), a dummy variable indicating whether there is a balcony or not in the rental house (BalconyB, where 1 for the case that the rental house has, and 0 otherwise), Type of information publisher of the rent house (InformationDistributorType, taking 0 for published by real estate agent, and 0 for individual, such as landlord and the second hand owner), and ten dummy variables that represent it has beds, wardrobes, sofa, TV, refrigerator, washer, airconditioner, heater, broadband, fuel gas, and heating, or not (BedB, WardrobeB, sofaB, TVB, RefrigeratorB, WasherB, AirConditionerB, HeaterB, BroadbandB, FuelGasB, and HeatingB, respectively, with 1 for has and 0 for not),); and four variables that represent the surrounding environment: numbers of schools and hospitals within 4 kilometers and 6 kilometers, respectively (SurroundingSchoolsNumber and SurroundingHospitalsNumber, respectively), and the nearest distances to the school within 4 kilometers and to the hospital within 6 kilometers, respectively, (DistanceNearestSchool and DistanceNearestHospital, respectively, where -1 means there is no school within within 4 kilometers or no hospital within within 6 kilometers). Notice that Toward is 30 % missing, and then we remove this variable. Thereafter, we consider above 24 variables as explanatory variable and the rent price of house as response variable. Moreover, the dataset we studied has 4732 5469, and 8685 observations in Beijing, Shanghai and Shenzhen, respectively.

In the real data analysis, we randomly select 70% of all cities as the training set with size 13220 (3312 records for Beijing, 3828 records for Shanghai, and 6080 records for Shenzhen), and the rest as the testing data set with size 5666 (1420 records for Beijing, 1641 records for Shanghai, and 2605 records for Shenzhen). We apply our proposed Mallows' model averaging method to combine the ridge estimators estimated from three local subsets: Beijing, Shanghai, and Shenzhen.

We compare the out-of-sample prediction errors for three cities Beijing, Shanghai, and Shenzhen and the total sample of twenty-two methods: (i) three local methods, which only use the data in their sub-data set: least squared estimators(local-LSE), ridge regression with tuning parameter set by (Hoerl and Kennard, 1970)(local-Ridge), and ridge regression with tuning parameter selected by $[0, 0.1k, 0.1k, \dots, k, 2k, \lfloor 50\log(n_k)\rfloor k]$ (local-Ridge-s) with $\lfloor\cdot\rfloor$ being round off; (ii) three Mallows model averaging methods with three local-LSE, local-Ridge, and local-Ridge-s, respectively (MMA-LSE, MMA-Ridge, and MMA-Ridge-s); (iii) three DC methods with three local-LSE, local-Ridge, and local-Ridge-s, respectively (DC-LSE, DC-Ridge, and DC-Ridge-s); (iv) three averaging methods with three local estimators, where the weights are set by their sample sizes over the total sample size (SDC-LSE, SDC-Ridge, and SDC-Ridge-s); (v) Mallows model averaging methods under heteroscedastic distributed subjects with three local estimators(MMAh-LSE, MMAh-Ridge, and MMAh-Ridge-s); (vi) three VDC methods with three local estimators(VDC-LSE, VDC-Ridge, and VDC-Ridge-s); and four global methods: least squared estimator, generalized least squared estimator, global ridge regression with tuning parameter set by (Hoerl and Kennard, 1970)(Ridge), and global ridge regression with tuning parameter selected by$[0, 0.1k, 0.2k, \dots, k, 2k, \lfloor 50\log(n_k)\rfloor k]$ (Ridge-s).

We repeat every method 200 times. The results are summarized in Tables 4. From the table, we can see that:

1) All three local estimators often perform similarly, and the mean squared prediction errors of local-Ridge estimator in three cities are always the smallest.

2) The mean squared prediction errors of MMA type estimators are smaller than that of local type estimators, respectively. MMA-Ridge and MMA-Ridge-s estimators behave similarly.

3) All three DC type estimators are worse than their MMA type estimators, respectively, and even worse than their local type estimators in Beijing and Shanghai.

4) All SDC type estimators perform worse than their DC type estimators, respectively, and still the worst in all distributed methods.

5) MMAh type methods always obtain the smaller mean squared prediction errors than that of their MMA type methods, respectively. For Shenzhen, MMAh-Ridge obtain the smallest mean squared errors in all distributed estimators.

6) VDC type methods generally performs better than their SDC type methods, and worse than their DC, MMA and MMAh types methods.

7) Global estimators in different districts often perform better than all their distributed estimator except for Shenzhen Cities where MMAh-Ridge and MMAh-Ridge-s perform the best in all of the compared methods.

In summary, compared to DC, SDC, VDC, and global data methods, our proposed method MMAh-Ridge and MMA-Ridge are two efficient averaging methods in this real data analysis.

Rental data analysis for each city

Note that the rental data set also records the district of the city in which the rental house is located. So we take the data of each city as a whole and the data of different districts as sub-data sets, and apply our proposed model averaging methods. The results are summarized in Tables 5 - 7.

Table 4: Mean squared prediction errors for rental data

Method	Beijing	Shanghai	Shenzhen	Total
local-LSE	3.137	1.739	1.486	1.973
local-Ridge	3.135	1.738	1.486	1.972
local-Ridge-s	3.137	1.739	1.486	1.973
MMA-LSE	3.052	1.658	1.302	1.844
MMA-Ridge	3.106	1.739	1.188	1.828
MMA-Ridge-s	3.111	1.739	1.187	1.829
DC-LSE	3.237	1.886	1.380	1.992
DC-Ridge	3.236	1.885	1.380	1.991
DC-Ridge-s	3.237	1.886	1.380	1.992
SDC-LSE	3.331	1.999	1.386	2.051
SDC-Ridge	3.330	1.998	1.385	2.050
SDC-Ridge-s	3.331	1.999	1.386	2.051
MMAh-LSE	3.051	1.657	1.299	1.842
MMAh-Ridge	3.118	1.738	1.186	1.830
MMAh-Ridge-s	3.119	1.739	1.186	1.831
VDC-LSE	3.260	1.911	1.288	1.963
VDC-Ridge	3.259	1.910	1.287	1.962
VDC-Ridge-s	3.260	1.911	1.288	1.963
LS	2.954	1.552	1.270	1.774
GLS	2.973	1.575	1.220	1.762
Ridge	2.953	1.551	1.270	1.773
Ridge-s	2.954	1.552	1.270	1.774

5.1. Rental data analysis for Beijing City

For Beijing, the dataset includes 13 districts rental data and we use the 70% of each district together as the training dataset with size 3314 (221 for Dongcheng District, 243 for Fengtai District, 204 for Daxing District, 113 for Huairou District, 174 for Fangshang District, 217 for Changping District, 869 for Chaoyang District, 370 for Haidian District, 188 for Shijingshan District, 216 for Xicheng District, 156 for Tongzhou District, 185 for Mengtougaou District, and 158 for Shunyi District), and the rest as the testing data set with size 1418 (94 for Dongcheng District, 104 for Fengtai District, 87 for Daxing District, 49 for Huairou District, 75 for Fangshang District, 93 for Changping District, 372 for Chaoyang District, 158 for Haidian District, 81 for Shijingshan District, 92 for Xicheng District, 67 for Tongzhou District, 79 for Mengtougaou District, and 67 for Shunyi District). From Table 5, it can been seen that:

1) For local estimators, the mean squared prediction errors of local-Ridge estimator in different districts are not more than that of local-LSE estimator. local-Ridge and local-Ridge-s behave similarly, except for the data in Tongzhou district, of which the sample size is the smallest. When the sample size is small, the model selection for ridge estimator is unstable, which will lead to big prediction errors.

2) For MMA type methods, the mean squared prediction errors of MMA-Ridge estimator are smaller than that of MMA-LSE estimator. MMA-Ridge and MMA-Ridge-s estimators behave similarly and better than their local type estimator, respectively. The mean squared prediction errors of MMA-LSE estimators in different districts are often bigger than that of local-LSE estimator, except for Dongcheng, Chaoyang, Xicheng districts.

3) DC-Ridge always performs the best in all DC type estimators, and followed by DC-LSE. DC-Ridge-s always performs the worst, the reason may be that there are some different characteristics of the rental data in different districts, and simple averaging the data driven local-Ridge-s estimators will catch some useless information for the local districts data.

4) SDC-LSE and SDC-Ridge perform similarly, and not better than that of DC-LSE and DC-Ridge, respectively. While SDC-Ridge-s behaves the worst in all SDC methods, but better than that of DC-Ridge-s, the reason may be that the sample size proportional weight vector for local-Ridge-s estimators will obtain more stable estimator than that of simple averaging weight vector.

5) MMAh type methods always obtain the similar mean squared prediction errors with that of their MMA type methods, respectively, and for Huairou and Fangshan Districts, MMAh type methods perform better than their MMA type methods, and otherwise, vice versa.

6) VDC type methods generally performs worse than their MMA or MMAh method, and better than their DC and SDC estimators. Only VDC-Ridge in Xicheng District achieve the smallest mean squared prediction errors in all distributed methods.

7) Global estimators in different districts often perform worse than their local estimator except for Dongcheng, Chaoyang, and Xicheng Districs. The reason may be that the data characteristics of those three districts are similar to that of the overall districts of Beijing.

8) MMA-Ridge or MMAh-Ridge always performs the best in all methods for different districts in Beijing.

In summary, compared to DC, SDC, VDC, and global data methods, our proposed methods MMA-Ridge and MMAh-Ridge are two efficient averaging methods not only for subset rental data of but also for global rental data in Beijing.

5.2. Rental data analysis for Shanghai City

The dataset collects 12 districts rental data of Shanghai city. 70% sample size of each district together form the training dataset with size 3826 (211 for Jiading District, 143 for Fengxian District, 255 for Baoshan District, 396 for Xuhui District, 291 for Putuo District, 221 for Yangpu District, 933 for Pudong District, 145 for Hongkou District, 302 for Changning District, 410 for Minhang District, 244 for Jingan District, and 275 for Huangpu District), and the rest as the testing data set with size 1643 (91 for Jiading District, 61 for Fengxian District, 110 for Baoshan District, 170 for Xuhui District, 125 for Putuo District, 95 for Yangpu District, 400 for Pudong District, 62 for Hongkou District, 130 for Changning District, 176 for Minhang District, 105 for Jingan District, and 118 for Huangpu District). From Table 6, we can find that:

1) All three local estimators often behave similarly, and the mean squared prediction errors of local-Ridge estimator in different districts are always the smallest, followed by that of local-Ridge-s estimator.

2) For MMA type methods, the mean squared prediction errors in different districts of MMA-Ridge estimator are smaller than that of MMA-LSE estimator, except for Jiading, Fengxian, and Baoshan Districts. MMA-Ridge and MMA-Ridge-s estimators in different districts always behave similarly, and sometimes better than their local type estimator, respectively, except for that in Jiading, Fengxian, Baoshan, Pudong, yangpu and Hongkou Districts.

3) DC-Ridge always performs the best in all DC type estimators, and DC-LSE behaves nearly to DC-Ridge, while DC-Ridge-s always performs the worst. DC-Ridge in different districts often performs better than their MMA-Ridge and local-Ridge estimators, except for that in Jiading, Fengxian, Baoshan, Pudong, and Yangpu Districts.

4) SDC-LSE and SDC-Ridge perform similarly, and worse than that of DC-LSE and DC-Ridge, respectively. While SDC-Ridge-s behaves the worst in all SDC methods, and worse than that of DC-Ridge-s.

5) MMAh type methods always obtain the smaller mean squared prediction errors than that of their MMA type methods, respectively, except for Xuhui District, and MMAh-Ridge and MMAh-Ridge-s better than their MMAh-LSE.

6) VDC type methods always performs better than their DC and SDC estimators. For Jiading, Baoshan, Putuo, Yangpu, Pudong and Huangpu Districts, VDC-LSE and VDC-Ridge behave the best in all distributed methods.

7) Global estimators in different districts often perform worse than their local estimator except for Xuhui, Pudong, Jingan, and Huangpu Districs. The reason may be that the sample size of above districts are big enough and the data characteristics of those three districts are similar to that of the overall districts of Beijing.

8) MMA-Ridge or MMAh-Ridge always performs the best in all methods for different districts in Shanghai. DC-Ridge-s, SDC-Ridge-s, and VDC-Ridge-s methods are less stable than MMA-Ridge-s and MMAh-Ridge-s methods. In some districts like Jiading, Baoshan and Huangpu, VDC-LSE and VDC-Ridge behave also well.

In summary, our proposed methods MMA-Ridge and MMAh-Ridge are two efficient averaging methods for global rental data in Shanghai.

5.3. Rental data analysis for Shenzhen City

This rental dataset contain all 7 districts of Shenzhen City data. we still use the 70% of different districts as the training dataset with size 6081 (528 for Guangming District, 1067 for Nanshan District, 776 for Baoan District, 515 for Buji District, 380 for Yantian District, 818 for Futian District, 600 for Luohu District, 797 for Longhua District, and 600 for Longgang District), and the rest as the testing data set with size 2604 (226 for Guangming District, 457 for Nanshan District, 332 for Baoan District, 220 for Buji District, 163 for Yantian District, 351 for Futian District, 257 for Luohu District, 341 for Longhua District, and 257 for Longgang District). Table 7 shows that:

1) All three local estimators often perform similarly, and the mean squared prediction errors of local-LSE estimator in different districts are always the smallest.

2) For MMA type methods, the mean squared prediction errors of MMA-Ridge and MMA-Ridge-s estimators are smaller than that of MMA-LSE estimator. MMA-Ridge and MMA-Ridge-s estimators behave similarly and better than their local type estimator, respectively, except for that of Guangming and Longhua Districts. The mean squared prediction errors of MMA-LSE estimators in different districts are always bigger than that of local-LSE estimator, except for Nanshan District. The reason may be that the sample size of Nanshan District data is big enough and can effectively verify and utilize the local least squares estimators of other districts.

3) DC-LSE always performs the best in all DC type estimators, and followed by DC-Ridge. DC-LSE in different districts perform better than MMA-LSE, while DC-Ridge and DC-Ridge-s perform worse than their MMA type estimators, respectively, except for that in Nanshang District, where all three DC type estimators are better than their MMA type estimators, respectively.

4) All SDC type estimators perform not better than their DC type estimators, respectively, and still the worst in all distributed methods.

5) MMAh-Ridge always obtain the smallest mean squared errors in all MMAh type estimators. MMAh type methods always obtain the smaller mean squared prediction errors than that of their MMA type methods, respectively, and except for Nanshan, Luohuand and Longhua Districts, where MMAh-Ridge and MMA-Ridge methods perform nearly well.

6) VDC type methods generally performs better than their DC and SDC methods, while worse than their MMA and MMAh estimators, except for Nanshang District where vice versa, and Luohu and Longhua Districts where VDC-Ridge performs the best in all distributed methods.

7) Global estimators in different districts often perform worse than their local estimator except for Nanshang District. The reason may be that the data characteristics of Nanshang District is similar to that of the overall districts of Shenzhen and the sample size of Nanshang District is the biggest compared with that of other districts.

8) MMAh-Ridge always performs the best in all methods for different districts in Shenzhen, and followed by MMA-Ridge estimator. For Luohu and Longhua Districts, VDC-Ridge performs the best. indicate that there is heteroscedasticity between the data of different districts in Shenzhen, which is consistent with the reality.

In summary, compared to DC, SDC, VDC, and global data methods, our proposed method MMAh-Ridge is always the most efficient averaging methods and followed by MMA-Ridge.

6. Concluding remarks

We propose a suitable Mallows C_L averaging method for distributed data and prove the asymptotic optimality and consistency of the proposed algorithm. Numerical simulation and real data analysis show that our proposed methods always have advantages in reduce the mean squared prediction errors.

As for heteroscedastic data in each subsets, method in Liu and Okui (2013) can be calibrated and the the corresponding asymptotic properties can be studied. In addition, GCV criterion can also be considered (Li, 1987; Craven and Wahba, 1978) to determine weights, which does not require an estimator of the covariance matrix while Liu and Okui (2013) needs.

Appendix A. Proofs

Proof of Theorem 1

Following the formula(10) and (12), it can be obtained that

$$\begin{aligned} & HC_g(W) - L(W) - \varepsilon^T \varepsilon \\ & = -2\varepsilon^T \left(A\overline{W} - I\right)\mu - 2\varepsilon^T A\overline{W}\varepsilon + 2tr\left(A\overline{W}\Omega\right). \end{aligned} \tag{A.1}$$

With (A.1), Theorem 1 is available if we prove the following conclusions:

$$\sup_{W\in H} \left|\varepsilon^T \left(A\overline{W} - I\right)\mu\right| / R(W) \xrightarrow{p} 0, \tag{A.2}$$

$$\sup_{W\in H} \left|\varepsilon^T A\overline{W}\varepsilon - tr(A\overline{W}\Omega)\right| / R(W) \xrightarrow{p} 0, \tag{A.3}$$

Table 5: Mean squared prediction errors for rental data of Beijing

Method	Dongcheng	Fengtai	Daxing	Huairou	Fangshang	Changping	Chaoyang	Haidian	Shijingshan	Xicheng	Tongzhou	Mengtougaou	Shunyi	Total
local-LSE	17.829	1.892	0.166	0.445	0.016	0.154	4.706	1.032	0.088	8.153	0.697	0.217	2.480	3.403
local-Ridge	17.796	1.885	0.166	0.015	0.016	0.154	4.691	1.032	0.088	8.135	0.691	0.217	2.392	3.376
local-Ridge-s	17.829	1.892	0.166	0.015	0.016	0.154	4.706	1.032	0.088	8.153	1.860E+5	0.217	2.444	8.788E+3
MMA-LSE	11.443	2.364	0.829	0.456	0.459	0.421	3.774	1.189	0.502	3.761	2.059	1.993	2.763	2.785
MMA-Ridge	11.276	0.411	0.087	0.022	0.021	0.135	3.898	1.131	0.045	3.300	0.423	0.066	1.598	2.258
MMA-Ridge-s	11.276	0.411	0.087	0.022	0.021	0.135	3.898	1.131	0.045	3.300	0.423	0.066	1.598	2.258
DC-LSE	11.252	1.266	0.389	0.205	0.203	0.255	3.789	1.152	0.349	3.616	1.471	1.291	2.424	2.530
DC-Ridge	11.247	1.262	0.386	0.202	0.201	0.254	3.782	1.146	0.345	3.608	1.460	1.279	2.413	2.524
DC-Ridge-s	5.230E+2	1.182E+2	5.659E+2	2.512E+2	6.555E+2	1.325E+2	2.647E+2	3.114E+2	6.067E+2	9.355E+2	1.101E+3	9.332E+2	3.673E+2	4.509E+2
SDC-LSE	11.265	1.630	0.508	0.250	0.248	0.291	3.905	1.241	0.409	3.781	1.850	1.715	2.752	2.683
SDC-Ridge	11.264	1.630	0.508	0.250	0.249	0.291	3.898	1.237	0.407	3.774	1.840	1.704	2.741	2.678
SDC-Ridge-s	1.958E+2	4.428E+1	2.119E+2	9.409E+1	2.455E+2	4.963E+1	9.913E+1	1.166E+2	2.272E+2	3.503E+2	4.121E+2	3.495E+2	1.375E+2	1.689E+2
MMAh-LSE	11.446	2.365	0.831	0.459	0.461	0.423	3.766	1.185	0.505	3.758	2.053	1.986	2.757	2.782
MMAh-Ridge	11.411	0.433	0.093	0.015	0.015	0.139	3.912	1.146	0.048	3.320	0.443	0.079	1.613	2.278
MMAh-Ridge-s	11.411	0.433	0.093	0.015	0.015	0.139	3.912	1.146	0.048	3.320	0.443	0.079	1.613	2.278
VDC-LSE	11.403	0.562	0.181	0.109	0.111	0.210	3.927	1.191	0.136	3.383	0.551	0.209	1.681	2.339
VDC-Ridge	11.292	0.442	0.096	0.026	0.025	0.135	3.857	1.106	0.049	3.279	0.467	0.124	1.597	2.253
VDC-Ridge-s	7.596	1.719	8.238	3.659	9.520	1.934	3.855	4.529	8.808	1.360E+1	1.599E+1	1.355E+1	5.345	6.556
LS	12.111	3.472	1.181	0.545	0.550	0.512	3.996	1.336	0.545	3.919	2.289	2.236	2.940	3.066
GLS	10.787	0.617	0.187	0.119	0.116	0.168	3.830	1.087	0.103	3.253	0.710	0.470	1.650	2.274
Ridge	12.110	3.468	1.180	0.545	0.550	0.512	3.994	1.335	0.545	3.917	2.286	2.233	2.937	3.064
Ridge-s	12.111	3.472	1.181	0.545	0.550	0.512	3.996	1.336	0.545	3.919	2.289	2.236	2.940	3.066

Table 6: Mean squared prediction errors for rental data of Shanghai

Method	Jiading	Fengxian	Baoshan	Xuhui	Putuo	Yangpu	Pudong	Hongkou	Changning	Minhang	Jingan	Huangpu	Total
local-LSE	0.020	0.013	0.027	0.694	0.060	0.137	1.583	0.142	0.417	0.479	0.727	1.652	0.728
local-Ridge	0.020	0.013	0.027	0.693	0.058	0.137	1.583	0.134	0.417	0.478	0.726	1.588	0.723
local-Ridge-s	0.020	0.013	0.027	0.694	0.058	0.137	1.583	0.134	0.417	0.479	0.727	1.592	0.723
MMA-LSE	0.071	0.078	0.077	0.572	0.112	0.256	1.541	0.171	0.424	0.473	0.694	1.523	0.714
MMA-Ridge	0.109	0.140	0.113	0.494	0.099	0.161	1.496	0.134	0.409	0.434	0.664	1.181	0.662
MMA-Ridge-s	0.109	0.139	0.113	0.494	0.099	0.161	1.496	0.134	0.409	0.434	0.664	1.183	0.662
DC-LSE	0.048	0.058	0.054	0.538	0.082	0.169	1.473	0.129	0.393	0.428	0.641	1.239	0.650
DC-Ridge	0.047	0.053	0.052	0.535	0.078	0.164	1.463	0.128	0.392	0.424	0.637	1.239	0.646
DC-Ridge-s	5.191	0.053	8.588	8.335	0.079	1.740E+13	1.653E+1	0.128	0.392	8.051	6.748	6.004	8.481
SDC-LSE	0.055	0.062	0.061	0.549	0.092	0.205	1.493	0.141	0.398	0.443	0.656	1.350	0.671
SDC-Ridge	0.054	0.058	0.060	0.546	0.089	0.201	1.485	0.141	0.398	0.440	0.653	1.349	0.668
SDC-Ridge-s	4.324	0.058	7.154	6.944	0.089	1.450E+1	1.377E+1	0.141	0.398	6.707	5.621	5.002	7.065
MMAh-LSE	0.071	0.077	0.077	0.575	0.113	0.255	1.543	0.171	0.425	0.473	0.696	1.523	0.715
MMAh-Ridge	0.020	0.013	0.033	0.573	0.074	0.087	1.440	0.094	0.378	0.383	0.601	1.170	0.621
MMAh-Ridge-s	0.020	0.013	0.033	0.573	0.073	0.087	1.438	0.094	0.377	0.383	0.600	1.169	0.620
VDC-LSE	0.019	0.025	0.025	0.528	0.059	0.086	1.436	0.098	0.380	0.388	0.603	1.106	0.611
VDC-Ridge	0.019	0.025	0.025	0.528	0.059	0.086	1.437	0.098	0.380	0.389	0.603	1.107	0.611
VDC-Ridge-s	1.575	0.025	2.606	2.530	0.059	5.281	5.017	0.098	0.380	2.443	2.048	1.822	2.574
LS	0.071	0.074	0.078	0.580	0.114	0.262	1.546	0.174	0.424	0.480	0.703	1.541	0.719
GLS	0.030	0.030	0.039	0.557	0.072	0.140	1.432	0.114	0.374	0.407	0.616	1.185	0.627
Ridge	0.071	0.074	0.078	0.580	0.114	0.262	1.546	0.174	0.424	0.480	0.703	1.541	0.719
Ridge-s	0.071	0.074	0.078	0.580	0.114	0.262	1.546	0.174	0.424	0.480	0.703	1.541	0.719

Table 7: Mean squared prediction errors for rental data of Shenzhen

Method	Guangming	Nanshan	Baoan	Buji	Yantian	Futian	Luohu	Longhua	Longgang	Total
local-LSE	0.020	4.924	0.723	0.069	0.458	0.937	0.086	0.242	0.029	1.162
local-Ridge	0.020	4.922	0.722	0.069	1.225	0.936	0.086	0.241	0.029	1.209
local-Ridge-s	0.020	4.924	0.723	0.069	1.225	0.937	0.086	0.242	0.029	1.210
MMA-LSE	0.363	3.612	1.143	1.044	0.976	0.948	0.174	0.329	0.102	1.159
MMA-Ridge	0.027	3.621	0.201	0.056	0.093	0.638	0.073	0.214	0.032	0.798
MMA-Ridge-s	0.027	3.622	0.201	0.056	0.093	0.638	0.073	0.214	0.032	0.798
DC-LSE	0.154	3.489	0.524	0.408	0.387	0.679	0.094	0.242	0.053	0.889
DC-Ridge	0.171	3.422	0.566	0.456	0.425	0.667	0.097	0.247	0.065	0.891
DC-Ridge-s	0.171	3.423	0.566	0.457	0.425	0.667	0.097	0.247	0.065	0.891
SDC-LSE	0.215	3.531	0.657	0.544	0.513	0.738	0.109	0.258	0.060	0.950
SDC-Ridge	0.224	3.494	0.683	0.574	0.536	0.732	0.111	0.261	0.066	0.952
SDC-Ridge-s	0.224	3.495	0.683	0.574	0.537	0.732	0.111	0.261	0.066	0.952
MMAh-LSE	0.358	3.608	1.134	1.036	0.968	0.944	0.173	0.328	0.101	1.154
MMAh-Ridge	0.020	3.669	0.186	0.031	0.083	0.672	0.077	0.217	0.030	0.807
MMAh-Ridge-s	0.020	3.669	0.186	0.031	0.083	0.672	0.077	0.217	0.030	0.807
VDC-LSE	0.037	3.554	0.220	0.081	0.108	0.617	0.067	0.209	0.030	0.789
VDC-Ridge	0.041	3.507	0.227	0.091	0.112	0.597	0.062	0.205	0.031	0.779
VDC-Ridge-s	0.041	3.507	0.227	0.091	0.112	0.597	0.062	0.205	0.031	0.779
LS	0.291	3.739	1.345	1.246	1.168	1.065	0.200	0.354	0.110	1.252
GLS	0.031	3.550	0.245	0.114	0.130	0.598	0.069	0.213	0.036	0.794
Ridge	0.290	3.739	1.344	1.245	1.168	1.065	0.200	0.354	0.110	1.251
Ridge-s	0.291	3.739	1.345	1.246	1.168	1.065	0.200	0.354	0.110	1.252

and

$$\sup_{W\in H} |L(W)/R(W) - 1| \xrightarrow{p} 0. \tag{A.4}$$

According to Triangle inequality, Bonferroni inequality, Chebyshev inequality, and Whittle (1960), for any $\delta > 0$, there is

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^T\left(A\overline{W}-I\right)\mu\right|/R(W)>\delta\right\}\\
&\leqslant \mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^T\left(A\overline{W}-I\right)\mu\right|>\delta\xi_N\right\}\\
&=\mathbf{P}\left\{\sup_{W\in H}\left|\sum_{k=1}^{K}w_k\varepsilon^T\left(A\overline{W_k^0}-I\right)\mu\right|>\delta\xi_N\right\}\\
&\leqslant \mathbf{P}\left\{\sup_{W\in H}\sum_{k=1}^{K}w_k\left|\varepsilon^T\left(A\overline{W_k^0}-I\right)\mu\right|>\delta\xi_N\right\}\\
&=\mathbf{P}\left\{\max_{1\leqslant k\leqslant K}\left|\varepsilon^T\left(A\overline{W_k^0}-I\right)\mu\right|>\delta\xi_N\right\}\\
&\leqslant \sum_{k=1}^{K}\mathbf{P}\left\{\left|\varepsilon^T\left(A\overline{W_k^0}-I\right)\mu\right|>\delta\xi_N\right\}\\
&\leqslant \sum_{k=1}^{K}\mathbf{E}\left\{\frac{\left|\varepsilon^T\left(A\overline{W_k^0}-I\right)\mu\right|^{2G}}{\delta^{2G}\xi_N^{2G}}\right\}\\
&\leqslant C_1\delta^{-2G}\xi_N^{-2G}\sum_{k=1}^{K}\left\|\left(A\overline{W_k^0}-I\right)\mu\right\|^{2G}\\
&\leqslant C_1\delta^{-2G}\xi_N^{-2G}\sum_{k=1}^{K}R^G(W_k^0).
\end{aligned}$$

Hence, (A.2) holds with Condition 3. By the same way, we observe that

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^TA\overline{W}\varepsilon-tr\left(A\overline{W}\Omega\right)\right|/R(W)>\delta\right\}\\
&\leqslant \mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^TA\overline{W}\varepsilon-tr\left(A\overline{W}\Omega\right)\right|>\delta\xi_N\right\}\\
&\leqslant \sum_{k=1}^{K}\mathbf{P}\left\{\left|\varepsilon^TA\overline{W_k^0}\varepsilon-tr\left(A\overline{W_k^0}\Omega\right)\right|>\delta\xi_N\right\}\\
&\leqslant \sum_{k=1}^{K}\mathbf{E}\left\{\frac{\left|\varepsilon^TA\overline{W_k^0}\varepsilon-tr\left(A\overline{W_k^0}\Omega\right)\right|^{2G}}{\delta^{2G}\xi_N^{2G}}\right\}\\
&\leqslant C_1\delta^{-2G}\xi_N^{-2G}\sum_{k=1}^{K}\left[tr\left(\Omega\overline{W_k^0}A^TA\overline{W_k^0}\Omega\right)\right]^G\\
&\leqslant C_1\delta^{-2G}\xi_N^{-2G}\sum_{k=1}^{K}R^G(W_k^0).
\end{aligned} \tag{A.5}$$

Because of condition 3, then we deduce that (A.3) is true.

We will prove (A.4) in the next step. It is obvious that from the formulas (10) and (11), we only need to prove

$$\sup_{W\in H}\left\{\left|\varepsilon^T\overline{W}A^T\left(A\overline{W}-I\right)\mu\right|/R(W)\right\}\xrightarrow{p}0, \tag{A.6}$$

and

$$\sup_{W\in H}\left\{\left|\varepsilon^T\overline{W}A^TA\overline{W}\varepsilon-tr\left(\overline{W}A^TA\overline{W}\Omega\right)\right|/R(W)\right\}\xrightarrow{p}0. \tag{A.7}$$

Notice that $A\overline{W_l^0W_l^0}A^T=XD_l^{-1}X_l^TX_lD_l^{-1}X^T$, and then

$$\begin{aligned}&\frac{1}{K^2}\sum_{l=1}^K\lambda_{\max}\left(A\overline{W_l^0W_l^0}A^T\right)\\&=\frac{1}{K^2}\sum_{l=1}^K\lambda_{\max}\left(XD_l^{-1}X_l^TX_lD_l^{-1}X^T\right)\\&=\frac{1}{K}\sum_{l=1}^K\lambda_{\max}\left(D_l^{-1}X_l^TX_lD_l^{-1}\frac{1}{K}\sum_{k=1}^KX_k^TX_k\right),\end{aligned} \tag{A.8}$$

so $K^{-2}\lambda_{\max}\left(A\overline{W_l^0W_l^0}A^T\right)$ is uniformly bounded with the help of Condition 2. The next step, we have for (A.6)

$$\begin{aligned}&\mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^T\overline{W}A^T\left(A\overline{W}-I\right)\mu\right|/R(W)>\delta\right\}\\&\leqslant\mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^T\overline{W}A^T\left(A\overline{W}-I\right)\mu\right|>\delta\xi_N\right\}\\&=\mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^T\left(\sum_{l=1}^Kw_l\overline{W_l^0}\right)A^T\sum_{k=1}^Kw_k\left(A\overline{W_k^0}-I\right)\mu\right|>\delta\xi_N\right\}\\&\leqslant\mathbf{P}\left\{\sup_{W\in H}\sum_{l=1}^K\sum_{k=1}^Kw_lw_k\left|\varepsilon^T\overline{W_l^0}A^T\left(A\overline{W_k^0}-I\right)\mu\right|>\delta\xi_N\right\}\\&=\mathbf{P}\left\{\max_{1\leqslant l\leqslant K}\max_{1\leqslant k\leqslant K}\left|\varepsilon^T\overline{W_l^0}A^T\left(A\overline{W_k^0}-I\right)\mu\right|>\delta\xi_N\right\}\\&\leqslant\sum_{l=1}^K\sum_{k=1}^K\mathbf{P}\left\{\left|\varepsilon^T\overline{W_l^0}A^T\left(A\overline{W_k^0}-I\right)\mu\right|>\delta\xi_N\right\}\\&\leqslant\sum_{l=1}^K\sum_{k=1}^K\mathbf{E}\left\{\frac{\left|\varepsilon^T\overline{W_l^0}A^T\left(A\overline{W_k^0}-I\right)\mu\right|^{2G}}{\delta^{2G}\xi_N^{2G}}\right\}\\&\leqslant C_1\delta^{-2G}\xi_N^{-2G}\sum_{l=1}^K\sum_{k=1}^K\left\|\overline{W_l^0}A^T\left(A\overline{W_k^0}-I\right)\mu\right\|^{2G}\\&\leqslant C_1\delta^{-2G}\xi_N^{-2G}\left\{\sum_{l=1}^K\lambda_{\max}\left(A\overline{W_l^0W_l^0}A^T\right)\right\}\sum_{k=1}^KR^G(W_k^0)\\&=O\left(\xi_N^{-2G}K^2\sum_{k=1}^KR^G(W_k^0)\right)\to0,\end{aligned} \tag{A.9}$$

where the last inequality comes from condition 3.

It is also known that $tr\left(\overline{W_k^0}A^TA\overline{W_l^0W_l^0}A^TA\overline{W_k^0}\right) \leqslant \lambda_{\max}\left(A\overline{W_l^0W_l^0}A^T\right)tr\left(\overline{W_k^0}A^TA\overline{W_k^0}\right)$. Concerning (A.7), we find that

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^T\overline{W}A^TA\overline{W}\varepsilon - tr(\overline{W}A^TA\overline{W}\Omega)\right|/R(W) > \delta\right\}\\
&\leqslant \mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^T\overline{W}A^TA\overline{W}\varepsilon - tr(\overline{W}A^TA\overline{W}\Omega)\right| > \delta\xi_N\right\}\\
&= \mathbf{P}\left\{\sup_{W\in H}\left|\varepsilon^T\left(\sum_{l=1}^K w_l\overline{W_l^0}\right)A^TA\left(\sum_{k=1}^K w_k\overline{W_k^0}\right)\varepsilon\right.\right.\\
&\left.\left.-tr\left[\left(\sum_{l=1}^K w_l\overline{W_l^0}\right)A^TA\left(\sum_{k=1}^K w_k\overline{W_k^0}\right)\Omega\right]\right| > \delta\xi_N\right\}\\
&\leqslant \mathbf{P}\left\{\sup_{W\in H}\sum_{l=1}^K\sum_{k=1}^K w_lw_k\left|\varepsilon^T\overline{W_l^0}A^TA\overline{W_k^0}\varepsilon - tr\left(\overline{W_l^0}A^TA\overline{W_k^0}\Omega\right)\right| > \delta\xi_N\right\}\\
&= \mathbf{P}\left\{\max_{1\leqslant l\leqslant K}\max_{1\leqslant k\leqslant K}\left|\varepsilon^T\overline{W_l^0}A^TA\overline{W_k^0}\varepsilon - tr\left(\overline{W_l^0}A^TA\overline{W_k^0}\Omega\right)\right| > \delta\xi_N\right\}\\
&\leqslant \sum_{l=1}^K\sum_{k=1}^K\mathbf{P}\left\{\left|\varepsilon^T\overline{W_l^0}A^TA\overline{W_k^0}\varepsilon - tr\left(\overline{W_l^0}A^TA\overline{W_k^0}\Omega\right)\right| > \delta\xi_N\right\}\\
&\leqslant \sum_{l=1}^K\sum_{k=1}^K\mathbf{E}\left\{\frac{\left|\varepsilon^T\overline{W_l^0}A^TA\overline{W_k^0}\varepsilon - tr\left(\overline{W_l^0}A^TA\overline{W_k^0}\Omega\right)\right|^{2G}}{\delta^{2G}\xi_N^{2G}}\right\}\\
&\leqslant C_1\delta^{-2G}\xi_N^{-2G}\sum_{l=1}^K\sum_{k=1}^K tr\left(\overline{W_k^0}A^TA\overline{W_l^0W_l^0}A^TA\overline{W_k^0}\right)^G\\
&\leqslant C_1\delta^{-2G}\xi_N^{-2G}\left\{\sum_{l=1}^K\lambda_{\max}\left(A\overline{W_l^0W_l^0}A^T\right)\right\}\sum_{k=1}^K R^G(W_k^0)\\
&= O\left(\xi_N^{-2G}K^2\sum_{k=1}^K R^G(W_k^0)\right) \to 0.
\end{aligned}\tag{A.10}$$

Hence, with condition 3, we complete the proof of Theorem 1.

Proof of Theorem 2

Define $\overline{P}_k(W)$ as a diagonal matrix whose ith diagonal element is $w_kP_{k,ii}$. So $\sum_{k=1}^K w_kF_k = \sum_{k=1}^K tr\left(\overline{P}_k(W)\Omega_k\right)$. Since

$$\widehat{hC}_g(W) = hC_g(W) + 2\sum_{k=1}^K\left\{w_k\sum_{i=1}^{N_k}P_{k,ii}\left(\hat{\sigma}_k^2 - \sigma_k^2\right)\right\},$$

we only need to proof that

$$\sup_{W\in H}\left\{\left|\sum_{k=1}^K\left\{w_k\sum_{i=1}^{N_k}P_{k,ii}\left(\hat{\sigma}_k^2 - \sigma_k^2\right)\right\}\right|/R(W)\right\} \xrightarrow{p} 0. \tag{A.11}$$

It is easy to see that

$$
\begin{aligned}
&\sup_{W\in H}\left\{\left|\sum_{k=1}^{K}\left\{w_k\sum_{i=1}^{N_k}P_{k,ii}\left(\hat{\sigma}_k^2-\sigma_k^2\right)\right\}\right|/R(W)\right\}\\
&\leqslant \xi_N^{-1}\sup_{W\in H}\left\{\left|\sum_{k=1}^{K}\left\{w_k\sum_{i=1}^{N_k}P_{k,ii}\left(\hat{\sigma}_k^2-\sigma_k^2\right)\right\}\right|\right\}\\
&\leqslant \xi_N^{-1}\sup_{W\in H}\left\{\sum_{k=1}^{K}w_k\left|\sum_{i=1}^{N_k}P_{k,ii}\right|\left|\hat{\sigma}_k^2-\sigma_k^2\right|\right\}\\
&\leqslant \xi_N^{-1}\max_k N_k\tilde{p}\sup_{W\in H}\left\{\sum_{k=1}^{K}w_k\left|\hat{\sigma}_k^2-\sigma_k^2\right|\right\}\\
&\leqslant \xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|\hat{\sigma}_k^2-\sigma_k^2\right|\right\}\\
&= \xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}Y_k^T\left(I-P_k\right)^2Y_k-\sigma_k^2\right|\right\}\\
&\leqslant \xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\mu_k^T\left(I-P_k\right)^2\mu_k\right|\right\}+2\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\mu_k^T\left(I-P_k\right)^2\varepsilon_k\right|\right\}\\
&\quad+\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\varepsilon_k^T\left(I-P_k\right)^2\varepsilon_k-\sigma_k^2\right|\right\}\\
&\leqslant \xi_N^{-1}\frac{\max_k N_k}{\min_k N_k}\tilde{p}\left\{\sum_{k=1}^{K}\left|\mu_k^T\left(I-P_k\right)^2\mu_k\right|\right\}+2\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\mu_k^T\left(I-P_k\right)^2\varepsilon_k\right|\right\}\\
&\quad+\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\varepsilon_k^T\left(I-P_k\right)^2\varepsilon_k-\sigma_k^2\right|\right\}\\
&= o_p(1)+2\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\mu_k^T\left(I-P_k\right)^2\varepsilon_k\right|\right\}+\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\varepsilon_k^T\left(I-P_k\right)^2\varepsilon_k-\sigma_k^2\right|\right\},
\end{aligned}
$$

where the last equation is by Condition 4 (ii) and $\max_k N_k\left\{\min_k N_k\right\}^{-1}=O(1)$. Then we need to prove that

$$
\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\mu_k^T\left(I-P_k\right)^2\varepsilon_k\right|\right\}=o_p(1) \tag{A.12}
$$

and

$$
\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\varepsilon_k^T\left(I-P_k\right)^2\varepsilon_k-\sigma_k^2\right|\right\}=o_p(1), \tag{A.13}
$$

are valid.

Notice that $\max_k N_k\left\{\min_k N_k\right\}^{-1}=O(1)$, and $\mathbf{E}\left\{\mu_k^T\left(I-P_k\right)^2\varepsilon_k\right\}=0, k=1,2,\ldots,K$. Let $\gamma_{k,i}$ be the ith element of

$\mu_k^T(I-P_k)^2$. Then by Chebyshev's inequality and Theorem 2 of Whittle (1960), for any $\delta > 0$, we have

$$\begin{aligned}
&\mathbf{P}\left\{\tilde{p}\sum_{k=1}^{K}\left|\mu_k^T(I-P_k)^2\varepsilon_k\right|/\xi_N > \delta\right\}\\
&\leqslant \sum_{k=1}^{K}\mathbf{P}\left\{\left|\mu_k^T(I-P_k)^2\varepsilon_k\right| > \delta\xi_N\tilde{p}^{-1}\right\}\\
&\leqslant \frac{\sum_{k=1}^{K}\mathbf{E}\left|\mu_k^T(I-P_k)^2\varepsilon_k\right|^{2G}}{\delta^{2G}\xi_N^{2G}\tilde{p}^{-2G}}\\
&\leqslant \delta^{-2G}\xi_N^{-2G}\tilde{p}^{2G}\sum_{k=1}^{K}C_k\left\{\sum_{i=1}^{N_k}\gamma_{k,i}^2\left[\mathbf{E}(\varepsilon_{k,i}^{2G})\right]^{1/G}\right\}^G\\
&\leqslant \kappa\delta^{-2G}\xi_N^{-2G}\tilde{p}^{2G}\sum_{k=1}^{K}C_k\left\{\sum_{i=1}^{N_k}\gamma_{k,i}^2\right\}^G\\
&= \kappa\delta^{-2G}\xi_N^{-2G}\tilde{p}^{2G}\sum_{k=1}^{K}C_k\left\{\mu_k^T(I-P_k)^4\mu_k\right\}^G\\
&\leqslant \kappa\delta^{-2G}\xi_N^{-2G}\tilde{p}^{2G}\sum_{k=1}^{K}C_k\lambda_{max}^G\left((I-P_k)^2\right)\left\{\mu_k^T(I-P_k)(I-P_k)\mu_k\right\}^G\\
&\leqslant C\kappa\delta^{-2G}\xi_N^{-2G}\tilde{p}^{2G}s^{2G}\sum_{k=1}^{K}\left\{\mu_k^T(I-P_k)(I-P_k)\mu_k\right\}^G, \qquad \text{(A.14)}
\end{aligned}$$

which with Condition 4 (iii), (A.12) can be proved.

Next, we will prove (A.13). Notice that

$$\begin{aligned}
&N_k^{-1}\varepsilon_k^T(I-P_k)^2\varepsilon_k - \sigma_k^2\\
&= N_k^{-1}\varepsilon_k^T(I-P_k)^2\varepsilon_k - N_k^{-1}\varepsilon_k^T\varepsilon_k + N_k^{-1}\varepsilon_k^T\varepsilon_k - \sigma_k^2\\
&= -2N_k^{-1}\varepsilon_k^TP_k\varepsilon_k + N_k^{-1}\varepsilon_k^TP_k^2\varepsilon_k + N_k^{-1}\varepsilon_k^T\varepsilon_k - \sigma_k^2,
\end{aligned}$$

then we only need to prove that

$$\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\varepsilon_k^TP_k\varepsilon_k\right|\right\} = o_p(1), \qquad \text{(A.15)}$$

$$\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\varepsilon_k^TP_k^2\varepsilon_k\right|\right\} = o_p(1), \qquad \text{(A.16)}$$

and

$$\xi_N^{-1}\max_k N_k\tilde{p}\left\{\sum_{k=1}^{K}\left|N_k^{-1}\varepsilon_k^T\varepsilon_k - \sigma_k^2\right|\right\} = o_p(1), \qquad \text{(A.17)}$$

Since $\max_k N_k\{\min_k N_k\}^{-1} = O(1)$, with Condition 4 (iv), we have

$$\xi_N^{-1}\tilde{p}\sum_{k=1}^{K}\mathbf{E}\left\{\varepsilon_k^TP_k^2\varepsilon_k\right\} = \xi_N^{-1}\tilde{p}\sum_{k=1}^{K}tr\left(P_k^2\Omega_k\right) \leqslant \xi_N^{-1}\tilde{p}\max_k\sigma_k^2\sum_{k=1}^{K}tr\left(P_k^2\right) = o_p(1),$$

and for any $\delta > 0$

$$\begin{aligned}
&\mathbf{P}\left\{\xi_N^{-1} \max_k N_k \tilde{p} \left\{\sum_{k=1}^{K} \left|N_k^{-1} \varepsilon_k^T P_k \varepsilon_k\right|\right\} > \delta\right\} \\
&\leqslant \sum_{k=1}^{K} \left(\mathbf{P}\left\{\left|\varepsilon_k^T P_k \varepsilon_k - \mathbf{E}\left[\varepsilon_k^T P_k \varepsilon_k\right]\right| > \delta \xi_N \tilde{p}^{-1}\right\} + \mathbf{P}\left\{\left|\mathbf{E}\left(\varepsilon_k^T P_k \varepsilon_k\right)\right| > \delta \xi_N \tilde{p}^{-1}\right\}\right),
\end{aligned}$$

with $\sum_{k=1}^{K} \mathbf{P}\left\{\left|\mathbf{E}\left(\varepsilon_k^T P_k \varepsilon_k\right)\right| > \delta \xi_N \tilde{p}^{-1}\right\} = o(1)$ by Condition 4 (iv), and

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbf{P}\left\{\left|\varepsilon_k^T P_k \varepsilon_k - \mathbf{E}\left[\varepsilon_k^T P_k \varepsilon_k\right]\right| > \delta \xi_N \tilde{p}^{-1}\right\} \\
&\leqslant \delta^{-2G} \xi_N^{-2G} \tilde{p}^{2G} \sum_{k=1}^{K} \mathbf{E}\left\{\varepsilon_k^T P_k \varepsilon_k - \mathbf{E}\left[\varepsilon_k^T P_k \varepsilon_k\right]\right\}^{2G} \\
&\leqslant C \delta^{-2G} \xi_N^{-2G} \tilde{p}^{2G} \sum_{k=1}^{K} R(W_k^0)^G \to 0, \qquad \text{(A.18)}
\end{aligned}$$

where the last step is by Condition 2 and Condition 4 (i), and

$$\begin{aligned}
&\mathbf{P}\left\{\xi_N^{-1} \max_k N_k \tilde{p} \left\{\sum_{k=1}^{K} \left|N_k^{-1} \varepsilon_k^T P_k^2 \varepsilon_k\right|\right\} > \delta\right\} \\
&\leqslant \sum_{k=1}^{K} \left(\mathbf{P}\left\{\left|\varepsilon_k^T P_k^2 \varepsilon_k - \mathbf{E}\left[\varepsilon_k^T P_k^2 \varepsilon_k\right]\right| > \delta \xi_N \tilde{p}^{-1}\right\} + \mathbf{P}\left\{\left|\mathbf{E}\left(\varepsilon_k^T P_k^2 \varepsilon_k\right)\right| > \delta \xi_N \tilde{p}^{-1}\right\}\right) \\
&\leqslant \delta^{-2G} \xi_N^{-2G} \tilde{p}^{2G} \sum_{k=1}^{K} \mathbf{E}\left\{\varepsilon_k^T P_k^2 \varepsilon_k - \mathbf{E}\left[\varepsilon_k^T P_k^2 \varepsilon_k\right]\right\}^{2G} + \sum_{k=1}^{K} \mathbf{P}\left\{\left|\mathbf{E}\left(\varepsilon_k^T P_k^2 \varepsilon_k\right)\right| > \delta \xi_N \tilde{p}^{-1}\right\} \\
&\leqslant C \delta^{-2G} \xi_N^{-2G} \tilde{p}^{2G} \sum_{k=1}^{K} \sigma_k^2 \left[tr(P_k^2)\right]^G + \sum_{k=1}^{K} \mathbf{P}\left\{\left|\mathbf{E}\left(\varepsilon_k^T P_k^2 \varepsilon_k\right)\right| > \delta \xi_N \tilde{p}^{-1}\right\} \to 0.
\end{aligned}$$

Thus, (A.15) and (A.16) follow.

By $\max_k N_k \{\min_k N_k\}^{-1} = O(1)$, Condition 1, Condition 4 (i) and Law of Large Numbers, We deduce that (A.17) is true. Consequently, we complete the proof of (A.13). Hence, Theorem 2 is proved.

Proof of Theorem 3

Since

$$\widehat{C}_g(W) = C_g(W) + 2\left(\hat{\sigma}^2 - \sigma^2\right) \sum_{k=1}^{K} \left\{w_k \sum_{i=1}^{N_k} P_{k,ii}\right\},$$

we only need to proof that

$$\sup_{W \in H} \left\{\left|\left(\hat{\sigma}^2 - \sigma^2\right) \sum_{k=1}^{K} \left\{w_k \sum_{i=1}^{N_k} P_{k,ii}\right\}\right| / R(W)\right\} \xrightarrow{p} 0. \qquad \text{(A.19)}$$

It is obvious that

$$\begin{aligned}
&\sup_{W\in H}\left\{\left|\left(\hat{\sigma}^2-\sigma^2\right)\sum_{k=1}^{K}\left\{w_k\sum_{i=1}^{N_k}P_{k,ii}\right\}\right|/R(W)\right\}\\
&\leqslant \xi_N^{-1}\max_k N_k\tilde{p}\left|N^{-1}\sum_{k=1}^{K}Y_k^T\left(I-P_k\right)^2Y_k-\sigma^2\right|\\
&\leqslant \xi_N^{-1}\max_k N_k\tilde{p}N^{-1}\sum_{k=1}^{K}\mu_k^T\left(I-P_k\right)^2\mu_k+2\xi_N^{-1}\max_k N_k\tilde{p}\left|N^{-1}\sum_{k=1}^{K}\mu_k^T\left(I-P_k\right)^2\varepsilon_k\right|\\
&\quad+\xi_N^{-1}\max_k N_k\tilde{p}\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\left(I-P_k\right)^2\varepsilon_k-\sigma^2\right|\\
&=o_p(1)+2\max_k N_k\xi_N^{-1}\tilde{p}\left|N^{-1}\sum_{k=1}^{K}\mu_k^T\left(I-P_k\right)^2\varepsilon_k\right|+\xi_N^{-1}\max_k N_k\tilde{p}\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\left(I-P_k\right)^2\varepsilon_k-\sigma^2\right|,
\end{aligned}$$

where the last equation is by Condition 4 (ii) and $N^{-1}\max_k N_k \leqslant 1$. Then we only need to show

$$\xi_N^{-1}\max_k N_k\tilde{p}\left|N^{-1}\sum_{k=1}^{K}\mu_k^T\left(I-P_k\right)^2\varepsilon_k\right|=o_p(1), \tag{A.20}$$

and

$$\xi_N^{-1}\max_k N_k\tilde{p}\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\left(I-P_k\right)^2\varepsilon_k-\sigma^2\right|=o_p(1). \tag{A.21}$$

It is easy to see that (A.20) holds with Condition 4 (iii), $N^{-1}\max_k N_k\leqslant 1$ and (A.14).

For (A.21), with , we have

$$\begin{aligned}
&\xi_N^{-1}\max_k N_k\tilde{p}\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\left(I-P_k\right)^2\varepsilon_k-\sigma^2\right|\\
&\leqslant \xi_N^{-1}\max_k N_k\tilde{p}\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\left(I-P_k\right)^2\varepsilon_k-N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\varepsilon_k\right|+\xi_N^{-1}\max_k N_k\tilde{p}\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\varepsilon_k-\sigma^2\right|\\
&\leqslant \xi_N^{-1}\tilde{p}\max_k N_kN^{-1}\left|\sum_{k=1}^{K}\varepsilon_k^TP_k^2\varepsilon_k-2\sum_{k=1}^{K}\varepsilon_k^TP_k\varepsilon_k\right|+\xi_N^{-1}\tilde{p}\max_k N_k\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\varepsilon_k-\sigma^2\right|\\
&\leqslant \xi_N^{-1}\tilde{p}\left|\sum_{k=1}^{K}\varepsilon_k^TP_k^2\varepsilon_k-2\sum_{k=1}^{K}\varepsilon_k^TP_k\varepsilon_k\right|+\xi_N^{-1}\tilde{p}\max_k N_k\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\varepsilon_k-\sigma^2\right|.
\end{aligned}$$

By (A.18), we have $\xi_N^{-1}\tilde{p}\left|\sum_{k=1}^K\varepsilon_k^TP_k^2\varepsilon_k-2\sum_{k=1}^K\varepsilon_k^TP_k\varepsilon_k\right|=o_p(1)$. With the help of Chebshev inequality amd Theorem 2 of Whittle (1960), for any $\delta>0$, we have

$$\begin{aligned}
&\mathbf{P}\left\{\xi_N^{-1}\tilde{p}\max_k N_k\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\varepsilon_k-\sigma^2\right|>\delta\right\}\\
&\leqslant \xi_N^{-2G}\tilde{p}^{2G}\max_k N_k^{2G}\mathbf{E}\left|N^{-1}\sum_{k=1}^{K}\varepsilon_k^T\varepsilon_k-\sigma^2\right|^{2G}\\
&\leqslant C\xi_N^{-2G}\tilde{p}^{2G}\max_k N_k^{2G}N^{-G}\\
&=o_p(1).
\end{aligned} \tag{A.22}$$

So (A.21) follows. From the above results, Theorem 3 holds.

Proof of Theorem 4

Let $\alpha_N = N^{-\frac{1}{2}+\delta}Kp^2$ and $\boldsymbol{u} = (u_1, ..., u_K)'$ be an K-dimensional vector. Following the proofs of Chen et al. (2018) and Fan and Peng (2004) , to show Theorem 4, it suffices to prove that for any $\epsilon > 0$, there exists a constant C such that for large n,

$$\mathrm{P}\left\{\inf_{\|\boldsymbol{u}\|=C,(\boldsymbol{w}^0+\alpha_N\boldsymbol{u})\in\mathcal{W}} HC_g(W^0+\alpha_N\boldsymbol{u}) > HC_g(W^0)\right\} \geq 1-\epsilon. \tag{A.23}$$

Observe that

$$\begin{aligned}
&HC_g(W^0+\alpha_N\boldsymbol{u}) - HC_g(W^0)\\
=&\sum_{k=1}^{K}(Y_k-\widehat{\boldsymbol{\mathcal{P}}}_kW^0-\alpha_N\widehat{\boldsymbol{\mathcal{P}}}_k\boldsymbol{u})^T(Y_k-\widehat{\boldsymbol{\mathcal{P}}}_kW^0-\alpha_N\widehat{\boldsymbol{\mathcal{P}}}_k\boldsymbol{u}) - \sum_{k=1}^{K}(Y_k-\widehat{\boldsymbol{\mathcal{P}}}_kW^0)^T(Y_k-\widehat{\boldsymbol{\mathcal{P}}}_kW^0)\\
&+2\alpha_N\sum_{k=1}^{K}u_k\sigma_k^2tr(P_k)\\
=&\alpha_N^2\boldsymbol{u}^T\left(\sum_{k=1}^{K}\widehat{\boldsymbol{\mathcal{P}}}_k^T\widehat{\boldsymbol{\mathcal{P}}}_k\right)\boldsymbol{u} - 2\alpha_N\boldsymbol{u}^T\sum_{k=1}^{K}\widehat{\boldsymbol{\mathcal{P}}}_k^T(Y_k-\widehat{\boldsymbol{\mathcal{P}}}_kW^0) + 2\alpha_N\sum_{k=1}^{K}u_k\sigma_k^2tr(P_k)\\
=&\alpha_N^2\boldsymbol{u}^T\left(\sum_{k=1}^{K}\widehat{\boldsymbol{\mathcal{P}}}_k^T\widehat{\boldsymbol{\mathcal{P}}}_k\right)\boldsymbol{u} - 2\alpha_N\sum_{k=1}^{K}(\boldsymbol{\mathcal{P}}_k-\widehat{\boldsymbol{\mathcal{P}}}_kW^0)^T\widehat{\boldsymbol{\mathcal{P}}}_k\boldsymbol{u} - 2\alpha_N\sum_{k=1}^{K}(Y_k-\boldsymbol{\mathcal{P}}_k)^T\boldsymbol{\mathcal{P}}_k^*\boldsymbol{u}\\
&-2\alpha_N\sum_{k=1}^{K}(Y_k-\boldsymbol{\mathcal{P}}_k)^T(\widehat{\boldsymbol{\mathcal{P}}}_k-\boldsymbol{\mathcal{P}}_k^*)\boldsymbol{u} + 2\alpha_N\sum_{k=1}^{K}u_k\sigma_k^2tr(P_k)\\
\equiv&\Lambda_1+\Lambda_2+\Lambda_3+\Lambda_4+\Lambda_5,
\end{aligned}$$

where $\boldsymbol{\mathcal{P}}_k^*$ is the same as $\widehat{\boldsymbol{\mathcal{P}}}_k$ except that $\widehat{\boldsymbol{\beta}}_m(\lambda_m)$ in $\widehat{\boldsymbol{\mathcal{P}}}_k$ is replaced by $\boldsymbol{\beta}_m^*(\lambda_m)$ for all $m = 1, ..., K$. By Assumption 1, we have

$$\begin{aligned}
\Lambda_1 &\geq \alpha_N^2\sum_{k=1}^{K}\lambda_{\min}\left(\widehat{\boldsymbol{\mathcal{P}}}_k^T\widehat{\boldsymbol{\mathcal{P}}}_k\right)\|\boldsymbol{u}\|^2\\
&\geq c\alpha_N^2\sum_{k=1}^{K}N_k\|\boldsymbol{u}\|^2\\
&= cN\alpha_N^2\|\boldsymbol{u}\|^2.
\end{aligned} \tag{A.24}$$

Thus, (A.23) holds if Λ_2, Λ_3, Λ_4 and Λ_5 are asymptotically dominated by Λ_1.

We first consider Λ_2. According to Cauchy-Schwarz inequality, it is seen that

$$\begin{aligned}
|\Lambda_2| &\leq 2\alpha_N\sum_{k=1}^{K}\|\boldsymbol{\mathcal{P}}_k-\widehat{\boldsymbol{\mathcal{P}}}_kW^0\|\|\widehat{\boldsymbol{\mathcal{P}}}_k\boldsymbol{u}\|\\
&\leq 2\alpha_N\left(\sum_{k=1}^{K}\|\boldsymbol{\mathcal{P}}_k-\widehat{\boldsymbol{\mathcal{P}}}_kW^0\|^2\right)^{1/2}\left(\sum_{k=1}^{K}\|\widehat{\boldsymbol{\mathcal{P}}}_k\boldsymbol{u}\|^2\right)^{1/2}.
\end{aligned}$$

Since $\psi_N = \mathrm{E}\left\{\sum_{k=1}^{K} \|\boldsymbol{\mathcal{P}}_k - \widehat{\boldsymbol{\mathcal{P}}}_k W^0\|^2\right\}$, we have $\left(\sum_{k=1}^{K} \|\boldsymbol{\mathcal{P}}_k - \widehat{\boldsymbol{\mathcal{P}}}_k W^0\|^2\right)^{1/2} = O_p(\psi_N^{1/2})$. By Assumption 1, it is seen that

$$\begin{aligned}\sum_{k=1}^{K} \|\widehat{\boldsymbol{\mathcal{P}}}_k \boldsymbol{u}\|^2 &= \boldsymbol{u}^T \left(\sum_{k=1}^{K} \widehat{\boldsymbol{\mathcal{P}}}_k^T \widehat{\boldsymbol{\mathcal{P}}}_k\right) \boldsymbol{u} \\ &\leq \sum_{k=1}^{K} \lambda_{\max}\left(\widehat{\boldsymbol{\mathcal{P}}}_k^T \widehat{\boldsymbol{\mathcal{P}}}_k\right) \|\boldsymbol{u}\|^2 \\ &= O_p(Np)\|\boldsymbol{u}\|^2.\end{aligned}$$

Hence, we obtain

$$|\Lambda_2| = O_p(\alpha_N \psi_N^{1/2} N^{1/2} p^{1/2})\|\boldsymbol{u}\|. \tag{A.25}$$

Second, we consider Λ_3. Since $\mathrm{E}(\varepsilon_{k,i}|\boldsymbol{x}_{k,i}) = 0$, it is seen that

$$\mathrm{E}\left\{\varepsilon_{k,i}\boldsymbol{x}_{k,i}^T \boldsymbol{\beta}_m^*(\lambda_m)|\boldsymbol{x}_{k,i}\right\} = 0,$$

which implies that $\mathrm{E}\left\{\varepsilon_{k,i}\boldsymbol{x}_{k,i}^T \boldsymbol{\beta}_m^*(\lambda_m)\right\} = 0$. Note that $\varepsilon_{k,i}\boldsymbol{x}_{k,i}^T \boldsymbol{\beta}_m^*(\lambda_m)$, $i = 1, ..., N_k$, are mutually independent. Thus, by Assumption 2 and Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}&\mathrm{E}\left(\sum_{k=1}^{K} \|\boldsymbol{\mathcal{P}}_k^{*T}(Y_k - \boldsymbol{\mathcal{P}}_k)\|\right)^2 \\ \leq& K\sum_{k=1}^{K} \mathrm{E}\left(\|\boldsymbol{\mathcal{P}}_k^{*T}(Y_k - \boldsymbol{\mathcal{P}}_k)\|^2\right) \\ =& K\sum_{k=1}^{K}\sum_{m=1}^{K} \mathrm{E}\left(\sum_{i=1}^{N_k} \varepsilon_{k,i}\boldsymbol{x}_{k,i}^T \boldsymbol{\beta}_m^*(\lambda_m)\right)^2 \\ =& K\sum_{k=1}^{K}\sum_{m=1}^{K}\sum_{i=1}^{N_k} \mathrm{E}\left(\varepsilon_{k,i}\boldsymbol{x}_{k,i}^T \boldsymbol{\beta}_m^*(\lambda_m)\right)^2 \\ \leq& K\sum_{k=1}^{K}\sum_{m=1}^{K}\sum_{i=1}^{N_k} \mathrm{E}\left(\varepsilon_{k,i}\|\boldsymbol{x}_{k,i}\|\|\boldsymbol{\beta}_m^*(\lambda_m)\|\right)^2 \\ =& O(NK^2p^2).\end{aligned}$$

This indicates that

$$\sum_{k=1}^{K} \|\boldsymbol{\mathcal{P}}_k^{*T}(Y_k - \boldsymbol{\mathcal{P}}_k)\| = O_p(N^{1/2}Kp).$$

Thus, it is readily seen that

$$\begin{aligned}|\Lambda_3| &\leq 2\alpha_N \sum_{k=1}^{K} \|\boldsymbol{\mathcal{P}}_k^{*T}(Y_k - \boldsymbol{\mathcal{P}}_k)\|\|\boldsymbol{u}\| \\ &= O_p(\alpha_N N^{1/2}Kp).\end{aligned} \tag{A.26}$$

We now consider Λ_4. By Assumption 2 and Cauchy-Schwarz inequality, we find that

$$\begin{aligned}
\sum_{k=1}^{K} \|\widehat{\boldsymbol{\mathcal{P}}}_k - \boldsymbol{\mathcal{P}}_k^*\|^2 &= \sum_{k=1}^{K}\sum_{m=1}^{K}\sum_{i=1}^{N_k} \left\{\boldsymbol{x}_{k,i}^T(\widehat{\boldsymbol{\beta}}_m(\lambda_m) - \boldsymbol{\beta}_m^*(\lambda_m))\right\}^2 \\
&\leq \sum_{k=1}^{K}\sum_{m=1}^{K}\sum_{i=1}^{N_k} \|\boldsymbol{x}_{k,i}\|^2 \|\widehat{\boldsymbol{\beta}}_m(\lambda_m) - \boldsymbol{\beta}_m^*(\lambda_m)\|^2 \\
&\leq CNp \sum_{m=1}^{K} \|\widehat{\boldsymbol{\beta}}_m(\lambda_m) - \boldsymbol{\beta}_m^*(\lambda_m)\|^2. \qquad \text{(A.27)}
\end{aligned}$$

By the definition of $\widehat{\boldsymbol{\beta}}_m(\lambda_m)$, we can write

$$\begin{aligned}
&\widehat{\boldsymbol{\beta}}_m(\lambda_m) - \boldsymbol{\beta}_m^*(\lambda_m) \\
=&(X_m^T X_m + \lambda_m \mathrm{I}_p)^{-1}[X_m^T\{Y_m - X_m\boldsymbol{\beta}_m^*(\lambda_m)\} - \lambda_m\boldsymbol{\beta}_m^*(\lambda_m)] \\
=&(X_m^T X_m + \lambda_m \mathrm{I}_p)^{-1} \sum_{i=1}^{N_m} \left[\left\{y_{m,i} - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_m^*(\lambda_m)\right\}\boldsymbol{x}_{m,i} - \lambda_m\boldsymbol{\beta}_m^*(\lambda_m)\right]. \qquad \text{(A.28)}
\end{aligned}$$

Let $\xi_{i(m)} = \left\{y_{m,i} - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_m^*(\lambda_m)\right\}\boldsymbol{x}_{m,i} - \lambda_m\boldsymbol{\beta}_m^*(\lambda_m)$ be a p-dimensional vector. By the definition of $\boldsymbol{\beta}_m^*(\lambda_m)$, we obtain $\mathrm{E}(\xi_{i(m)}) = 0$. Further, it is obvious that $\xi_{i(m)}$ $(i = 1, ..., n)$ are mutually independent. So it follows from Assumption 2 that

$$\begin{aligned}
\mathrm{E}\left(\max_{1\leq m\leq K}\left\|\sum_{i=1}^{N_m}\xi_{i(m)}\right\|^2\right) &\leq \sum_{m=1}^{K}\mathrm{E}\left(\sum_{i=1}^{N_m}\xi_{i(m)}^{'}\sum_{j=1}^{N_m}\xi_{j(m)}\right) \\
&= \sum_{m=1}^{K}\sum_{i=1}^{N_m}\mathrm{E}\left(\|\xi_{i(m)}\|^2\right) \\
&\leq C\sum_{m=1}^{K}\sum_{i=1}^{N_m}\mathrm{E}\left(|y_{m,i} - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_m^*(\lambda_m)|^2\|\boldsymbol{x}_{m,i}\|^2 + \lambda_m^2\|\boldsymbol{\beta}_m^*(\lambda_m)\|^2\right) \\
&\leq Cp\sum_{m=1}^{K}\sum_{i=1}^{N_m}\mathrm{E}\left(|y_{m,i} - \boldsymbol{x}_{m,i}^T\boldsymbol{\beta}_m^*(\lambda_m)|^2\right) \\
&\leq Cp\sum_{m=1}^{K}\sum_{i=1}^{N_m}\mathrm{E}\left(\varepsilon_{m,i}^2 + \mu_{m,i}^2 + \|\boldsymbol{x}_{m,i}\|^2\|\boldsymbol{\beta}_m^*(\lambda_m)\|^2\right) \\
&= O(Np^3),
\end{aligned}$$

which implies that

$$\max_{1\leq m\leq K}\left\|\sum_{i=1}^{N_m}\xi_{i(m)}\right\|^2 = O_p(Np^3).$$

Hence, it is seen from (A.28) and Assumption 3 that

$$\begin{aligned}
&\sum_{m=1}^{K} \|\widehat{\boldsymbol{\beta}}_m(\lambda_m) - \boldsymbol{\beta}_m^*(\lambda_m)\|^2 \\
=&\sum_{m=1}^{K} \left(\sum_{i=1}^{N_m} \xi_{i(m)}\right)^T (X_m^T X_m + \lambda_m \mathrm{I}_p)^{-2} \left(\sum_{i=1}^{N_m} \xi_{i(m)}\right) \\
=&\sum_{m=1}^{K} \lambda_{\max}^2 \left\{(X_m^T X_m + \lambda_m \mathrm{I}_p)^{-1}\right\} \max_{1\le m\le K} \left\|\sum_{i=1}^{N_m} \xi_{i(m)}\right\|^2 \\
=&O_p(N^{-1}p^3).
\end{aligned}$$

Thus, according to (A.27), we obtain $\sum_{k=1}^{K} \|\widehat{\mathcal{P}}_k - \mathcal{P}_k^*\|^2 = O_p(p^4)$. Recognizing that $\sum_{k=1}^{K} \|Y_k - \mathcal{P}_k\|^2 = O_p(N)$, we have

$$\begin{aligned}
|\Lambda_4| &\le 2\alpha_N \sum_{k=1}^{K} \|Y_k - \mathcal{P}_k\| \|(\widehat{\mathcal{P}}_k - \mathcal{P}_k^*)\boldsymbol{u}\| \\
&\le 2\alpha_N \left(\sum_{k=1}^{K} \|Y_k - \mathcal{P}_k\|^2\right)^{1/2} \left(\sum_{k=1}^{K} \|(\widehat{\mathcal{P}}_k - \mathcal{P}_k^*)\boldsymbol{u}\|^2\right)^{1/2} \\
&= O_p(\alpha_N \sqrt{N} p^2)\|\boldsymbol{u}\|.
\end{aligned} \tag{A.29}$$

Finally, we consider Λ_5. By Assumption 5, we see that $\max_{1\le k\le K} tr(P_k) = O(1)$ almost surely. Hence, with Assumption 4 and Cauchy-Schwarz inequality, it is seen that

$$|\Lambda_5| \le 2\alpha_N \max_{1\le k\le K} tr(P_k) \max_{1\le k\le K} \widehat{\sigma}_k^2 K^{1/2} \|\boldsymbol{u}\|.$$

Observe that

$$\begin{aligned}
|\widehat{\sigma}_k^2| &= \left|\frac{Y_k^T (I - P_k)^2 Y_k}{N_k}\right| \\
&\le \left|\frac{\varepsilon_k^T (I - P_k)^2 \varepsilon_k}{N_k}\right| + \left|\frac{2\varepsilon_k^T (I - P_k)^2 \mu_k}{N_k}\right| + \left|\frac{\mu_k^T (I - P_k)^2 \mu_k}{N_k}\right| \\
&\equiv \Pi_{1k} + \Pi_{2k} + \Pi_{3k}.
\end{aligned}$$

By Assumption 5, we have $\lambda_{\max}\left\{(I - P_k)^2\right\}$ is bounded. So we obtain

$$\begin{aligned}
\mathrm{E}\left(\max_{1\le k\le K} \Pi_{1k}^2\right) &\le \sum_{k=1}^{K} \mathrm{E}\left(\Pi_{1k}^2\right) \\
&\le C \sum_{k=1}^{K} \mathrm{E}\left(\|\varepsilon_k\|^4\right) / N_k^2 \\
&\le C \sum_{k=1}^{K} \mathrm{E}\left(N_k \sum_{i=1}^{N_k} \varepsilon_{k,i}^4\right) / N_k^2 \\
&= O(K),
\end{aligned}$$

which implies that $\max_{1\le k\le K} \Pi_{1k} = O_p(K^{1/2})$. From Theorem 2 of Whittle (1960), Chebyshev's inequality, and Assumptions 4 and 6, we see that for any $\delta > 0$,

$$\begin{aligned}
&\mathrm{P}\left(\max_{1\le k\le K}\left|\frac{\varepsilon_k^T(I-P_k)^2\mu_k}{N_k}\right| > \delta|X\right) \\
&\le \sum_{k=1}^{K}\mathrm{P}\left(\left|\frac{\varepsilon_k^T(I-P_k)^2\mu_k}{N_k}\right| > \delta|X\right) \\
&\le \sum_{k=1}^{K}\mathrm{E}\left\{(\varepsilon_k^T(I-P_k)^2\mu_k)^2|X\right\}/N_k^2\delta^2 \\
&\le \sum_{k=1}^{K}\frac{1}{N_k^2\delta^2}\mu_k^T(I-P_k)^4\mu_k = o(1)
\end{aligned}$$

almost surely. This implies that $\max\limits_{1\le k\le K}\Pi_{2k} = o_p(1)$. By Assumption 4, it is seen that

$$\max_{1\le k\le K}\Pi_{3k} \le C\max_{1\le k\le K}\frac{\mu_k^T\mu_k}{N_k} = O(1)$$

almost surely. So we conclude that $\max\limits_{1\le k\le K}\widehat{\sigma}_k^2 = O_p(K^{1/2})$, and this implies that

$$\begin{aligned}
|\Lambda_5| &\le 2\alpha_N\max_{1\le k\le K}tr(P_k)\max_{1\le k\le K}\widehat{\sigma}_k^2K^{1/2}\|\boldsymbol{u}\| \\
&= O(\alpha_N K)\|\boldsymbol{u}\|. \qquad (A.30)
\end{aligned}$$

Combining (A.24), (A.25), (A.26), (A.29), (A.30) and Assumption 6, it is seen that Λ_2, Λ_3, Λ_4 and Λ_5 are asymptotically dominated by Λ_1. This completes the proof Theorem 4.

References

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

Chen, J., Li, D.G., Linton, O., Lu, Z.D.. Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. Journal of the American Statistical Association 2018;113(522):919–932.

Chen, L., Zhou, Y.. Quantile regression in big data: A divide and conquer based strategy. Computational Statistics & Data Analysis 2019;144:106892.

Chen, X., Xie, M.. A split-and-conquer approach for analysis of extraordinarily large data. Statistica Sinica 2014;24(4):1655–1684.

Craven, P., Wahba, G.. Smoothing noisy data with spline functions. Numerische Mathematik 1978;31(4):377–403.

Dempster, A.P., Schatzoff, M., Wermuth, N.. A simulation study of alternatives to ordinary least squares. Journal of the American Statistical Association 1977;72(357):77–91.

Deng, Z., Choi, K.S., Jiang, Y., Wang, S.. Generalized hidden-mapping ridge regression, knowledge-leveraged inductive transfer learning for neural networks, fuzzy systems and kernel methods. IEEE Transactions on Cybernetics 2014;44(12):2585–2599.

Dobriban, E., Sheng, Y.. Wonder: Weighted one-shot distributed ridge regression in high dimensions. Journal of Machine Learning Research 2020;21(66):1–52.

Fan, J., Guo, Y., Wang, K.. Communication-efficient accurate statistical estimation. Journal of the American Statistical Association 2021;0(0):1–11.

Fan, J., Peng, H.. On nonconcave penalized likelihood with diverging number of parameters. Annals of Statistics 2004;32(3):928–961.

Fan, J., Wang, D., Wang, K., Zhu, Z., et al. Distributed estimation of principal eigenspaces. Annals of Statistics 2019;47(6):3009–3031.

Fomby, T.B., Johnson, S.R., Hill, R.C.. Feasible generalized least squares estimation. In: Advanced econometric methods. Springer; 1984. p. 147–169.

Garber, D., Shamir, O., Srebro, N.. Communication-efficient algorithms for distributed stochastic principal component analysis. In: Proceedings of the 34th International Conference on Machine Learning. JMLR. org; 2017. p. 1203–1212.

Golub, G.H., Heath, M., Wahba, G.. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 1979;21(2):215–223.

Hansen, B.E.. Least squares model averaging. Econometrica 2007;75(4):1175–1189.

Hansen, B.E., Racine, J.S.. Jackknife model averaging. Journal of Econometrics 2012;167(1):38–46.

Hjort, N.L., Claeskens, G.. Frequentist model average estimators. Journal of the American Statistical Association 2003;98(464):879–899.

Hoerl, A.E., Kennard, R.W.. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970;12(1):55–67.

Jordan, M.I., Lee, J.D., Yang, Y.. Communication-efficient distributed statistical inference. Journal of the American Statistical Association 2019;114(526):668–681.

Kibria, B.G.. Performance of some new ridge regression estimators. Communications in Statistics-Simulation and Computation 2003;32(2):419–435.

Kibria, B.G.. Some liu and ridge-type estimators and their properties under the ill-conditioned gaussian linear regression model. Journal of Statistical Computation and Simulation 2012;82(1):1–17.

Lee, J.D., Liu, Q., Sun, Y., Taylor, J.E.. Communication-efficient sparse regression. Journal of Machine Learning Research 2017;18(1):115–144.

Lee, T.S.. Algorithm as 223: optimum ridge parameter selection. Journal of the Royal Statistical Society Series C (Applied Statistics) 1987;36(1):112–118.

Li, G.R., Peng, H., Zhu, L.. Nonconcave penalized M-estimation with a diverging number of parameters. Statistica Sinica 2011;21(1):391–419.

Li, K.C.. Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. The Annals of Statistics 1986;:1101–1112.

Li, K.C.. Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set. The Annals of Statistics 1987;15(3):958–975.

Li, X.M., Zou, G.H., Zhang, X.Y., Zhao, S.W.. Least squares model averaging based on generalized cross validation. Acta Mathematicae Applicatae Sinica, English Series 2021;37(3):495–509.

Lian, H., Fan, Z.. Divide-and-conquer for debiased l 1-norm support vector machine in ultra-high dimensions. Journal of Machine Learning Research 2017;18(1):6691–6716.

Lin, N., Xi, R.. Aggregated estimating equation estimation. Statistics and Its Interface 2011;4(1):73–83.

Liu, Q., Okui, R.. Heteroskedasticity-robust C_p model averaging. The Econometrics Journal 2013;16(3):463–472.

Mallows, C.L.. Some comments on C_p. Technometrics 1973;15(4):661–675.

Mirakyan, A., Meyer-Renschhausen, M., Koch, A.. Composite forecasting approach, application for next-day electricity price forecasting. Energy Economics 2017;66:228–237.

Schomaker, M.. Shrinkage averaging estimation. Statistical Papers 2012;53(4):1015–1034.

Shen, X., Alam, M., Fikse, F., Rönnegård, L.. A novel generalized ridge regression method for quantitative genetics. Genetics 2013;193(4):1255–1268.

Shi, C., Lu, W., Song, R.. A massive data framework for m-estimators with cubic-rate. Journal of the American Statistical Association 2018;113(524):1698–1709.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 2014;15(1):1929–1958.

Tang, L., Zhou, L., Song, P.X.K.. Distributed simultaneous inference in generalized linear models via confidence distribution. Journal of Multivariate Analysis 2020;176:104567.

Ullah, A., Wan, A., Wang, H., Zhang, X., Zou, G.. A semiparametric generalized ridge estimator and link with model averaging. Econometric Reviews 2017;36:370–384.

Volgushev, S., Chao, S.K., Cheng, G., et al. Distributed inference for quantile regression processes. Annals of Statistics 2019;47(3):1634–1662.

Wan, A.T., Zhang, X., Zou, G.. Least squares model averaging by Mallows criterion. Journal of Econometrics 2010;156(2):277–283.

Wang, C., Chen, M.H., Schifano, E., Wu, J., Yan, J.. Statistical methods and computing for big data. Statistics and Its Interface 2016;9(4):399–411.

Wang, C., Chen, M.H., Wu, J., Yan, J., Zhang, Y., Schifano, E.. Online updating method with new variables for big data streams. Canadian Journal of Statistics 2018;46(1):123–146.

Wang, H.X.J., Zhu, Z.Y., Zhou, J.H.. Quantile regression in partially linear varying coefficient models. Annals of Statistics 2009;37(6B):3841–3866.

Wang, J., Kolar, M., Srebro, N., Zhang, T.. Efficient distributed learning with sparsity. In: Proceedings of the 34th International Conference on Machine Learning. JMLR. org; 2017. p. 3636–3645.

Wang, L., Wu, Y.C., Li, R.Z.. Quantile regression for analyzing heterogeneity in ultra-high dimension. Journal of the American Statistical Association 2012;107(497):214–222.

Wang, X., Yang, Z., Chen, X., Liu, W.. Distributed inference for linear support vector machine. Journal of machine learning research 2019;20.

Whittle, P.. Bounds for the moments of linear and quadratic forms in independent variables. Theory of Probability and Its Applications 1960;5(3):302–305.

Xi, R., Nan, L., Chen, Y.. Compression and aggregation for logistic regression analysis in data cubes. IEEE Transactions on Knowledge and Data Engineering 2009;21(4):479–492.

Xiang, D., Wahba, G.. A generalized approximate cross validation for smoothing splines with non-gaussian data. Statistica Sinica 1996;6(3):675–692.

Xu, G., Shang, Z., Cheng, G.. Distributed generalized cross-validation for divide-and-conquer kernel ridge regression and its asymptotic optimality. Journal of Computational and Graphical Statistics 2019;28(4):891–908.

Xue, H., Zhu, Y., Chen, S.. Local ridge regression for face recognition. Neurocomputing 2009;72(4-6):1342–1346.

Yang, Y.. Adaptive regression by mixing. Journal of the American Statistical Association 2001;96:574–588.

Zhan, H., Xu, S.. Adaptive ridge regression for rare variant detection. PloS One 2012;7(8).

Zhang, T., Yang, B.. An exact approach to ridge regression for big data. Computational Statistics 2017;32(3):909–928.

Zhang, Y., Duchi, J., Wainwright, M.. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. Journal of Machine Learning Research 2015;16(1):3299–3340.

Zhang, Y., Duchi, J.C., Wainwright, M.J.. Communication-efficient algorithms for statistical optimization. Journal of Machine Learning Research 2013;14(1):3321–3363.

Zhao, S., Liao, J., Yu, D.. Model averaging estimator in ridge regression and its large sample properties. Statistical Papers 2020;61(4):1719–1739.

Zhao, T., Cheng, G., Liu, H.. A partially linear framework for massive heterogeneous data. The Annals of statistics 2016;44(4):1400–1437.