Journal of Machine Learning Research 24 (2023) 1-59 Submitted 5/22; Revised 4/23; Published 7/23

Least Squares Model Averaging for Distributed Data

Haili Zhang† ZHANGHL@SZPT.EDU.CN
Institute of Applied Mathematics
Shenzhen Polytechnic University
Shenzhen, 518055, China

Zhaobo Liu† LIUZHAOBO@SZU.EDU.CN
Institute for Advanced Study
Shenzhen University
Shenzhen, 518060, China

Guohua Zou* GHZOU@AMSS.AC.CN
School of Mathematical Sciences
Capital Normal University
Beijing, 100048, China

Editor: Sayan Mukherjee

Abstract

Divide and conquer algorithm is a common strategy applied in big data. Model averaging has the natural divide-and-conquer feature, but its theory has not been developed in big data scenarios. The goal of this paper is to fill this gap. We propose two divide-and-conquer-type model averaging estimators for linear models with distributed data. Under some regularity conditions, we show that the weights from Mallows model averaging criterion converge in L_2 to the theoretically optimal weights minimizing the risk of the model averaging estimator. We also give the bounds of the in-sample and out-of-sample mean squared errors and prove the asymptotic optimality for the proposed model averaging estimators. Our conclusions hold even when the dimensions and the number of candidate models are divergent. Simulation results and a real airline data analysis illustrate that the proposed model averaging methods perform better than the commonly used model selection and model averaging methods in distributed data cases. Our approaches contribute to model averaging theory in distributed data and parallel computations, and can be applied in big data analysis to save time and reduce the computational burden.

Keywords: consistency, distributed data, divide and conquer algorithm, Mallows' criterion, model averaging, optimality.

1. Introduction

Modern science and technology make data collection easier and easier, and thus more and more big data have been obtained and stored. Usually, such data are with complicated, structured, varied, and various characteristics in economy, finance, biology, medicine, industry, agriculture, transportation, and other fields. See, for example, Misra et al. (2019), who provided plenty of real data examples that reflect the overall outlook of big data era.

. † Haili Zhang and Zhaobo Liu contribute equally to this work.
. *Guohua Zou is corresponding author.

In this world of explosively large data, estimation faces big computational and statistical challenges, especially in scalability and storage bottlenecks of hardware and software issues, and invalidated exogeneous assumptions brought by incidental endogeneity in big data, seeing Fan et al. (2014) for a review. In big data applications, one often prefers to suggest a specific methodology for the problems he/she faces but without theoretical analysis. For example, Sienkiewicz et al. (2017) solved the computational problems on a single, multi-core server to describe spiking activity in non-linear dynamic systems with the software MapReduce and Hoodop, but no theoretical property is discussed. Hence the effective distributed estimation procedures with theoretical supports are urgently needed to deal with the computational challenges arisen from large sample size and large number of parameters in massive data analysis. In this regard, some distributed statistical computing methods have been proposed. See, for example, Varian (2014) and Wang et al. (2016).

The large-scale datasets may not fit the memory of a single computer and thus are distributedly stored in multiple machines or servers. So statistical methods should be adjusted and modified to accommodate distributed data. The divide and conquer trick is a practicable and common approach to handle the massive data computation with memory constraints. It divides data into several groups and then aggregates all group estimators by a simple average to lessen the computational burden (Zhang et al., 2013b; Chen and Xie, 2014; Zhang et al., 2015; Xu et al., 2019). A number of problems have been studied for the divide and conquer method, including variable selection (Chen and Xie, 2014), statistical optimization (Zhang et al., 2013b), logistic regression (Xi et al., 2009), estimation equation (Lin and Xi, 2011), kernel ridge regression (Zhang et al., 2015; Xu et al., 2019), quantile regression (Chen et al., 2019, 2020), linear support vector machine (Wang et al., 2019), and distributed principal component analysis (PCA) (Balcan et al., 2012; Garber et al., 2017). Some distributed statistical methods based on likelihood framework are also proposed, and the theoretical upper bound of the information loss for the distributed algorithm is obtained (c.f., Battey et al., 2018). For data distributed over the nodes, Safarinejadian et al. (2010) proposed a distributed expectation maximization (DEM) algorithm with two important advantages of scalability and fault tolerance for density estimation and clustering in sensor networks, which can also be seen as a divide and conquer method. The DEM algorithm is scalable and robust under the Gaussian mixture model assumption, where the addition of more nodes does not affect the performance of the DEM algorithm and it can still produce the right results even if failures of some nodes occur. The diffusion speed and convergence of the DEM algorithm have also been studied in Safarinejadian et al. (2010).

However, numerous papers on the divide and conquer algorithm are not involved with model selection uncertainty. Model averaging is a feasible method to avoid such an uncertainty. There are four main reasons prompting us to choose model averaging instead of model selection. First, choosing a single model may not take full information provided by the training data, especially when it is hard to get a best model. For example, there may be more than one candidate model with similar quantitative scores under some model selection criteria. On the other hand, different candidate models capture different data characteristics. In this dilemma, combining all of those models will not lose the information from each candidate model and thus may be a better choice. Simple averaging of different machine learning models to get a more accurate prediction has been a popular method in some big data applications. Model averaging can result in a smaller risk and get a more accurate prediction generally. In fact, model averaging often performs at least as well as the best algorithm in the candidate models. As commented by Schomaker and Heumann (2020), model averaging can improve the predictions and should be regarded as attractive complements for the machine learning and forecasting. Second, model averaging can be more stable. Based on different statistic analysis goals, model averaging can stabilize estimation and forecast by assigning different weights to candidate models, and is regarded as a smoothed extension of model selection. Third, model averaging can avoid selecting the worst candidate model. Last but not the least, model selection criteria based on likelihood, such as AIC (Akaike, 1974; Matsuda et al., 2021), BIC (Schwarz, 1978), and minimum description length (MDL, Maggioni and Murphy, 2019), can be invalid for some singular candidate models including artificial neural networks, normal mixtures, binomial mixtures, reduced rank regressions, Bayesian networks, and hidden Markov models, as the likelihood functions of these singular statistical models and learning machines cannot be approximated by any normal distribution (Watanabe, 2010, 2013). For so many singular models, model averaging, a valid solution, can be used to get more robust estimates and generalized machine learning methods. For all these reasons, compared with model selection, model averaging estimators often get higher prediction precision and better robustness, and thus have received extensive attention in recent years.

In the frequentist viewpoint, a key problem with the model averaging is the choice of weights assigned to different models. A variety of model averaging criteria have been suggested. See, for example, smoothed information criteria including smoothed AIC, smoothed BIC (Buckland et al., 1997), and smoothed FIC (Hjort and Claeskens, 2003; Claeskens and Carroll, 2007; Zhang and Liang, 2011; Zhang et al., 2012; Xu et al., 2014); adaptive method (Yang, 2001; Yuan and Yang, 2005); and asymptotically optimal methods, such as Mallows model averaging (MMA) method (Hansen, 2007; Wan et al., 2010), OPT method (Liang et al., 2011), jackknife model averaging (JMA) method (Hansen and Racine, 2012; Zhang et al., 2013a; Zhang and Zou, 2020), and leave-subject-out cross-validation method for time series data (Gao et al., 2016; Liao et al., 2019).

In this paper, we will focus on Hansen's MMA, which is the first model averaging method with optimality. Hansen (2007) proved that the Mallows criterion is asymptotically optimal in the sense of achieving the lowest possible squared error for the nested candidate models and discrete weight set. Further, Wan et al. (2010) provided an alternative proof for the non-nested candidate models and continuous model weights. Liu and Okui (2013) proposed a modified Mallows model averaging for heteroscedasticity data. Gao et al. (2019) suggested an adjusted MMA criterion for threshold auto-regressive model. Zhu et al. (2019) developed a Mallows-type model averaging estimator for the varying-coefficient partially linear model. A corrected Mallows model averaging method for small sample sizes can be found in Liao and Zou (2020).

In recent years, the property of model weight has attracted much attention. For model averaging, there are few articles on the uniqueness of the optimal weight choice except Hansen (2014) in which a unique empirical weight vector is obtained if the candidate models are appropriately restricted. Hansen (2014) investigated the asymptotic risks of nested least-squares averaging estimators with minimum mean squared error criterion in a local asymptotic framework and gave an explicit form of optimal weights based on asymptotic risk in some common situations. Hansen (2014) also suggested a practical rule that model averaging estimators should be based on models where the regressors have been grouped. This rule will lead to a better implementation of averaging. Charkhi et al. (2016) noticed the uniqueness of weights of model averaging based on likelihood frameworks and recommended a suitable class of models which are so-called singleton models where each model includes only one candidate variable. This singleton model trick can result in a drastic reduction in the computational cost of model averaging and can be applied in big data area. Another interesting problem with the model averaging is the consistency of weights. There are a few articles on this topic (c.f., Chen et al., 2018; Liao et al., 2019; Liao and Zou, 2020). Chen et al. (2018) proposed a semi-parametric penalized model averaging method for marginal regressions of time series and derived the consistency and oracle property with the assumption that the weights are sparse and some other regularity conditions. Each candidate model in Chen et al. (2018) can be regarded as a projection from response variable to marginal regressions, and the weights assigned to different models are without any constraints, as in Li et al. (2015), who proposed a forecasting method by combining all marginal regressions in applications. Liao et al. (2019) derived the convergence rate of the weights based on leave-subject-out cross-validation model averaging method for VAR model. Liao and Zou (2020) proved the consistency of MMA weights. Some articles also focus on the other statistical limiting properties of Mallows model averaging. For example, Liu (2015) derived the limiting distributions of the weights based on Mallows criterion and nested least squares averaging estimators under the local misspecification framework.

For distributed and massive data, except simple averaging and Fang's et al (2018) approximating calculations, no model averaging theory is developed. The purpose of this paper is to fill this gap. We will propose efficient computational strategies and theory for model averaging on distributed data and divergent dimensional regressions. The contributions of this article are threefold. First, we prove that the weight vector selected by Mallows model averaging criterion for least squares estimators in linear regression models is L_2 convergent to the theoretical optimal weight. Our results of convergence type are different from those in Hansen (2014), Liu (2015), Chen et al. (2018), Liao et al. (2019), and Zhang et al. (2020). Second, we propose two types of model averaging estimators for distributed or parallel data. From our theoretical analysis, we find that the two tricks of grouping regressors and singleton models can be used to reduce the computation cost. Before model averaging, using model selection can throw away some clearly unreasonable models and will relieve of the computational burden. Based on some suitable candidate models, we may be able to get a better model averaging estimator. The grouping regressors models and singleton models can be used as some alternative tricks to build the candidate model set. Such tricks have been used by, say, Hansen (2014) and Charkhi et al. (2016) in the literature on model averaging. In fact, the idea of grouping regressors has been investigated previously in statistical literature, including Efron and Morris (1973), Berger and Dey (1983), Dey and Berger (1983), George (1986a), George (1986b), and Mougeot et al. (2013). Grouping strategies have been shown to improve the prediction performance and interpretability of the candidate models (Lounici et al., 2011). In model averaging, each learner based on grouping some similar regressors will be more useful and all these learners can comprise a candidate model set which will lead to a drastic reduction in the computational cost. Both singleton models and averaging across singletons are also two popular methods in data analysis. For example, Hjort and Claeskens (2003) observed that the averaging across singletons method does quite well in achieving the smallest risk and leads to low standard deviation and short confidence intervals. In summary, both group strategies and singleton models can reduce the computational burden, make the optimal weights be identical and unique, and lead to a parsimonious model averaging estimator, and thus are useful tools in big data area. Third, it is inspired that we can use the weight calculated from a simple random sample without replacement with large size drawn from the extremely massive data as the weight estimation for the whole collected data.

The remainder of this paper is organized as follows. In Section 2, we build a general Mallows model averaging framework for distributed data. Then, we investigate the theoretical properties of the proposed weights and model averaging estimators in Section 3. Section 4 covers simulations. In Section 5, we apply our model averaging methods for distributed data to the real airline data. Theoretical proofs are included in Appendices.

2. Model Averaging Based on Distributed Data

2.1 Model averaging for subject

Let $\{(y_i, x_i) : i = 1, 2, \ldots, N\}$ be an i.i.d. sample from the following data-generating process,

$$y_i = \mu_i + e_i = \sum_{j=1}^{\infty} \theta_j x_{ij} + e_i, \quad i = 1, 2, \ldots, N,$$

where $y_i \in \mathbb{R}$, $x_i = (x_{i1}, x_{i2}, \ldots)^T$ is countably infinite, and e_i is an error term. We assume that $\{e_i\}_{i\geq 1}$ are mutually independent with $\mathbf{E}\left(e_i|x_i\right) = 0$ and $\mathbf{E}\left(e_i^2|x_i\right) = \sigma^2$, and $\mathbf{E}\mu_i^2 < \infty$. The model set-up follows Hansen (2007). We assume a sequence of linear approximating models, where the sth model uses the first p_s regressors of x_i, $s = 1, \ldots, S$. That is, the sth candidate model is

$$y_i = \sum_{j=1}^{p_s} \theta_j x_{ij} + e_i, \quad i = 1, 2, \ldots, N.$$

The approximating error of the sth candidate model is $b_{i(s)} = \sum_{j=p_s+1}^{\infty} \theta_j x_{ij}$. Let $\beta_{(s)} = (\theta_1, \theta_2, \ldots, \theta_{p_s})^T$, $s = 1, \ldots, S$.

Since N is extremely large, we apply the divide and conquer trick to treat the collected data. Without loss of generality, we let $N = Kn$, where both K and n are positive integers. Then we divide the collected data set $\{(y_i, x_i), i = 1, \ldots, N\}$ evenly and uniformly at random among a total of K subjects. At each subject, denote the resultant data as $\left\{Y_k, X_{(k)}\right\}, k = 1, 2, \ldots, K$, where $Y_k = (y_{k,1}, y_{k,2}, \cdots, y_{k,n})^T$ and $X_{(k)} = (x_{k,1}, x_{k,2}, \cdots, x_{k,n})^T$ with $(y_{k,j}, x_{k,j})$, $j = 1, 2, \cdots, n$, being a random sample from the $\{(y_i, x_i), i = 1, \ldots, N\}$. Denote $\boldsymbol{\mu}_k = (\mu_{k,1}, \mu_{k,2}, \cdots, \mu_{k,n})^T$ and the error term for the kth subject as $e_{(k)} = (e_{k,1}, e_{k,2}, \cdots, e_{k,n})^T$ accordingly. At subject k, we consider model averaging procedure.

The estimator of $\beta_{(s)}$ in the sth candidate model under the kth subject is given by

$$\hat{\beta}_{k,s} = (X_{k,s}^T X_{k,s})^{-1} X_{k,s}^T Y_k,$$

where $X_{k,s}$ is an $n \times p_s$ matrix with full column rank, including the first p_s columns of $X_{(k)}$ related to the sth candidate model, $s = 1, \ldots, S$. For simplicity, we denote $X_k = X_{k,S}$.

Then the model averaging estimator for $\boldsymbol{\mu}_k$ has the form

$$\hat{\boldsymbol{\mu}}_k(W_k) = \sum_{s=1}^{S} w_{k,s} X_{k,s} \hat{\beta}_{k,s}$$

with $W_k = (w_{k,1}, \ldots, w_{k,S})^T \in Q$ and

$$Q \triangleq \left\{ w = (w_1, \ldots, w_S)^T : \sum_{s=1}^{S} w_s = 1, w_s \geq 0, s = 1, 2, \ldots, S \right\}.$$

A key problem with the estimator $\hat{\boldsymbol{\mu}}_k(W_k)$ is the choice of weights. To choose a proper W_k, we minimize the following Mallows criterion

$$C_{k,n}(W_k) = \frac{1}{n} \left(Y_k - \hat{\boldsymbol{\mu}}_k(W_k)\right)^T \left(Y_k - \hat{\boldsymbol{\mu}}_k(W_k)\right) + \frac{2}{n}\sigma^2 tr\left[P_k(W_k)\right] \tag{1}$$

in Q to get

$$\hat{W}_k = (\hat{w}_{k,1}, \hat{w}_{k,2}, \cdots, \hat{w}_{k,S})^T = \underset{w \in Q}{\operatorname{argmin}}\, C_{k,n}(w), \tag{2}$$

where

$$P_k(W_k) \triangleq \sum_{s=1}^{S} w_{k,s} X_{k,s} \left(X_{k,s}^T X_{k,s}\right)^{-1} X_{k,s}^T \triangleq \sum_{s=1}^{S} w_{k,s} P_{k,s},$$

and $tr\left[P_k(W_k)\right] = \sum_{s=1}^{S} w_{k,s} p_s$. When σ^2 is unknown, (1) needs to be computed with a sample estimate. There are several ways to estimate σ^2. We use the following estimator

$$\hat{\sigma}_k^2 = \frac{\left(Y_k - X_{k,S}\hat{\beta}_{k,S}\right)^T \left(Y_k - X_{k,S}\hat{\beta}_{k,S}\right)}{n - p_S},$$

which is based on the largest candidate model (Hansen, 2007; Wan et al., 2010) for the kth subject. The resultant Mallows model averaging estimator for $\boldsymbol{\mu}_k$ is given by

$$\hat{\boldsymbol{\mu}}_k(\hat{W}_k) = \sum_{s=1}^{S} \hat{w}_{k,s} X_{k,s} \hat{\beta}_{k,s}.$$

2.2 Model averaging for distributed data

Let Π_s be a selection matrix for the sth candidate model, so that $X_{k,s} = X_k \Pi_s^T$ and $\Pi_s \Pi_s^T = I_{p_s}$, where I_{p_s} is an identity matrix of order p_s. The model averaging estimator of $\beta_{(S)}$ at subject k is

$$\hat{\beta}_k(\hat{W}_k) = \sum_{s=1}^{S} \hat{w}_{k,s} \Pi_s^T \hat{\beta}_{k,s}$$

for $k = 1, \ldots, K$. In the following, we construct two types of model averaging estimators.

(1^0) Simple aggregation of model averaging estimators

We aggregate the K local estimators together by simple averaging to obtain the simple aggregated model averaging estimator of $\beta_{(S)}$, that is,

$$\overline{\beta} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k(\hat{W}_k). \tag{3}$$

Accordingly, the simple aggregated model averaging estimator of $\boldsymbol{\mu}_k$ is given by

$$\overline{\boldsymbol{\mu}}_k = X_k\overline{\beta}.$$

(2^0) Doubly simple aggregation of model averaging estimators

The other aggregated model averaging procedure is as follows. First, we aggregate the least squares estimators $\hat{\beta}_{k,s}$ and the weights $\hat{w}_{k,s}$ respectively, that is,

$$\tilde{\beta}_s = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_{k,s}, \quad s = 1, \ldots, S,$$

and

$$\overline{w}_s = \frac{1}{K}\sum_{k=1}^{K}\hat{w}_{k,s}, \quad s = 1, \ldots, S. \tag{4}$$

Second, we aggregate $\tilde{\beta}_s$ and $\overline{w}_s$ of each candidate model to obtain the doubly simple aggregated model averaging estimator of $\beta_{(S)}$, that is

$$\overline{\tilde{\beta}} = \sum_{s=1}^{S}\overline{w}_s\Pi_s^T\tilde{\beta}_s. \tag{5}$$

The doubly simple aggregated model averaging estimator of $\boldsymbol{\mu}_k$ is

$$\overline{\tilde{\boldsymbol{\mu}}}_k = X_k\overline{\tilde{\beta}}.$$

3. Theoretical Results

We first introduce some notations. We use ℓ_2 to denote the usual Euclidean norm $\|\theta\| = \sqrt{\sum_{j=1}^{d}\theta_j^2}$ with $\theta = (\theta_1, \theta_2, \cdots, \theta_d)^T$. The ℓ_2-operator norm of a matrix $A \in \mathbb{R}^{d_1\times d_2}$ is its maximum singular value, defined by

$$\|A\|_2 \triangleq \sup_{v\in\mathbb{R}^{d_2}, \|v\|\le 1}\|Av\|.$$

Let λ_1, λ_2, $\cdots$, and λ_d be the real eigenvalues of a matrix $A \in \mathbb{R}^{d\times d}$. In particular, we denote its minimum and maximum eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. Then its spectral radius $\rho_r(A)$ is defined as

$$\rho_r(A) \triangleq \max_{1\le i\le d}|\lambda_i|.$$

A convex function F is λ-strongly convex on a set $U \subseteq \mathbb{R}^d$ if for arbitrary $u \in U$ and $v \in U$, we have

$$F(u) \geq F(v) + \langle \nabla F(v), u - v \rangle + \frac{\lambda}{2}\|u - v\|_2^2,$$

where ∇F is the derivative of the function F. In addition, if F is not differentiable, we may replace ∇F by any subgradient of F.

Consider the quadratic loss function

$$\begin{aligned} L_N(w) &= \frac{1}{N}\left\|\hat{\boldsymbol{\mu}}\left(w\right) - \boldsymbol{\mu}\right\|^2 \\ &= \frac{1}{nK}\sum_{k=1}^{K}\|\hat{\boldsymbol{\mu}}_k(w) - \boldsymbol{\mu}_k\|^2, \end{aligned}$$

where $w = (w_1, w_2, \cdots, w_S)^T \in Q$, and $\hat{\boldsymbol{\mu}}\left(w\right) = \left(\hat{\boldsymbol{\mu}}_1^T\left(w\right), \hat{\boldsymbol{\mu}}_2^T\left(w\right), \cdots, \hat{\boldsymbol{\mu}}_K^T\left(w\right)\right)^T$ and $\boldsymbol{\mu} = \left(\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, \cdots, \boldsymbol{\mu}_K^T\right)^T$ are two $N \times 1$ vectors. The population risk R_N is given by

$$\begin{aligned} R_N(w) &= \mathbf{E}L_N(w) \\ &= \frac{1}{n}\mathbf{E}\left\|\hat{\boldsymbol{\mu}}_1\left(w\right) - \boldsymbol{\mu}_1\right\|^2, \end{aligned}$$

where the second equality is by the assumption that the data are independent and identically distributed. For the weight $w \in Q$, we can rewrite $R_N(w)$ as

$$R_0\left(w_1, w_2, \cdots, w_{S-1}\right) = \frac{1}{n}\mathbf{E}\left\|\sum_{s=1}^{S-1} w_s X_{1,s}\hat{\beta}_{1,s} + \left(1 - \sum_{s=1}^{S-1} w_s\right) X_{1,S}\hat{\beta}_{1,S} - \boldsymbol{\mu}_1\right\|^2 \tag{6}$$

with the constraint of $(w_1, w_2, \cdots, w_{S-1})^T \in Q_0$ and

$$Q_0 \triangleq \left\{(w_1, w_2, \cdots, w_{S-1})^T \,\middle|\, w_s \geq 0, s = 1, 2, \cdots, S-1; 0 \leq \sum_{s=1}^{S-1} w_s \leq 1\right\}.$$

Denote $w_0 = (w_1, w_2, \cdots, w_{S-1})^T$. For the model averaging framework, we need to determine the weights assigned to candidate models. So, our goal is to estimate the parameter vector minimizing the risk $R_0\left(w_0\right)$, namely the quantity

$$w_0^* \triangleq \underset{w_0 \in Q_0}{\operatorname{argmin}}\, R_0\left(w_0\right),$$

which is equivalent to estimating

$$w^* \triangleq \underset{w \in Q}{\operatorname{argmin}}\, R_N(w).$$

By first calculating the weight W_k at the kth subject by (2), and then averaging the weights by (4) to get the averaged weight $\overline{w}$, where $\overline{w} = (\overline{w}_1, \overline{w}_2, \ldots, \overline{w}_S)^T$, we can show the consistency of the weight $\overline{w}_0 = (\overline{w}_1, \overline{w}_2, \ldots, \overline{w}_{S-1})^T$ to w_0^*. We will establish the theoretical properties of $\overline{w}_0$ in Subsection 3.1 and the proposed estimators (3) and (5) in Subsection 3.2.

3.1 Convergence of the weight estimator

In the following, we assume the candidate models $s = 1, 2, \ldots, S$ are nested, and then $0 < p_1 < p_2 < \cdots < p_S$. Without loss of generality, we assume that $\mathbf{E}x_{k,i,j} = 0$, and the covariance of $x_{k,i,j}$ and x_{k,i,j_1} is σ_{j,j_1} with $j \neq j_1$. Denote $\Sigma_s = \{\sigma_{j,j_1}\}_{1\leq j,j_1 \leq p_s}$ and let the pseudo-true value of $\beta_{(s)}$ be

$$\begin{aligned}
\beta_{\star,s} &\triangleq \arg\min_{\beta_{k,s}\in\mathbb{R}^{p_s}} \frac{1}{n}\mathbf{E}\left\|X_{k,s}\beta_{k,s} - \boldsymbol{\mu}_k\right\|^2 \\
&= \left\{\mathbf{E}\left(X_{k,s}^T X_{k,s}\right)\right\}^{-1}\mathbf{E}\left(X_{k,s}^T\boldsymbol{\mu}_k\right) \\
&= (\theta_1, \theta_2, \ldots, \theta_{p_s})^T + \Sigma_s^{-1}\left(\sum_{j=p_s+1}^{\infty}\theta_j\sigma_{1,j}, \sum_{j=p_s+1}^{\infty}\theta_j\sigma_{2,j}, \ldots, \sum_{j=p_s+1}^{\infty}\theta_j\sigma_{p_s,j}\right)^T \\
&\triangleq \beta_{(s)} + \Sigma_s^{-1}\gamma_s.
\end{aligned}$$

Further, define

$$R_N^*(w) = \frac{1}{N}\sum_{k=1}^{K}\mathbf{E}\left\|X_k\beta_\star(w) - \boldsymbol{\mu}_k\right\|^2,$$

where

$$\beta_\star(w) \triangleq \sum_{s=1}^{S} w_s \Pi_s^T \beta_{\star,s}.$$

Accordingly,

$$R_0^*(w_0) = \frac{1}{N}\sum_{k=1}^{K}\mathbf{E}\left\|X_k\left\{\sum_{s=1}^{S-1} w_s\Pi_s^T\beta_{\star,s} + \left(1 - \sum_{s=1}^{S-1} w_s\right)\Pi_S^T\beta_{\star,S}\right\} - \boldsymbol{\mu}_k\right\|^2.$$

We now define the error of the pseudo-true model as

$$\begin{aligned}
\delta_{k,s} &= \boldsymbol{\mu}_k - X_{k,s}\beta_{\star,s} \\
&= \left(\boldsymbol{\mu}_k - X_{k,s}\beta_{(s)}\right) - X_{k,s}\Sigma_s^{-1}\gamma_s \\
&\triangleq \boldsymbol{b}_{k,s} - X_{k,s}\Sigma_s^{-1}\gamma_s,
\end{aligned}$$

and

$$\Sigma_{\infty|s} \triangleq \mathbf{E}\left(\delta_{k,s}\delta_{k,s}^T|X_{k,s}\right), \quad \Sigma'_{\infty|s} \triangleq \frac{1}{n}\mathbf{E}(\delta_{k,s}\delta_{k,s}^T\delta_{k,s}\delta_{k,s}^T|X_{k,s}).$$

To derive the consistency of our weight estimator, we need the following regularity conditions.

Condition 1 $w_0^* \in intQ_0$.

Condition 2 $\max_{1\leq s\leq S}\mathbf{E}\left(|x_{(i)}^T\Pi_s^T\beta_{(s)}|^{\eta+2} + |x_{(i)}^T\Pi_s^T\beta_{\star,s}|^{\eta+2}\right) < C_b < \infty$ for some $\eta \geq 2$, where $x_{(i)} = (x_{i,1}, x_{i,2}, \ldots, x_{i,p_S})^T$, and $\mathbf{E}e_{k,i}^4 \leq \omega < \infty$ for $k = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n$.

Condition 3 There is a constant σ_n^2 bounded away from zero such that $\lambda_{\max}\left(\Sigma_{\infty|s}\right) + \lambda_{\max}\left(\Sigma'_{\infty|s}\right) \leq \sigma_n^2$ for $s = 1, \ldots, S$.

Condition 4 $\frac{S^2 p_S \sigma_n^2}{n \overline{\lambda}_S^2} = o(1)$, where $\overline{\lambda}_S = \lambda_{\min}\left[\nabla^2 R_0^*\left(w_0^*\right)\right]$.

Remark 1 *Condition 1 is common in optimization theory to ensure the solution can be calculated by some gradient descent algorithms or iterative algorithms. Since $R_0\left(w_0\right)$ is twice differentiable with respect to w_0, and Condition 1 requires that $R_0\left(w_0\right)$ have a local minimum at the interior point w_0^* of Q_0, which means that $R_N\left(w\right)$ has a local minimum at the interior point w^* of the simplex Q, we have*

$$\lambda_n \triangleq \lambda_{\min}[\nabla^2 R_0\left(w_0^*\right)] > 0.$$

Condition 1 may hold when all the candidate models are useful or competitive. This condition is a valuable alternative to Definition 2 of Watanabe (2010), by which Bayesian learning theory can be investigated directly.

Condition 2 places some bounds on the moments of error term $e_{k,i}$, candidate models and pseudo-true candidate models. When $x_{k,i,j}$ are independent and Gaussian, with the assumption $\mathbf{E}\left|\mu_i\right|^{\eta+2} < \infty$, Condition 2 is easily satisfied even for S tending to ∞.

Condition 3 gives an upper bound for the maximum eigenvalues of $\Sigma_{\infty|s}$ and $\Sigma'_{\infty|s}$ that depends on n (here we omit a set with zero probability). Noting that

$$\Sigma_{\infty|s} = \left\{\mathbf{E}\left[\left(\mu_{k,i} - x_{(k,i,s)}^T \beta_{\star,s}\right) \cdot \left(\mu_{k,j} - x_{(k,j,s)}^T \beta_{\star,s}\right) \middle| x_{(k,i,s)}, x_{(k,j,s)}\right]\right\}_{1 \leq i,j \leq n},$$

where $x_{(k,i,s)}$ is the transpose of the ith row of the matrix $X_{k,s}$, it is not difficult to show that

$$\begin{aligned}\lambda_{\max}\left(\Sigma_{\infty|s}\right) \leq & \max_{1 \leq j \leq n} \left\{\mathbf{E}\left[\left(\mu_{k,j} - x_{(k,j,s)}^T \beta_{\star,s}\right)^2 \middle| x_{(k,j,s)}\right]\right. \\ & \left. + \sum_{i \neq j} \left|\mathbf{E}\left[\left(\mu_{k,i} - x_{(k,i,s)}^T \beta_{\star,s}\right) \cdot \left(\mu_{k,j} - x_{(k,j,s)}^T \beta_{\star,s}\right) \middle| x_{(k,i,s)}, x_{(k,j,s)}\right]\right|\right\}.\end{aligned}$$

Let us consider a special scenario where $\mu_{k,i} - x_{(k,i,s)}^T \beta_{\star,s}$ are mutually independent random variables conditionally given $X_{k,s}$. Then it follows that

$$\begin{aligned}& \mathbf{E}\left[\left(\mu_{k,i} - x_{(k,i,s)}^T \beta_{\star,s}\right) \cdot \left(\mu_{k,j} - x_{(k,j,s)}^T \beta_{\star,s}\right) \middle| x_{(k,i,s)}, x_{(k,j,s)}\right] \\ = & \mathbf{E}\left[\left(\mu_{k,i} - x_{(k,i,s)}^T \beta_{\star,s}\right) \middle| x_{(k,i,s)}\right] \mathbf{E}\left[\left(\mu_{k,j} - x_{(k,j,s)}^T \beta_{\star,s}\right) \middle| x_{(k,j,s)}\right].\end{aligned}$$

Clearly, as $n \to \infty$, $\mathbf{E}\left[\mu_{k,i} - x_{(k,i,s)}^T \beta_{\star,s} \middle| x_{(k,i,s)}\right]$ plays a decisive role in the size of $\lambda_{\max}\left(\Sigma_{\infty|s}\right)$. Similarly, let $q_{il} = \left|\mathbf{E}\left[\left(\mu_{k,i} - x_{(k,i,s)}^T \beta_{\star,s}\right)^l \middle| x_{(k,i,s)}\right]\right|$, $l = 1, 2, 3, 4$, then

$$\lambda_{\max}\left(\Sigma'_{\infty|s}\right) \leq \frac{1}{n} \max_{1 \leq j \leq n} \left(q_{j4} + \sum_{i \neq j} q_{i2} q_{j2} + \sum_{i \neq j} (q_{i3} q_{j1} + q_{j3} q_{i1}) + \sum_{h \neq i,j} q_{h2} q_{i1} q_{j1}\right).$$

As $n^2 - n$ of the n^2 terms on the right-hand side of the above inequality contain q_{i1}, $i = 1, \dots, n$, $\lambda_{\max}\left(\Sigma'_{\infty|s}\right)$ is also influenced by $\mathbf{E}\left[\mu_{k,i} - x^T_{(k,i,s)}\beta_{\star,s}\middle| x_{(k,i,s)}\right]$, $i = 1, \dots, n$. Observing that $\mathbf{E}\left[\mu_{k,i}\middle| x_{(k,i,s)}\right]$ is the optimal estimator of $\mu_{k,i}$ in L_2 sense, $\mathbf{E}\left[\mu_{k,i} - x^T_{(k,i,s)}\beta_{\star,s}\middle| x_{(k,i,s)}\right]$ represents the gap between the optimal L_2 estimator and the linear minimum variance estimator based on the sth candidate model. Specially, when $x_{k,i,j}$ is jointly Gaussian, it follows that $\mathbf{E}\left[\mu_{k,i} - x^T_{(k,i,s)}\beta_{\star,s}\middle| x_{(k,i,s)}\right] = 0$ and

$$\Sigma_{\infty|s} = \left\{ \sum_{j_1,j_2=p_s+1}^{\infty} \theta_{j_1}\theta_{j_2}\sigma_{j_1,j_2} - \gamma_s^T \Sigma_s^{-1}\gamma_s \right\} I_n = \sigma^2_{\infty|s} I_n,$$

hence $\Sigma'_{\infty|s} = \frac{n+2}{n}\sigma^4_{\infty|s} I_n$. Thus, $\sigma_n^2 = \max_{1\le s\le S}(\sigma^2_{\infty|s} + 3\sigma^4_{\infty|s})$ satisfies Condition 3.

Condition 4 allows $\overline{\lambda}_S$ to tend to zero at a rate slower than $\sqrt{S^2 p_S \sigma_n^2 n^{-1}}$ with the dimension of regressor vector and/or the number of candidate models being divergent when n tends to ∞. Further, with the assumption that the data are independent and identically distributed, after some calculations, it can be seen that,

$$\begin{aligned}
\nabla^2 R_0^*(w_0^*) &= 2\mathbf{E}\left[\left\{\left(\Pi_{s_1}^T\beta_{\star,s_1} - \Pi_S^T\beta_{\star,S}\right)^T x_{(i)}x_{(i)}^T\left(\Pi_{s_2}^T\beta_{\star,s_2} - \Pi_S^T\beta_{\star,S}\right)\right\}_{1\le s_1,s_2\le S-1}\right] \\
&= 2\left\{\left(\Pi_{s_1}^T\beta_{\star,s_1} - \Pi_S^T\beta_{\star,S}\right)^T \Sigma_S \left(\Pi_{s_2}^T\beta_{\star,s_2} - \Pi_S^T\beta_{\star,S}\right)\right\}_{1\le s_1,s_2\le S-1},
\end{aligned}$$

which is similar to Condition A6 of Chen et al. (2018). Like Chen et al. (2018), if we do not take account of the constraint $\sum_{s=1}^S w_s = 1$, then

$$\nabla^2 R_N^*(w^*) = 2\mathbf{E}\left[\left\{\left(\Pi_{s_1}^T\beta_{\star,s_1}\right)^T x_{(i)}x_{(i)}^T\left(\Pi_{s_2}^T\beta_{\star,s_2}\right)\right\}_{1\le s_1,s_2\le S}\right].$$

In this case, Condition 4 only requires that

$$\frac{S\sqrt{p_S}\sigma_n}{\sqrt{n}\lambda_{\min}\left[\nabla^2 R_N^*(w^*)\right]} = o(1),$$

which is weaker than Condition A6 of Chen et al. (2018) when $S^2 p_S = o(n)$.

Now, denoting $\hat{W}_{k,0} = (\hat{w}_{k,1}, \hat{w}_{k,2}, \cdots, \hat{w}_{k,S-1})^T$ and then $\overline{w}_0 = \frac{1}{K}\sum_{k=1}^K \hat{W}_{k,0}$, we have the following theoretical results.

Theorem 1 *Under Conditions 1-4, we have*

$$\mathbf{E}\left\|\overline{w}_0 - w_0^*\right\|^2 = O\left(\frac{Sp_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2}\right) + O\left(\frac{S^3 p_S^2(S+\sigma_n^2)}{n^2\overline{\lambda}_S^4}\right),$$

and so

$$\mathbf{E}\left\|\overline{w} - w^*\right\|^2 = O\left(\frac{S^2 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2}\right) + O\left(\frac{S^4 p_S^2(S+\sigma_n^2)}{n^2\overline{\lambda}_S^4}\right).$$

Proof See Appendix B. ■

Corollary 1 *Under Conditions 1,2 and 4, if the covariates $x_{k,i,j}$ are jointly Gaussian, and S and $p_s, s = 1, 2, \cdots, S$ are fixed, then*

$$\mathbf{E}\left\|\overline{w} - w^*\right\|^2 = O\left(\frac{1}{Kn}\right) + O\left(\frac{1}{n^2}\right).$$

3.2 Mean squared errors of model averaging estimators for regression coefficients

In this subsection, we first show some limiting results about $\min_{w \in Q} R_N(w)$ and the proposed two model averaging estimators of $\beta_{(S)}$, and then provide the upper bounds of the mean squared errors of Mallows model averaging estimators.

Condition 5 $\frac{p_S \sigma_n^2}{n \xi_{\star,N}} = o(1)$, where $\xi_{\star,N} = \inf_{w \in Q} R_N^*(w)$.

Remark 2 *Condition 5 requires that the rate of $n\xi_{\star,N}$ tending to ∞ should be faster than that of $p_S\sigma_n^2$, which is similar to Condition (C4) of Zhang et al. (2020). If $x_{k,i,j}$ is jointly Gaussian, then $\sigma_n^2 = \max\limits_{1 \le s \le S}(\sigma_{\infty|s}^2 + 3\sigma_{\infty|s}^4)$, and in this case, this condition is easily satisfied.*

Theorem 2 *Under Conditions 1-5, we have*

$$\sup_{w \in Q}\left|\frac{R_N(w)}{R_N^*(w)} - 1\right| = o(1), \tag{7}$$

and

$$\frac{R_N(w^*)}{\xi_{\star,N}} = 1 + o(1). \tag{8}$$

Proof See Appendix B. ■

Remark 3 *From Theorem 2, $\xi_{\star,N}$ can be seen as the limit of $R_N(w^*)$, the optimal risk of Mallows model averaging estimator. Condition 5 and (8) show that the rate of $NR_N(w^*)$ tending to ∞ should be faster than that of $Kp_S\sigma_n^2$. This property is also consistent with the requirement that the true model should not be in the candidate model set, which is a condition commonly arisen in optimal model averaging. When the true model is an infinite dimensional model, $NR_N(w^*)/Kp_S\sigma_n^2 \to \infty$ is an alternative to Assumption 2 of Zhang (2021).*

In the following, we derive the differences between (3), (5) and $\beta_\star(w^*)$, respectively. Define

$$m_S = \max_{1 \le s \le S}\left\|\mathbf{E}\left[\Sigma_S^{1/2}\left\{\hat{\beta}_{k,s} - \beta_{\star,s}\right\}\right]\right\|.$$

Condition 6 *tr* $\left(\mathbf{E}\left[\left(X_{k,s}^T X_{k,s}\right)^{-1}\right]\Sigma_s\right) = O\left(\frac{p_s}{n}\right)$, $1 \leq s \leq S$.

Remark 4 *This condition places restriction on the upper bound of* $tr(\mathbf{E}[(X_{k,s}^T X_{k,s})^{-1}]\Sigma_s)$*. The upper bound nearly matches the risk for Gaussian design. The sufficient conditions for Condition 6 are given in Theorem 3 of Mourtada (2022). From Mourtada (2022), it can be seen that our Condition 6 is mild. When* $x_{(k,i,s)}$ *is Gaussian with the covariance matrix is* Σ_s*, it is easy to verify that*

$$\mathbf{E}\left[\left(X_{k,s}^T X_{k,s}\right)^{-1}\right]\Sigma_s = (n - p_s - 1)^{-1} I_{p_s},$$

and so

$$tr\left(\mathbf{E}\left[\left(X_{k,s}^T X_{k,s}\right)^{-1}\right]\Sigma_s\right) = \frac{p_s}{n - p_s - 1}.$$

Theorem 3 *Under Conditions 1-4 and 6, we have*

$$\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\overline{\beta} - \beta_\star\left(w^*\right)\right\}\right\|^2 = O\left(\frac{p_S\sigma_n^2}{n}\right) + O\left(\frac{S^3 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2}\right) + O\left(\frac{S^5 p_S^2\left(S+\sigma_n^2\right)}{n^2\overline{\lambda}_S^4}\right),$$

and

$$\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\overline{\overline{\beta}} - \beta_\star\left(w^*\right)\right\}\right\|^2 = O\left(m_S^2\right) + O\left(\frac{S^3 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2}\right) + O\left(\frac{S^5 p_S^2(S+\sigma_n^2)}{n^2\overline{\lambda}_S^4}\right).$$

Proof See Appendix B. ■

Remark 5 *When* $x_{k,j,i}$ *is jointly Gaussian, the ordinary least squares estimator* $\hat{\beta}_{k,s}$ *is an unbiased estimator of pseudo-true parameter* $\beta_{\star,s}$*, i.e.,* $m_S = 0$*. So Theorem 3 means that when* $\overline{\lambda}_S$ *has a uniform lower bound away from zero, if* $K = O(1)$*, then* (3) *and* (5) *have the same convergence rates to* $\beta_\star(w^*)$*; if* K *tends to* ∞*, then* (5) *has a faster convergence rate to* $\beta_\star(w^*)$ *than* (3)*.*

3.3 Mean squared errors of model averaging estimators for conditional mean

We now consider the mean squared errors of model averaging estimators for estimating conditional mean.

(1^0) Out-of-sample mean squared errors

Let (y_v, x_v) be an independent copy of (y_i, x_i), where $x_v = (x_{v1}, x_{v2}, \ldots)$ is countably infinite, $x_{v,s} = (x_{v1}, x_{v2}, \ldots, x_{vp_s})^T$, $\mu_v = \sum_{j=1}^{\infty}\theta_j x_{vj}$. The simple aggregated model averaging estimator of μ_v is

$$\overline{\mu}_v = x_{v,S}^T\overline{\beta}.$$

The doubly simple aggregated model averaging estimator for μ_v is

$$\overline{\overline{\mu}}_v = x_{v,S}^T\overline{\overline{\beta}}.$$

Define the out-of-sample mean squared errors for $\overline{\mu}_v$ and $\overline{\overline{\mu}}_v$ as $\mathbf{E}\left(\overline{\mu}_v - \mu_v\right)^2$ and $\mathbf{E}\left(\overline{\overline{\mu}}_v - \mu_v\right)^2$, respectively, for which we give bounds in the following theorem.

Theorem 4 *Under Conditions 1-6, we obtain*

$$\frac{\mathbf{E}\left(\overline{\mu}_v - \mu_v\right)^2}{\xi_{\star,N}} = 1 + O\left(\sqrt{\frac{p_S\sigma_n^2}{n\xi_{\star,N}} + \frac{S^3 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2\xi_{\star,N}} + \frac{S^5 p_S^2(S+\sigma_n^2)}{n^2\overline{\lambda}_S^4\xi_{\star,N}}}\right),$$

and

$$\frac{\mathbf{E}\left(\overline{\widetilde{\mu}}_v - \mu_v\right)^2}{\xi_{\star,N}} = 1 + O\left(\sqrt{\frac{m_S^2}{\xi_{\star,N}} + \frac{S^3 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2\xi_{\star,N}} + \frac{S^5 p_S^2(S+\sigma_n^2)}{n^2\overline{\lambda}_S^4\xi_{\star,N}}}\right).$$

Proof See Appendix B. ■

Remark 6 *Theorems 4 suggests the following points:*

1. *Noting that by Lemma 6 in Appendix A*

$$m_S^2 \leq \max_{1\leq s\leq S} \mathbf{E}\left\|\Sigma_S^{1/2}\left\{\hat{\beta}_{k,s} - \beta_{\star,s}\right\}\right\|^2 = O\left(\frac{p_S\sigma_n^2}{n}\right),$$

   the doubly simple aggregation may be a better choice than simple aggregation since it has a smaller out-of-sample mean squared errors bound. Specifically, when the least squares is close to the unbiased estimator of the pseudo-true value (for example, when $x_{k,i,j}$ is jointly Gaussian, all $m_s^2 = 0$), the advantages of doubly simple aggregation will be more prominent.

2. *The results can be used to determine the optimal number of subjects for the fixed total number of observations $N = nK$. To obtain a simple solution, we treat the term $\frac{S^3 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2\xi_{\star,N}}$ as fixed, and then let the remaining two terms be equal to this term, respectively, to reach the minimum upper bounds of MSEs. Thus, we suggest that the optimal choice of the number of subjects K with the two methods satisfy*

$$K^* \asymp \begin{cases} \left(\frac{\sigma_n^2\overline{\lambda}_S^2}{(S+\sigma_n^2)S^3} + \sqrt{\frac{S^2 p_S}{N\overline{\lambda}_S^2}}\right)^{-1}, & \textit{for simple aggregation,} \\ \left(\frac{m_S^2 n\overline{\lambda}_S^2}{(S+\sigma_n^2)S^3 p_S} + \sqrt{\frac{S^2 p_S}{N\overline{\lambda}_S^2}}\right)^{-1}, & \textit{for doubly simple aggregation.} \end{cases}$$

   In fact, as long as σ_n^2 is monotonically increasing and σ_n^2/n is monotonically decreasing with respect to n (by Remark 1, this condition is easily satisfied), the above K^* minimizes the upper bounds of MSEs for the fixed N with S and $p_1, p_2, \ldots, p_S$ depending only on N. Here we use symbol $a_n \asymp b_n$, which means both $a_n = O(b_n)$ and $b_n = O(a_n)$. The optimality of K implies that choosing any $K = O(K^*)$ cannot reduce the upper bound of out-of-sample mean squared errors (instead, n will increase and so more storage space and computational resources are needed at each subject), while choosing any K with K/K^* tending to ∞ will increase the upper bound of out-of-sample mean squared errors. If $x_{k,i,j}$ is jointly Gaussian, it follows that the optimal *choice of K with the doubly simple aggregation method satisfies $K^* \asymp \left(\frac{N\overline{\lambda}_S^2}{S^2 p_S}\right)^{1/2}$. In such a setting, the boundedness of σ_n^2 can be obtained, so that the optimal K^* with the simple aggregation method satisfies*

$$K^* \asymp \min\left\{\frac{S^4}{\overline{\lambda}_S^2}, \left(\frac{N\overline{\lambda}_S^2}{S^2 p_S}\right)^{1/2}\right\},$$

as $n \to \infty$.

3. *In practical prediction, it is difficult to determine the value of $\overline{\lambda}_S$. To facilitate the selection of K, we can assume that $\overline{\lambda}_S \asymp 1$ holds and $x_{k,i,j}$ is jointly Gaussian. In this case, the optimal $K^* \asymp \min\left\{S^4, \left(\frac{N}{S^2 p_S}\right)^{1/2}\right\}$ and $K^* \asymp \left(\frac{N}{S^2 p_S}\right)^{1/2}$ for the proposed methods, respectively. Assumption $\overline{\lambda}_S \asymp 1$ is not restricted, and it is consistent with Condition A6 of Chen et al. (2018). Essentially, $\overline{\lambda}_S \asymp 1$ represents the eigenvalues of a positive definite information matrix based on S pseudo-true models to be away from 0.*

(2^0) In-sample mean squared errors

The in-sample mean squared errors for simple aggregated model averaging estimator and doubly simple aggregated model averaging estimator are defined as

$$\overline{MSE} = \frac{1}{N}\sum_{k=1}^{K} \mathbf{E}\left\|\overline{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\right\|^2,$$

and

$$\widetilde{\overline{MSE}} = \frac{1}{N}\sum_{k=1}^{K} \mathbf{E}\left\|\tilde{\overline{\boldsymbol{\mu}}}_k - \boldsymbol{\mu}_k\right\|^2,$$

respectively.

Theorem 5 *Under Conditions 1-6, we obtain*

$$\frac{\overline{MSE}}{\xi_{\star,N}} = 1 + O\left(\xi_{\star,N}^{-\frac{1}{2}} \cdot \left\{\frac{S^3 p_S(S+\sigma_n^2)}{n\overline{\lambda}_S^2} + \frac{S^{\frac{\eta+4}{\eta+2}}}{K}\left(\frac{S^2 p_S\left(S+\sigma_n^2\right)}{n\overline{\lambda}_S^2}\right)^{\frac{\eta}{\eta+2}}\right\}^{\frac{1}{2}}\right), \tag{9}$$

and

$$\frac{\widetilde{\overline{MSE}}}{\xi_{\star,N}} = 1 + O\left(\xi_{\star,N}^{-\frac{1}{2}} \cdot \left\{\overline{m}_S^2 + S^{\frac{\eta+4}{\eta+2}}\left(\frac{S^2 p_S\left(S+\sigma_n^2\right)}{Kn\overline{\lambda}_S^2} + \frac{S^4 p_S^2\left(S+\sigma_n^2\right)}{n^2\overline{\lambda}_S^4}\right)^{\frac{\eta}{\eta+2}}\right\}^{\frac{1}{2}}\right), \tag{10}$$

where

$$\overline{m}_S = \max_{1\le s\le S}\left\|\mathbf{E}\left[x_{k,1}\Pi_s^T\left\{\hat{\beta}_{k,s} - \beta_{\star,s}\right\}\right]\right\|.$$

Proof See Appendix B. ■

Remark 7 *Theorems 5 suggests the following points:*

1. *By Lemma 5 in Appendix A, we see that* $\overline{m}_s^2 = O\left(p_S \sigma_n^2 / n\right)$*, so by some simple calculations and Condition 4, for any* K *with*

$$\left(\frac{n\overline{\lambda}_S^2}{Sp_S(S+\sigma_n^2)}\right)^{\frac{2}{\eta}} / K = o(1), \tag{11}$$

the doubly simple aggregation is a more appropriate choice when we focus on the in-sample mean squared errors. When η *is large,* (11) *is easy to satisfy.*

2. *Similar to Remark 6, for the total number of observations* $N = nK$*, the optimal selection of* K *is given by*

$$K^* \asymp \begin{cases} \left(\frac{N\overline{\lambda}_S^2}{Sp_S(S+\sigma_n^2)}\right)^{\frac{2}{\eta+4}}, & \text{for simple aggregation,} \\ \min\left\{\sqrt{\frac{N\overline{\lambda}_S^2}{S^2 p_S}}, \frac{N\overline{\lambda}_S^2 \overline{m}_S^{\frac{\eta+2}{\eta}}}{\sqrt{S^{5+4/\eta}(S+\sigma_n^2)p_S}}\right\}, & \text{for doubly simple aggregation.} \end{cases}$$

When $x_{k,i,j}$ *is jointly Gaussian,* $\overline{m}_S = 0$*, and* η *can be sufficiently large. So it follows that the optimal values of* K^* *for simple aggregation and doubly simple aggregation satisfy*

$$K^* \asymp \left(\frac{N\overline{\lambda}_S^2}{S^2 p_S}\right)^{2/(\eta+4)} \quad \text{and} \quad K^* \asymp \left(\frac{N\overline{\lambda}_S^2}{S^2 p_S}\right)^{1/2},$$

respectively. This shows that for simple aggregation, as N *tends to infinity, the optimal* K *is always smaller than that for doubly simple aggregation.*

3.4 Asymptotic optimality

This subsection focuses on the optimality of the proposed methods in the asymptotic sense. In the distributed data framework, the definition of asymptotic optimality of model averaging estimator differs a little from the traditional definition. Since the least squares $\hat{\beta}_{k,s}$ at each subject uses only n observations, the average loss

$$R_N(w^*) = \frac{1}{n}\mathbf{E}\left\|\sum_{s=1}^{S} w_s^* X_{1,s}\hat{\beta}_{1,s} - \boldsymbol{\mu}_1\right\|^2$$

cannot represent the risk of the optimal model averaging estimator using the full N observations. To address this issue, we note that as long as $n \to \infty$, the least squares estimator based on either n observations or N observations converges to the pseudo-true value, so $R_N^*(w^*)$ defined in Subsection 3.1 can be used to represent the minimum risk of model averaging estimator in the distributed data case. Thus, we define that a model averaging estimator has asymptotic optimality if its mean squared error (MSE) satisfies $MSE/R_N^*(w^*) \to 1$.

Theorem 6 below reveals that under the framework of distributed data, our proposed model averaging methods are asymptotically optimal for both out-of-sample and in-sample estimations.

Theorem 6 *Under Conditions 1-6, and* $\frac{S^3 p_S (S+\sigma_n^2)}{\xi_{\star,N} \underline{\lambda}_S^2 n} = o(1)$*, we have*

(i) for out-of-sample mean squared errors,

$$\frac{\mathbf{E}\left(\overline{\mu}_v - \mu_v\right)^2}{R_N^*(w^*)} = 1 + o(1) \quad \textit{and} \quad \frac{\mathbf{E}\left(\widetilde{\overline{\mu}}_v - \mu_v\right)^2}{R_N^*(w^*)} = 1 + o(1);$$

(ii) for in-sample mean squared errors,

$$\frac{\overline{MSE}}{R_N^*(w^*)} = 1 + o(1) \quad \textit{and} \quad \frac{\widetilde{MSE}}{R_N^*(w^*)} = 1 + o(1). \tag{12}$$

Specifically, when $N \to \infty$ *and* $K = 1$*,* (12) *degenerates to a typical form*

$$\frac{R_N(\hat{w})}{\inf_{w \in Q} R_N(w)} = 1 + o(1). \tag{13}$$

Proof This theorem is a direct corollary of Theorems 2, 4 and 5. ■

Remark 8 *Theorem 6 shows that our simple aggregated model averaging estimators and doubly simple aggregated model averaging estimators achieve the optimality in out-of-sample and in-sample mean squared errors.*

Unlike the existing literature on the asymptotic optimality based on the loss function, (13) *indicates that the Mallows' model averaging method has an asymptotic optimality in the sense of minimizing risk. This is an interesting finding, which also shows that* $R_N(\hat{w}) \le C \inf_{w \in Q} R_N(w)$*, where* C *is a constant that depends only on* $\theta_j, j = 1, 2, \ldots, \infty$.

3.5 Minimaxity of model averaging estimators

In this subsection, we turn to deriving the minimax optimal convergence rate of proposed estimators. For simple and doubly simple aggregation model averaging estimators, denote

$$\overline{Mse}\left(W_1, W_2, \ldots, W_K\right) = \frac{1}{N} \sum_{k=1}^{K} \mathbf{E} \left\| X_k \frac{1}{K} \sum_{l=1}^{K} \hat{\beta}_l(W_l) - \boldsymbol{\mu}_k \right\|^2,$$

$$\widetilde{Mse}(W_1, W_2, \ldots, W_K) = \frac{1}{N} \sum_{k=1}^{K} \mathbf{E} \left\| X_k \sum_{s=1}^{S} \left(\frac{1}{K} \sum_{l=1}^{K} w_{l,s} \right) \Pi_s^T \left(\frac{1}{K} \sum_{j=1}^{K} \hat{\beta}_{j,s} \right) - \boldsymbol{\mu}_k \right\|^2,$$

and then $\overline{MSE} = \overline{Mse}(\hat{W}_1, \hat{W}_2, \ldots, \hat{W}_K)$ and $\widetilde{MSE} = \widetilde{Mse}(\hat{W}_1, \hat{W}_2, \ldots, \hat{W}_K)$. We assume that the candidate models are fixed, i.e., $p_1, \ldots, p_S$ and S are fixed integers.

Now, for any $q \geq 4$, we consider the true parameter $\theta = (\theta_1, \theta_2, \ldots)^T$ in the Banach space

$$\ell_q = \left\{ \theta : \sum_{j=1}^{\infty} |\theta_j|^q < \infty \right\}$$

with the norm

$$\|\theta\| = \left(\sum_{j=1}^{\infty} |\theta_j|^q \right)^{\frac{1}{q}}.$$

We construct

$$\Theta = \Theta(\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3) = S_1 \cap S_2 \cap S_3 \tag{14}$$

with

$$S_1 = \left\{ \theta \in \ell_q : w_s^{**} \in [\epsilon_0, 1 - \epsilon_1], s = 1, \ldots, S-1, (w_1^{**}, \ldots, w_{S-1}^{**})^T = \operatorname*{argmin}_{w \in Q_0} R_N^*(w) \right\},$$

$$S_2 = \left\{ \theta \in \ell_q : \lambda_{\min} \left[\nabla^2 R_0^* \left(w_0^* \right) \right] \geq \epsilon_2, \inf_{w \in Q} R_N^*(w) \geq \epsilon_2 \right\},$$

and

$$S_3 = \{ \theta \in \ell_q : \|\theta\| \leq \epsilon_3 \},$$

where $\epsilon_0, \epsilon_1 \in (0, 1)$ with $\epsilon_0 < 1 - \epsilon_1$, and $\epsilon_2, \epsilon_3 > 0$ are constants.

Theorem 7 *Assume Conditions 3 and 6 hold, and $\sigma_n^2 = o(n)$. If $\sup_{j \geq 1} \mathbf{E}|x_{k,i,j}|^q < \infty$, and $\mathbf{E}e_{k,i}^4 \leq \omega < \infty$ for $k = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n$, then we have*

$$\sup_{\theta \in \Theta} \overline{MSE} = \left\{ 1 + O\left(\frac{\sigma_n^2}{n} + \frac{1}{K} \left(\frac{\sigma_n^2}{n} \right)^{\frac{q-2}{q}} \right) \right\} \inf_{w \in Q} \sup_{\theta \in \Theta} R_N^*(w) \tag{15}$$

and

$$\sup_{\theta \in \Theta} \widetilde{MSE} = \left\{ 1 + O\left(\frac{\sigma_n^2}{n} + \left(\frac{\sigma_n^2}{N} + \frac{\sigma_n^2}{n^2} \right)^{\frac{q-2}{q}} \right) \right\} \inf_{w \in Q} \sup_{\theta \in \Theta} R_N^*(w), \tag{16}$$

where Θ is defined by (14)*. Moreover,*

$$\inf_{W_1 \in Q, W_2 \in Q, \ldots, W_K \in Q} \sup_{\theta \in \Theta} \overline{Mse}\left(W_1, W_2, \ldots, W_K \right) = \left\{ 1 + O\left(\frac{\sigma_n^2}{n} \right) \right\} \inf_{w \in Q} \sup_{\theta \in \Theta} R_N^*(w) \tag{17}$$

and

$$\inf_{W_1 \in Q, W_2 \in Q, \ldots, W_K \in Q} \sup_{\theta \in \Theta} \widetilde{Mse}\left(W_1, W_2, \ldots, W_K \right) = \left\{ 1 + O\left(\frac{\sigma_n^2}{n} \right) \right\} \inf_{w \in Q} \sup_{\theta \in \Theta} R_N^*(w). \tag{18}$$

Proof See Appendix B. ■

Remark 9 (15) *and* (16) *of Theorem 7 imply that the simple and doubly simple aggregation model averaging estimators proposed in this paper are both asymptotically minimax for the parameter set* Θ*. Further,* (17) *and* (18) *illustrate that in the large sample sense, the maximum risks of our proposed model averaging estimators cannot be improved.*

4. Simulation

In this section, we conduct simulation experiments to compare the finite sample performance of our distributed model averaging methods and some commonly used model selection and model averaging methods. In detail, we compare three simple aggregated model selection estimators: (i) AIC model selection (AIC), (ii) BIC model selection (BIC), (iii) Mallows' model selection (Mallows C_p); three simple aggregated model averaging estimators: (iv) simple aggregated smoothed AIC estimator (SAIC), (v) simple aggregated smoothed BIC estimator (SBIC), (vi) simple aggregated Mallows' model averaging estimator (MMA); and three doubly simple aggregated model averaging estimators: (vii) doubly simple aggregated smoothed AIC estimator (dSAIC), (viii) doubly simple aggregated smoothed BIC estimator (dSBIC), (ix) doubly simple aggregated Mallows' model averaging estimator (dMMA). Thus, we compare totally nine estimators.

4.1 Simulation setup

We report the simulation studies of the infinite order regression first. The data generating process is exactly the same as that in Hansen (2007):

$$y_i = \sum_{j=1}^{\infty} \theta_j x_{j,i} + e_i,$$

where $x_{1,i} = 1$, and $x_{j,i}(j = 1, 2, \ldots)$ and error e_i are independent and identically distributed as $N(0, 1)$. We set $\theta_j = c\sqrt{2\alpha} j^{-\alpha-1/2}$, and consider the parameter α varied at $0.5, 1.0$ and 1.5. As in Hansen (2007), the parameter c is selected such that $R^2 = c^2/(1 + c^2)$ changes from 0.1 to 0.9.

4.2 Results on in-sample risk

In this subsection, we compare the in-sample risks of the above nine distributed estimators. For the distributed data, we set the sample size for each subject to be varied at $n = 50, 150, 400, 1000, 5000$ and 10000. The number of subjects is set as $K = 1, 2, 3, 5$ and 10. Let p_S equal to $\left[4n^{1/2}\right]_+$ ($[\cdot]_+$ means round to get an integer, and so $p_S = 28, 49, 80, 126, 283$ and 400 for the above six sample sizes), and the number of candidate models S be $\lceil n^{1/3} \rceil + 1(\lceil \cdot \rceil$ means round up to get an integer, and so $S = 5, 7, 9, 11, 19$ and 22 for the six sample sizes). All the candidate models are nested and the dimension for the sth candidate model is $1 + d \times (s - 1)$, where $d = \lceil (p_S - 1)/(S - 1) \rceil$ and $s = 1, 2, \ldots, S - 1$, while the dimension for the Sth candidate model is p_S.

To evaluate different estimators, similar to Hansen (2007), we normalize the risk based on average across 5000 simulation draws by dividing by the risk of the best-fitting simple aggregated estimator $\tilde{\beta}_s$ (i.e., (2.2)). For the simulated in-sample risk, we define it as

$$\frac{1}{D}\sum_{r=1}^{D}\sum_{k=1}^{K}\left\|\hat{\boldsymbol{\mu}}_{k,(j)}^{(r)}-\boldsymbol{\mu}_{k}^{(r)}\right\|^{2},$$

where r means the rth simulation replication, $D = 5000$, and j means the jth method considered in our simulation.

For $j =$i, ii, and iii, $\hat{\boldsymbol{\mu}}_{k,(j)}^{(r)}$ is determined by AIC, BIC, and Mallows' model selection methods, respectively, i.e.,

$$\hat{\boldsymbol{\mu}}_{k,(j)}^{(r)} = X_k\left\{\frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k(W_k)\right\},$$

where $W_k = (w_{k,1}, w_{k,2}, \ldots, w_{k,S}) \in \{0,1\}^S$ with $\sum_{s=1}^{S} w_{k,s} = 1$. The AIC for the sth model at the kth subject is given by

$$\mathrm{AIC}_k^{(s)} = n_k \log\left(\hat{\sigma}_{k,(s)}^2\right) + 2p_s$$

with

$$\hat{\sigma}_{k,(s)}^2 = \left\|Y_k - X_{k,s}\hat{\beta}_{k,s}\right\|^2 / n_k,$$

and the model selected by AIC is

$$\hat{W}_k^{AIC} = \arg\min_{\substack{W_k\in\{0,1\}^S\\ \sum_{s=1}^{S} w_{k,s}=1}} \sum_{s=1}^{S} w_{k,s}\mathrm{AIC}_k^{(s)}.$$

Similarly, the BIC for the sth model at the kth subject is

$$\mathrm{BIC}_k^{(s)} = n_k \log\left(\hat{\sigma}_{k,(s)}^2\right) + \log(n_k)p_s,$$

and the model selected by BIC is

$$\hat{W}_k^{BIC} = \arg\min_{\substack{W_k\in\{0,1\}^S\\ \sum_{s=1}^{S} w_{k,s}=1}} \sum_{s=1}^{S} w_{k,s}\mathrm{BIC}_k^{(s)}.$$

Furthermore, the Mallows' C_p of the sth model at the kth subject is

$$\textsc{Mallows}_k^{(s)} = \left\|Y_k - X_{k,s}\hat{\beta}_{k,s}\right\|^2 + 2\tilde{\sigma}^2 p_s,$$

where

$$\tilde{\sigma}^2 = (n - p_S)^{-1}\left\|Y_k - X_{k,S}\hat{\beta}_{k,S}\right\|^2.$$

The model selected by Mallows' C_p is

$$\hat{W}_k^{Mallows} = \arg \min_{\substack{W_k \in \{0,1\}^S \\ \sum_{s=1}^S w_{k,s}=1}} \sum_{s=1}^{S} w_{k,s} \text{Mallows}_k^{(s)}.$$

For j =iv, v, and vi, $\hat{\boldsymbol{\mu}}_{k,(j)}^{(r)}$ is determined by three simple aggregated model averaging estimators SAIC, SBIC, and MMA, respectively, i.e.,

$$\hat{\boldsymbol{\mu}}_{k,(j)}^{(r)} = X_k \left\{ \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_k(W_k) \right\},$$

where $W_k \in Q$ is calculated by

$$\left(\frac{\exp(-\text{AIC}_k^{(1)}/2)}{\sum_{s=1}^S \exp(-\text{AIC}_k^{(s)}/2)}, \ldots, \frac{\exp(-\text{AIC}_k^{(S)}/2)}{\sum_{s=1}^S \exp(-\text{AIC}_k^{(s)}/2)} \right),$$

$$\left(\frac{\exp(-\text{BIC}_k^{(1)}/2)}{\sum_{s=1}^S \exp(-\text{BIC}_k^{(s)}/2)}, \ldots, \frac{\exp(-\text{BIC}_k^{(S)}/2)}{\sum_{s=1}^S \exp(-\text{BIC}_k^{(s)}/2)} \right),$$

and (2) with σ^2 being replaced by $\tilde{\sigma}^2$, respectively.

As for j =vii, viii, and ix, $\hat{\boldsymbol{\mu}}_{k,(j)}^{(r)}$ is generated by three doubly simple aggregated model averaging estimators dSAIC, dSBIC, and dMMA, respectively, i.e.,

$$\hat{\boldsymbol{\mu}}_{k,(j)}^{(r)} = X_k \left\{ \sum_{s=1}^{S} w_s \Pi_s^T \tilde{\beta}_s \right\},$$

where w_s is calculated by

$$\frac{1}{K} \sum_{k=1}^{K} \frac{\exp(-\text{AIC}_k^{(s)}/2)}{\sum_{s=1}^S \exp(-\text{AIC}_k^{(s)}/2)},$$

$$\frac{1}{K} \sum_{k=1}^{K} \frac{\exp(-\text{BIC}_k^{(s)}/2)}{\sum_{s=1}^S \exp(-\text{BIC}_k^{(s)}/2)},$$

and (4), respectively.

The simulation results for $K = 1, 2$ and 3 are similar. When $K = 1$, the three doubly simple aggregated model averaging estimators are equal to the three simple aggregated model averaging estimators respectively, and all the model selection and averaging estimators perform closely to those in Hansen (2007) where all data are at the same subject. The performance of $K = 10$ is similar to that of $K = 5$. So to save space, we present only the results on $K = 2$ and 5, which are summarized in Figures 1-6. The risks of estimators under other settings are available from the authors upon request.

We reveal some interesting commonalities from Figures 1-6 as follows:

1) Model averaging methods are frequently better than their model selection counterparts, e.g., dMMA and MMA get smaller risks than Mallows, and dSAIC (dSBIC) and SAIC (SBIC) behave better than AIC (BIC), especially when $n = 50$ and 150. In model selection methods, Mallows performs the best in most of cases. The difference between AIC and Mallows gets small when n increases for all of figures. It is also observed that the difference of all methods becomes small as α varies from 0.5 to 1.5.

2) For model averaging methods, doubly simple aggregated model averaging estimators achieve lower risks than their corresponding simple aggregated model averaging estimators in most of cases. MMA often performs the best among the simple aggregated model averaging estimators. The difference between SAIC and MMA decreases with n tending to 10000.

3) It can be seen that our dMMA gets the smallest risks in most of cases, especially in the case of $n = 50$, followed by MMA and dSAIC. An exception is when $\alpha = 0.5$, $K = 5$ and $n = 50$. In this situation, dSAIC is better than MMA and dMMA. MMA and dMMA always perform the best when the sample sizes are 5000 and 10000. With n tending to 10000, the difference between dMMA and MMA gets smaller and smaller, which is consistent with Theorem 5 that shows dMMA and MMA may not have big difference in the sense of in-sample risk. In addition, the behavior of dMMA and dSAIC becomes similar as n increases to 10000.

4) Observing the effect of n, we can see that when n is small ($n = 50$ and 150), MMA and dMMA perform the best in most of cases. When n becomes large ($n = 400, 1000$ and 5000), Mallows type methods and AIC type methods behave similarly. When $n = 10000$, the risks of all approaches are close.

5) BIC type methods (e.g., BIC, SBIC, and dSBIC) always fluctuate a lot as R^2 goes from 0.1 to 0.9, and often behave well when $n = 50$ and $R^2 = 0.1$. The risks of AIC type methods (e.g., AIC, SAIC, and dSAIC) and Mallows type methods (e.g., Mallows, MMA, and dMMA) regularly decrease slowly as R^2 increases, and Mallows type methods frequently have the smallest risks. These indicate that the Mallows type methods are the most favored methods in most of cases.

6) As for the number of subjects K, comparing figures with $K = 2$ and $K = 5$, for small n, like $n = 50$, the improvements of dMMA over MMA when $K = 2$ are larger than those when $K = 5$; for big n, like $n > 50$, the improvements of dMMA over MMA become smaller and smaller as K increases from 2 to 5. This means that in the cases of smaller n and smaller K, dMMA has greater advantages. For bigger n and bigger K, MMA is more applicable since MMA requires less computation and has similar performance to dMMA. This phenomenon is consistent with the results by Theorem 5 in the case of $\overline{m}_S = 0$ and $\eta = \infty$.

In summary, for in-sample estimation, our simulation results show that doubly simple aggregated model averaging methods are better than their simple aggregated counterparts and model selection methods. Further, dMMA performs the best in most of cases, followed by MMA or dSAIC.

4.3 Results on out-of-sample risk

For the simulated out-of-sample risk, we define it as

$$\frac{1}{D}\sum_{r=1}^{D}\sum_{i=1}^{n}\left(\hat{\mu}_{i,o,(j)}^{(r)}-\mu_{i,o}^{(r)}\right)^2,$$

where the definitions of D, r, and j are the same as before. We compare the normalized out-of-sample risks of the above nine distributed estimators. The sample size for every subject is set as $n = 50, 150, 400$ and 1000. The number of subjects is set as $K = 1, 2, 3, 5$ and 10. We let p_S equal to $\left[9n^{1/3}\right]_+$, and S be $\lceil (p-1)/5\rceil + 1$. All the candidate models are nested and the dimension for the sth candidate model is $5\times(s-1)+1$, $s = 1, 2, \ldots, M-1$, while the dimension for the Sth candidate model is p_S.

To save space, we still present only the results on $K = 2$ and 5, which are summarized in Figures 7-12. Other results are available from the authors. Some common phenomena, which are a frequent occurrence in Figures 7-12, are listed below:

1) It is clear that model averaging methods are better than their model selection counterparts in the sense of minimizing the out-of-sample risks. For example, SAIC is better than AIC, SBIC is better than BIC, and MMA is better than Mallows' C_p, especially for the cases where $n = 50$ and 150.

2) Comparing all model averaging methods, doubly simple aggregated model averaging estimators outdo their corresponding simple aggregated model averaging estimators in the most of scenarios, particularly when $n = 50$ and 150. For example, dSAIC is superior to SAIC, and dMMA is superior to MMA. BIC type methods are still not robust for different R^2. With $K = 5$, AIC and Mallows C_p type methods behave closely to each other when n increases from 150 to 400 and then 1000.

3) It is observed that dMMA often behaves the best in getting the smallest risks, followed by MMA and dSAIC. In particular, dMMA surpasses MMA more clearly when $K = 5$ than when $K = 2$ for $n = 50$. This phenomenon accords closely with Theorem 4. In addition, the difference between dMMA and MMA becomes small when n varies from 50 to $150, 400$, and 1000. This is expected because from Theorem 4, the difference between dMMA and MMA becomes smaller and smaller with n increasing.

4) Varying R from 0.1 to 0.9 causes significant variations for BIC type methods in a large number of simulation settings, except for the case of $n = 50$ and $K = 2$, where BIC type methods often behave well. BIC type methods are quite poor relative to the other methods when n increases from 150 to 400 and 1000, as shown in Figures 7-12. These indicate that the BIC type methods are not robust. AIC type methods and Mallows type methods gradually reduce the out-of sample risks as R tends to 0.9. Mallows type methods are the most stable methods in our simulations. Thus, dMMA and MMA are also efficient and stable in achieving minimum out-of-sample risks. On the other hand, dSAIC is frequently superior to dMMA in getting minimum risks when $K = 5$ and $n = 150$.

5) Comparing risks with the same n when $K = 2$ and $K = 5$, the difference between dMMA and dSAIC when $K = 2$ is smaller than that when $K = 5$, particularly for the case where $n = 50$. As for the effect of the number of subjects K, small K is perferred for dMMA and MMA with $n = 150, 400$ and 1000, but when $n = 50$, dMMA and MMA with big K have significant advantages over other methods.

In summary, for out-of-sample estimation, our methods dMMA and MMA perform the best in most of cases. Furthermore, Mallows and AIC type methods often perform equally well.

5. Real Data Analysis

In this section, we use our proposed distributed model averaging methods to analyze the airline on-time performance data from the 2009 ASA Data Expo (http://stat-computing.org/dataexpo/2009/the-data.html). The data set is publicly available and has been used for demonstration with big data in many papers. For instance, it was used as a case study to demonstrate a logistic model fitting with a massive dataset that exceeds the RAM of a single computer by Wang et al. (2016). This data set is collected from October 1987 to April 2008 for all commercial flights within the USA. It consists of 12 million flights with 29 variables. The big memory project (http://www.jstatsoft.org/index.php/jss/article /downloadSuppFile/v055i14/Airline.tar.bz2) presents a compressed version of the pre-processed data set, which is approximately 1.7 GB, and will take 12 GB when uncompressed.

The response variable of the regression is late time (in hours). We consider linear models, and the covariates include three continuous variables: departure delay time (DepDelay, in hours), scheduled elapsed time (CRSElapsedTime, in hours), and distance from origin to destination (Distance, in 1000 miles); and five dummy variables: Weekend, departure hour (Dephour), origin (Origin), and destination (Dest). Since we consider a series of linear candidate models, we first rank the continuous variables by absolute marginal correlation coefficients to the response variable. The top three variables are DepDelay, CRSElapsedTime, and Distance. We then consider two sets of models: (i) three models that range from the model with intercept and DepDelay to the model that includes all top three continuous variables, and (ii) three nested models that incorporate dummy variables such as Weekend, Dephour, and Oringe and Dest. Weekend and Dephour capture the impact of official and business activities, and Dephour also captures the effects of weather on flight delays, while Oringe and Dest capture the impact of different routes. Additionally, the regressor sets for the six nested candidate models are presented in Table 1.

Due to computer memory limitation, we sort the data by the date of boarding time on schedule and divide the whole data with the sample size of 123,534,969 into 124 subjects, where the first 123 subjects each contain $N = 1,000,000$ records covering a week's flight data, and the last one contains 534,969 records. We use the ith subject data as training data to predict the late time at the $(i + 1)$th subject data, $i = 1, 2, \ldots, 123$. For the ith subject, we apply simple random sampling scheme without replacement to the data and get K random samples, then we use our proposed distributed model averaging methods for data analysis. K is set to be $1, 2, 5, 10, 100, 200, 500$ and 1000. We compare the mean squared prediction errors (MSPEs) of the nine methods given in Section 4 for the $(i + 1)$ subject data. We conduct 123 rounds. Since when $K = 1$, SAIC, SBIC, and MMA are the

Table 1: Regressor sets for the six models used in Airline Data.

Model	Regressor Set
1	Intercept + DepDelay
2	Intercept + DepDelay + CRSElapsedTime
3	Intercept + DepDelay + CRSElapsedTime + Distance
4	Intercept + DepDelay + CRSElapsedTime + Distance+ Weekend
5	Intercept + DepDelay + CRSElapsedTime + Distance+ Weekend + DepHour
6	Intercept + DepDelay + CRSElapsedTime + Distance+ Weekend + DepHour + Oringe + Dest

same as dSAIC, dSBIC and dMMA, respectively, we omit the results for the doubly simple aggregated model averaging methods in the case. To save space, we present only the results on the mean, median and optimal rate of 123 rounds MSPEs for each method in Table 2, and Diebold and Mariano test (Diebold and Mariano, 2002) results for the differences of MSPEs in Tables 3 and 4. The results on the other estimators such as those of weights and coefficients of candidate models are available from the authors upon request.

From Table 2, we observe that MMA and dMMA always achieve the lowest MSPEs, followed by AIC or SAIC. dMMA has a significant advantage when $K = 200, 500$ and 1000. Basically, the MSPEs of all methods decrease as K increases from 1 to 100 and increase as K increases from 100 to 1000. When $K = 100$, MMA obtains the smallest MSPEs, followed by dMMA. In optimal rate, dMMA is superior to the rest methods in obtaining the highest optimal rates

From Diebold and Mariano test results in Tables 3 and 4, MMA and dMMA are statistically significantly superior to other methods, and the difference between MMA and dMMA is not significant.

In conclusion, MMA and dMMA are effective methods to reduce risks in prediction for big data analysis.

6. Concluding Remarks

In this paper, we proposed two aggregated model averaging estimators for distributed data and proved that the weights based on Mallows model averaging criterion are L_2 convergent to the theoretically optimal weights. The bounds of mean squared errors and the asymptotic optimality for the proposed model averaging estimators are also established. These are the first theoretical results of applying model averaging method to big data analysis with divide and conquer trick. Simulations and real data analysis show that simple aggregation and doubly simple aggregation methods for model averaging estimators are better than their model selection counterparts in situations where there are massive distributed or parallel data, and especially when K is large, dMMA has more advantages in getting the smallest mean squared errors. In practice, how to balance K and n is an unavoidable problem for big data computing. In our opinion, dMMA is more preferred in the cases of smaller n and bigger K where dMMA has more reduction in variances of both the coefficient estimators in

Table 2: MSPEs of different methods for Airline Data.

K		AIC	BIC	Mallows	SAIC	SBIC	MMA	dSAIC	dSBIC	dMMA
1	Mean ($\times 10^{-2}$)	6.181	6.181	6.181	6.181	6.181	**6.174**			
	Median ($\times 10^{-2}$)	5.833	5.833	5.833	5.833	5.833	**5.811**			
	Optimal rate	0.051	0.047	0.051	0.088	0.088	**0.674**			
2	Mean ($\times 10^{-2}$)	6.179	6.179	6.179	6.179	6.179	**6.173**	6.179	6.179	**6.173**
	Median ($\times 10^{-2}$)	5.833	5.833	5.833	5.833	5.833	**5.811**	5.833	5.833	**5.811**
	Optimal rate	0.054	0.050	0.054	0.053	0.045	**0.285**	0.046	0.046	**0.366**
5	Mean ($\times 10^{-2}$)	6.177	6.177	6.177	6.177	6.177	**6.172**	6.177	6.177	**6.172**
	Median ($\times 10^{-2}$)	5.835	5.835	5.835	5.835	5.835	**5.803**	5.835	5.835	**5.803**
	Optimal rate	0.052	0.056	0.052	0.062	0.067	**0.252**	0.040	0.038	**0.382**
10	Mean ($\times 10^{-2}$)	6.174	6.174	6.174	6.174	6.174	**6.170**	6.175	6.175	**6.170**
	Median ($\times 10^{-2}$)	5.839	5.839	5.839	5.839	5.839	**5.807**	5.839	5.839	**5.806**
	Optimal rate	0.061	0.061	0.061	0.047	0.069	**0.260**	0.026	0.050	**0.366**
100	Mean ($\times 10^{-2}$)	6.162	6.164	6.162	6.162	6.164	**6.159**	6.162	6.163	6.160
	Median ($\times 10^{-2}$)	5.843	5.843	5.843	5.843	5.843	**5.840**	5.843	5.843	**5.839**
	Optimal rate	0.088	0.027	0.080	0.033	0.057	**0.244**	0.073	0.073	**0.325**
200	Mean ($\times 10^{-2}$)	6.172	6.176	6.172	6.172	6.175	**6.169**	**6.169**	6.174	**6.166**
	Median ($\times 10^{-2}$)	5.843	5.843	5.843	5.843	5.842	**5.838**	5.843	5.841	**5.838**
	Optimal rate	0.130	0.033	0.057	0.024	0.024	**0.203**	0.089	0.081	**0.358**
500	Mean ($\times 10^{-2}$)	6.214	6.221	6.214	6.215	6.221	**6.209**	6.211	6.223	**6.206**
	Median ($\times 10^{-2}$)	**5.760**	5.774	**5.760**	5.761	5.774	5.762	5.763	5.779	5.764
	Optimal rate	**0.228**	0.008	0.024	0.016	0.008	0.138	0.073	0.081	**0.423**
1000	Mean ($\times 10^{-2}$)	6.263	6.273	6.263	6.263	6.272	**6.255**	6.262	6.275	**6.253**
	Median ($\times 10^{-2}$)	5.785	5.799	5.779	5.783	5.798	**5.777**	5.780	5.800	**5.778**
	Optimal rate	0.260	0.008	0.024	0.024	0.016	**0.195**	0.016	0.073	**0.382**

Table 3: Diebold–Mariano test results for the differences between MMA and other methods.

K		$\frac{\text{AIC}}{\text{MMA}}$	$\frac{\text{BIC}}{\text{MMA}}$	$\frac{\text{Mallows}}{\text{MMA}}$	$\frac{\text{SAIC}}{\text{MMA}}$	$\frac{\text{SBIC}}{\text{MMA}}$	$\frac{\text{dSAIC}}{\text{MMA}}$	$\frac{\text{dSBIC}}{\text{MMA}}$	$\frac{\text{dMMA}}{\text{MMA}}$
1	DM	5.073	5.078	5.073	5.073	5.078	5.073	5.078	
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
2	DM	4.306	4.321	4.306	4.306	4.318	4.461	4.473	1.263
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.103
5	DM	3.939	3.882	3.939	3.938	3.888	4.179	4.122	1.974
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.024
10	DM	3.783	3.835	3.783	3.773	3.798	4.165	4.192	2.343
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010
100	DM	3.948	6.028	3.959	4.101	5.932	1.864	3.348	0.543
	P-value	0.000	0.000	0.000	0.000	0.000	0.031	0.000	0.294
200	DM	6.438	11.997	6.667	7.054	11.695	8.530	11.685	-2.187
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.986
500	DM	5.024	8.579	5.058	5.252	8.543	0.095	5.735	-1.762
	P-value	0.000	0.000	0.000	0.000	0.000	0.462	0.000	0.961
1000	DM	6.023	10.681	6.132	6.401	10.552	2.828	10.630	-3.277
	P-value	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.999

Table 4: Diebold–Mariano test results for the differences between dMMA and other methods.

K		AIC/dMMA	BIC/dMMA	Mallows/dMMA	SAIC/dMMA	SBIC/dMMA	dSAIC/dMMA	dSBIC/dMMA	MMA/dMMA
2	DM	4.134	4.149	4.134	4.134	4.146	4.294	4.306	-1.263
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.897
5	DM	3.562	3.508	3.562	3.561	3.512	3.813	3.759	-1.974
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.976
10	DM	2.992	3.085	2.992	3.001	3.052	3.466	3.517	-2.343
	P-value	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.990
100	DM	1.069	2.243	1.073	1.119	2.215	2.387	3.157	-0.543
	P-value	0.143	0.012	0.142	0.132	0.013	0.008	0.001	0.706
200	DM	4.835	10.356	4.958	5.152	9.975	5.979	10.969	2.187
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.014
500	DM	3.216	4.933	3.229	3.278	4.926	3.829	5.699	1.762
	P-value	0.001	0.000	0.001	0.001	0.000	0.000	0.000	0.039
1000	DM	5.011	8.798	5.066	5.171	8.634	5.622	9.441	3.277
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001

candidate models and weight estimators in model averaging, and MMA is more applicable to the cases of bigger n and smaller K where MMA requires less computation and has similar performance to dMMA.

Our results hold in both the fixed and divergent dimensional models. For high-dimensional linear models, we can group the regressors in order and then build the nested group candidate models or single group candidate models to reduce the effect of dimension. For example, Ando and Li (2014) proposed cross-validation model averaging framework which groups variables by correlation first to reduce the dimension, and then combines the candidate models by model averaging.

Our theoretical results need a homoscedastic assumption on the error term. If the data at the subject are heteroscedastic or dependent, how to choose weights to aggregate each subject estimator is an important problem. In this regard, the Jackknife model averaging method in Hansen and Racine (2012), heteroscedasticity-robust C_p model averaging method in Liu and Okui (2013), and leave-subject-out cross-validation model averaging method in Gao et al. (2016) are useful and warrant our further research. Some other interesting researches can also be done in the next step. One is to extend the proposed aggregated model averaging methods to the case of big data streams (Xi et al., 2009; Wang et al., 2018). Investigating model averaging based on generalized linear model and other complex models for distributed data is another important topic.

Acknowledgments

The thoughtful and constructive comments and suggestions from Action Editor Sayan Mukherjee and two anonymous referees are gratefully acknowledged. Zhang's work was partially supported by the Research Foundation of Shenzhen Polytechnic University under Grant 6023312034K and Post-doctoral Later-stage Foundation Project of Shenzhen Polytechnic University under Grant 6023271021K. Zou's work was partially supported by the National Natural Science Foundation of China (Grant Nos. 11971323 and 12031016).

Appendices

To prove Theorems 1-5 and 7 in current paper, we first give some lemmas and their proofs in Appendix A, then provide the proofs of the theorems in Appendix B.

Appendix A. Lemmas and Proofs

For $w_0 = (w_1, \ldots, w_{S-1})^T \in \mathbb{R}^{S-1}$, denote

$$C_k(w_0) = C_{k,n}\left(\mathrm{col}\left\{w_0, 1 - \sum_{s=1}^{S-1} w_s\right\}\right).$$

Choosing any small radius $\delta_\rho \leq \rho$, $\rho \in (0,1)$, we define the events

$$\mathcal{E}_1 \triangleq \left\{\left\|\nabla^2 C_k(w_0^*) - \nabla^2 R_0(w_0^*)\right\|_2 \leq \rho\lambda_n\right\},$$

and

$$\mathcal{E}_2 \triangleq \left\{\left\|\nabla C_k\left(w_0^*\right)\right\| \leq \frac{(1-\rho)\,\lambda_n\delta_\rho}{2}\right\}.$$

Lemma 1 *Under the events $\mathcal{E}_1$ and $\mathcal{E}_2$, for $k \in \{1, 2, \ldots, K\}$, we have*

$$\left\|\hat{W}_k - w_0^*\right\| \leq \frac{2\left\|\nabla C_k\left(w_0^*\right)\right\|}{(1-\rho)\,\lambda_n}, \tag{19}$$

and

$$\lambda_{\min}\left[\nabla^2 C_k(w_0)\right] \geqslant (1-\rho)\,\lambda_n, \tag{20}$$

where

$$w_0 \in U_{\delta_\rho} \triangleq \left\{w_0 \in \mathbb{R}^{S-1} \middle| \left\|w_0 - w_0^*\right\| \leq \delta_\rho\right\} \subseteq Q_0.$$

Proof We first prove (20), which means that the function $C_k\left(w_0\right)$ is $(1-\rho)\,\lambda_n$-strongly convex over the feasible set U_{δ_ρ} under the conditions given in the lemma. In fact, for fixed $\tau \in U_{\delta_\rho}$, we have

$$\left\|\nabla^2 C_k\left(\tau\right) - \nabla^2 R_0\left(w_0^*\right)\right\|_2 = \left\|\nabla^2 C_k(w_0^*) - \nabla^2 R_0(w_0^*)\right\|_2 \leq \rho\lambda_n.$$

According to the properties of the spectral radius, it follows that

$$\begin{aligned} \left|\lambda_{\min}\left[\nabla^2 C_k(\tau) - \nabla^2 R_0(w_0^*)\right]\right| &\leq \rho_r\left[\nabla^2 C_k(\tau) - \nabla^2 R_0(w_0^*)\right] \leq \left\|\nabla^2 C_k(\tau) - \nabla^2 R_0(w_0^*)\right\|_2 \\ &\leq \rho\lambda_n. \end{aligned}$$

Hence

$$\begin{aligned} \lambda_{\min}\left[\nabla^2 C_k\left(w_0\right)\right] &\geq \lambda_{\min}\left[\nabla^2 C_k(\tau) - \nabla^2 R_0\left(w_0^*\right)\right] + \lambda_{\min}\left[\nabla^2 R_0\left(w_0^*\right)\right] \\ &\geq -\rho\lambda_n + \lambda_n = (1-\rho)\,\lambda_n, \end{aligned}$$

which implies that $C_k(w_0)$ is $(1-\rho)\lambda_n$-strongly convex on U_{δ_ρ}.

We now prove (19). Here we follow Zhang et al. (2013b) to provide a general proof strategy that can be easily generalized to other non-linear model averaging methods, such as cross-validation model averaging for quantile regression (Lu and Su, 2015) and generalized functional linear model (Zhang and Zou, 2020), although $C_k(w_0)$ is a quadratic function of w_0. Using the fact that $C_k(w_0)$ is strongly convex on the set U_{δ_ρ}, for any $w_0^{'} \in Q$, we obtain

$$C_k\left(w_0'\right) \;\geqslant\; C_k(w_0^*) + \left\langle \nabla C_k\left(w_0^*\right), w_0' - w_0^* \right\rangle + \frac{(1-\rho)\lambda_n}{2}\min\left\{\left\|w_0' - w_0^*\right\|^2, \delta_\rho^2\right\}.$$

Rewriting this inequality, it can be seen that

$$\begin{aligned}
&\min\left\{\left\|w_0' - w_0^*\right\|^2, \delta_\rho^2\right\} \\
&\le \frac{2}{(1-\rho)\lambda_n}\left[C_k\left(w_0'\right) - C_k\left(w_0^*\right) - \left\langle \nabla C_k\left(w_0^*\right), w_0' - w_0^*\right\rangle\right] \\
&\le \frac{2}{(1-\rho)\lambda_n}\left[C_k(w_0') - C_k\left(w_0^*\right) + \|\nabla C_k(w_0^*)\|\left\|w_0' - w_0^*\right\|\right].
\end{aligned} \tag{21}$$

Without loss of generality, let $w_0' = \kappa \hat{W}_{k,0} + (1-\kappa)w_0^*$ for $\kappa \in (0,1]$, then $\|w_0' - w_0^*\| > 0$ and $\|w_0' - w_0^*\|^2 = \kappa^2\left\|\hat{W}_{k,0} - w_0^*\right\|^2$. Dividing both sides of (21) by $\|w_0' - w_0^*\|$ leads to

$$\begin{aligned}
&\min\left\{\kappa\left\|\hat{W}_{k,0} - w_0^*\right\|, \frac{\delta_\rho^2}{\kappa\left\|\hat{W}_{k,0} - w_0^*\right\|}\right\} \\
\le\;& \frac{2\left[C_k\left(\kappa\hat{W}_{k,0} + (1-\kappa)\,w_0^*\right) - C_k(w_0^*)\right]}{\kappa\left\|\hat{W}_{k,0} - w_0^*\right\|(1-\rho)\,\lambda_n} + \frac{2\left\|\nabla C_k\left(w_0^*\right)\right\|}{(1-\rho)\,\lambda_n}.
\end{aligned}$$

By Jensen's inequality, we see that

$$C_k\left(\kappa\hat{W}_{k,0} + (1-\kappa)w_0^*\right) < C_k\left(w_0^*\right),$$

which gives the following inequality

$$\min\left\{\kappa\left\|\hat{W}_{k,0} - w_0^*\right\|, \frac{\delta_\rho^2}{\kappa\left\|\hat{W}_{k,0} - w_0^*\right\|}\right\} < \frac{2\left\|\nabla C_k\left(w_0^*\right)\right\|}{(1-\rho)\,\lambda_n} \le \delta_\rho, \tag{22}$$

where the last inequality follows from the definition of $\mathcal{E}_2$ and the conditions in Lemma 1.

Since (22) holds for any $\kappa \in (0,1]$, if $\left\|\hat{W}_{k,0} - w_0^*\right\| > \delta_\rho$, we can set $\kappa = \frac{\delta_\rho}{\left\|\hat{W}_{k,0} - w_0^*\right\|}$, which yields a contradiction that $\min\{\delta_\rho, \delta_\rho\} < \delta_\rho$. Thus, we have

$$\left\|\hat{W}_{k,0} - w_0^*\right\| \le \delta_\rho.$$

Therefore, from (21) and $w_0' = \kappa \hat{W}_{k,0} + (1-\kappa) w_0^*$ with $\kappa = 1$, we obtain

$$\begin{aligned}\left\|\hat{W}_{k,0} - w_0^*\right\|^2 &\leq \frac{2}{(1-\rho)\,\lambda_n}\left[C_k\left(\hat{W}_{k,0}\right) - C_k\left(w_0^*\right) + \left\|\nabla C_k\left(w_0^*\right)\right\|\left\|\hat{W}_{k,0} - w_0^*\right\|\right] \\ &\leq \frac{2\left\|\nabla C_k\left(w_0^*\right)\right\|}{(1-\rho)\,\lambda_n}\left\|\hat{W}_{k,0} - w_0^*\right\|,\end{aligned}$$

which implies the inequality (19) immediately. ■

Lemma 2 *Assume Conditions 1-3 hold, then*

$$\mathbf{E}\left\|\nabla C_k\left(w_0^*\right)\right\|^2 = O\left(\frac{\sigma_n^2 S p_S}{n}\right), \tag{23}$$

and

$$\mathbf{E}\left\|\nabla^2 C_k\left(w_0^*\right) - \nabla^2 R_0\left(w_0^*\right)\right\|_2^2 = O\left(\frac{S^2 p_S}{n}\right). \tag{24}$$

Proof By the definition of w_0^* and Condition 1, we see that $\nabla R_0(w_0^*) = 0$, which together with (6) gives

$$\begin{aligned} 0 &= \left.\frac{\partial R_0(w_0)}{\partial w_0}\right|_{w_0 = w_0^*} \\ &= \frac{2}{n}\mathbf{E}\begin{bmatrix} Y_k^T\left(P_{k,1} - P_{k,S}\right)\left\{P(w^*)Y_k - \boldsymbol{\mu}_k\right\} \\ \vdots \\ Y_k^T\left(P_{k,S-1} - P_{k,S}\right)\left\{P\left(w^*\right)Y_k - \boldsymbol{\mu}_k\right\}\end{bmatrix} \\ &= \frac{2}{n}\mathbf{E}\begin{bmatrix} \boldsymbol{\mu}_k^T\left(P_{k,1} - P_{k,S}\right)\left\{P(w^*) - I\right\}\boldsymbol{\mu}_k \\ \vdots \\ \boldsymbol{\mu}_k^T\left(P_{k,S-1} - P_{k,S}\right)\left\{P\left(w^*\right) - I\right\}\boldsymbol{\mu}_k\end{bmatrix} \\ &\quad + \frac{2}{n}\mathbf{E}\begin{bmatrix} e_{(k)}^T\left(P_{k,1} - P_{k,S}\right)P(w^*)e_{(k)} \\ \vdots \\ e_{(k)}^T\left(P_{k,S-1} - P_{k,S}\right)P(w^*)e_{(k)}\end{bmatrix}. \end{aligned} \tag{25}$$

Moreover,

$$\begin{aligned} &\nabla C_k(w_0^*) \\ = \;& \frac{2}{n}\begin{bmatrix} Y_k^T\left(P_{k,1} - P_{k,S}\right)\left\{P(w^*)Y_k - \boldsymbol{\mu}_k\right\} \\ \vdots \\ Y_k^T\left(P_{k,S-1} - P_{k,S}\right)\left\{P\left(w^*\right)Y_k - \boldsymbol{\mu}_k\right\}\end{bmatrix} - \frac{2}{n}\begin{bmatrix} Y_k^T\left(P_{k,1} - P_{k,S}\right)e_{(k)} \\ \vdots \\ Y_k^T\left(P_{k,S-1} - P_{k,S}\right)e_{(k)}\end{bmatrix} \\ &+ \frac{2\sigma^2}{n}\begin{bmatrix} tr(P_{k,1} - P_{k,S}) \\ \vdots \\ tr(P_{k,S-1} - P_{k,S})\end{bmatrix}\end{aligned}$$

$$
\begin{aligned}
= \quad & \frac{2}{n}\begin{bmatrix} \boldsymbol{\mu}_k^T\left(P_{k,1}-P_{k,S}\right)\left\{P(w^*)-I\right\}\boldsymbol{\mu}_k \\ \vdots \\ \boldsymbol{\mu}_k^T\left(P_{k,S-1}-P_{k,S}\right)\left\{P\left(w^*\right)-I\right\}\boldsymbol{\mu}_k \end{bmatrix} + \frac{2}{n}\begin{bmatrix} e_{(k)}^T\left(P_{k,1}-P_{k,S}\right)P(w^*)e_{(k)} \\ \vdots \\ e_{(k)}^T\left(P_{k,S-1}-P_{k,S}\right)P(w^*)e_{(k)} \end{bmatrix} \\
& +\frac{4}{n}\begin{bmatrix} e_{(k)}^T\left(P_{k,1}-P_{k,S}\right)\left\{P(w^*)-I\right\}\boldsymbol{\mu}_k \\ \vdots \\ e_{(k)}^T\left(P_{k,S-1}-P_{k,S}\right)\left\{P\left(w^*\right)-I\right\}\boldsymbol{\mu}_k \end{bmatrix} - \frac{2}{n}\begin{bmatrix} e_{(k)}^T\left(P_{k,1}-P_{k,S}\right)e_{(k)} \\ \vdots \\ e_{(k)}^T\left(P_{k,S-1}-P_{k,S}\right)e_{(k)} \end{bmatrix} \\
& +\frac{2\sigma^2}{n}\begin{bmatrix} tr(P_{k,1}-P_{k,S}) \\ \vdots \\ tr(P_{k,S-1}-P_{k,S}) \end{bmatrix} \\
\triangleq \quad & \frac{2}{n}\left(A+C_1+2B_1-D_2\right), \qquad (26)
\end{aligned}
$$

where

$$
A = \begin{bmatrix} \boldsymbol{\mu}_k^T\left(P_{k,1}-P_{k,S}\right)\left\{P(w^*)-I\right\}\boldsymbol{\mu}_k \\ \vdots \\ \boldsymbol{\mu}_k^T\left(P_{k,S-1}-P_{k,S}\right)\left\{P\left(w^*\right)-I\right\}\boldsymbol{\mu}_k \end{bmatrix},
$$

$$
B_1 = \begin{bmatrix} e_{(k)}^T\left(P_{k,1}-P_{k,S}\right)\left\{P(w^*)-I\right\}\boldsymbol{\mu}_k \\ \vdots \\ e_{(k)}^T\left(P_{k,S-1}-P_{k,S}\right)\left\{P\left(w^*\right)-I\right\}\boldsymbol{\mu}_k \end{bmatrix},
$$

$$
C_1 = \begin{bmatrix} e_{(k)}^T\left(P_{k,1}-P_{k,S}\right)P(w^*)e_{(k)} \\ \vdots \\ e_{(k)}^T\left(P_{k,S-1}-P_{k,S}\right)P(w^*)e_{(k)} \end{bmatrix},
$$

and

$$
D_2 = \begin{bmatrix} e_{(k)}^T\left(P_{k,1}-P_{k,S}\right)e_{(k)} \\ \vdots \\ e_{(k)}^T\left(P_{k,S-1}-P_{k,S}\right)e_{(k)} \end{bmatrix} - \sigma^2\begin{bmatrix} tr(P_{k,1}-P_{k,S}) \\ \vdots \\ tr(P_{k,S-1}-P_{k,S}) \end{bmatrix}.
$$

Plugging (25) into (26) and by C_r-inequality, we have

$$
\begin{aligned}
\mathbf{E}\left\|\triangledown C_k\left(w_0^*\right)\right\|^2 &= \mathbf{E}\left\|\frac{2}{n}\left(A-\mathbf{E}A+2B_1+C_1-\mathbf{E}C_1-D_2\right)\right\|^2 \\
&\leq \frac{16}{n^2}\left\{\mathbf{E}\left\|A-EA\right\|^2+4\mathbf{E}\left\|B_1\right\|^2+\mathbf{E}\left\|C_1-EC_1\right\|^2+\mathbf{E}\left\|D_2\right\|^2\right\}.
\end{aligned}
$$

We first estimate $\mathbf{E}\left\|A-EA\right\|^2$. For $s\in\{1,2,\ldots,S\}$, since the transpose of each row in $X_{(k)}$ is independent of each other, and $\mathbf{E}\left|x_{(i)}^T\Pi_s^T\beta_{(s)}\right|^4 \leq \mathbf{E}\left|x_{(i)}^T\Pi_s^T\beta_{(s)}\right|^{\eta+2}+1 < C_b+1$, we obtain

$$
\mathbf{Var}[\beta_{(s)}^T\Pi_s x_{(i)}x_{(i)}^T\Pi_s^T\beta_{(s)}] \leq \mathbf{E}\left|x_{(i)}^T\Pi_s^T\beta_{(s)}\right|^4 < C_b+1,
$$

and

$$
\mathbf{Var}\left[\beta_{(s)}^T X_{k,s}^T X_{k,s}\beta_{(s)}\right] = n\cdot\mathbf{Var}\left(\beta_{(s)}^T\Pi_s x_{(i)}x_{(i)}^T\Pi_s^T\beta_{(s)}\right) = O(n). \qquad (27)
$$

Notice that for any random vectors x and y with $\mathbf{E}\left(x^T y\right) = 0$,

$$
\begin{aligned}
& \mathbf{E}\left[\|x+y\|^2 - \mathbf{E}\|x+y\|^2\right]^2 \\
= \;& \mathbf{E}\left[\|x\|^4 + \left(\|y\|^2 + 2x^T y\right)^2 + 4\|x\|^2 x^T y + 2\|x\|^2\|y\|^2\right] - \left(\mathbf{E}\|x\|^2 + \mathbf{E}\|y\|^2\right)^2 \\
\leq \;& 3\mathbf{E}\|x\|^4 + \mathbf{Var}\left[\|y\|^2 + 2x^T y\right] + 4\mathbf{E}\|x\|^2\|y\|^2. \qquad (28)
\end{aligned}
$$

Hence, by $\mathbf{E}[\delta_{k,s}^T P_{k,s} X_{k,s} \Sigma_s^{-1}\gamma_s] = 0$, and letting $x = P_{k,s}\delta_{k,s}$ and $y = P_{k,s}X_{k,s}\Sigma_s^{-1}\gamma_s$ in (28), it is seen that

$$
\begin{aligned}
\mathbf{Var}\left[\boldsymbol{b}_{k,s}^T P_{k,s}\boldsymbol{b}_{k,s}\right] \;\leq\;& 3\mathbf{E}\left[\delta_{k,s}^T P_{k,s}\delta_{k,s}\right]^2 + \mathbf{Var}\left[\left(2\boldsymbol{b}_{k,s} - X_{k,s}\Sigma_s^{-1}\gamma_s\right)^T X_{k,s}\Sigma_s^{-1}\gamma_s\right] \\
& + 4\mathbf{E}\left[\delta_{k,s}^T P_{k,s}\delta_{k,s}\left(X_{k,s}\Sigma_s^{-1}\gamma_s\right)^T P_{k,s}X_{k,s}\Sigma_s^{-1}\gamma_s\right]. \qquad (29)
\end{aligned}
$$

Since $(2\boldsymbol{b}_{k,s} - X_{k,s}\Sigma_s^{-1}\gamma_s)^T X_{k,s}\Sigma_s^{-1}\gamma_s$ is a sum of n i.i.d. random variables, and Condition 2 implies

$$
\begin{aligned}
\mathbf{E}\left|x_{(i)}^T \Pi_s^T \Sigma_s^{-1}\gamma_s\right|^2 \;\leq\;& \frac{1}{2} + \frac{1}{2}\mathbf{E}\left|x_{(i)}^T \Pi_s^T\left(\beta_{(s)} - \beta_{\star,s}\right)\right|^4 \\
\leq\;& \frac{1}{2} + 4\mathbf{E}\left(\left|x_{(i)}^T \Pi_s^T \beta_{(s)}\right|^4 + \left|x_{(i)}^T \Pi_s^T \beta_{\star,s}\right|^4\right) < \frac{17}{2} + 4C_b,
\end{aligned}
$$

it follows that

$$
\begin{aligned}
& \mathbf{Var}\left[\left(2\boldsymbol{b}_{k,s} - X_{k,s}\Sigma_s^{-1}\gamma_s\right)^T X_{k,s}\Sigma_s^{-1}\gamma_s\right] \\
\leq\;& \sum_{i=1}^n \mathbf{E}\left|(2b_{k,i,s} - x_{(i)}^T\Pi_s^T\Sigma_s^{-1}\gamma_s)^T x_{(i)}^T \Pi_s^T \Sigma_s^{-1}\gamma_s\right|^2 \\
\leq\;& 4\sigma_n^2 n\mathbf{E}\left[\gamma_s^T\Sigma_s^{-1}\Pi_s x_{(i)} x_{(i)}^T \Pi_s^T \Sigma_s^{-1}\gamma_s\right] + 2n\mathbf{E}\left|x_{(i)}^T\Pi_s^T\Sigma_s^{-1}\gamma_s\right|^4 \\
\leq\;& O\left(\sigma_n^2 n\right) + O(n), \qquad (30)
\end{aligned}
$$

which holds uniformly for $1 \leq s \leq S$. Moreover, Conditions 2 and 3 lead to

$$
\begin{aligned}
& \mathbf{E}\left[\delta_{k,s}^T P_{k,s}\delta_{k,s}(X_{k,s}\Sigma_s^{-1}\gamma_s)^T P_{k,s}X_{k,s}\Sigma_s^{-1}\gamma_s\right] \\
=\;& \mathbf{E}\left[\delta_{k,s}^T P_{k,s}\delta_{k,s}\gamma_s^T\Sigma_s^{-1}X_{k,s}^T X_{k,s}\Sigma_s^{-1}\gamma_s\right] \\
\leq\;& \sigma_n^2 p_s \mathbf{E}\left[\gamma_s^T\Sigma_s^{-1}X_{k,s}^T X_{k,s}\Sigma_s^{-1}\gamma_s\right] \\
=\;& O(\sigma_n^2 p_s n), \qquad (31)
\end{aligned}
$$

which is uniformly true for all $1 \leq s \leq S$, and

$$
\begin{aligned}
& \mathbf{E}\left[\delta_{k,s}^T P_{k,s}\delta_{k,s}\right]^2 \\
\leq\;& p_s\mathbf{E}\left[\lambda_{\max}\left(\mathbf{E}\left(\delta_{k,s}\delta_{k,s}^T\delta_{k,s}\delta_{k,s}^T \,\middle|\, X_{k,s}\right)\right)\right] \\
\leq\;& \sigma_n^2 n p_s. \qquad (32)
\end{aligned}
$$

Combining (29)–(32), we obtain

$$
\mathbf{E}\left[\boldsymbol{b}_{k,s}^T P_{k,s}\boldsymbol{b}_{k,s} - \mathbf{E}\left\{\boldsymbol{b}_{k,s}^T P_{k,s}\boldsymbol{b}_{k,s}\right\}\right]^2 = O\left(np_s\sigma_n^2\right) + O\left(n\right). \qquad (33)
$$

Further, from Condition 2, we have

$$\begin{aligned}
& \mathbf{Var}\left[\boldsymbol{b}_{k,s}^T X_{k,s}\beta_{(s)}\right] \\
= & \sum_{i=1}^{n} \mathbf{Var}\left[b_{k,i,s} x_{(k,i,s)}^T \Pi_s^T \beta_{(s)}\right] \\
\leq & 2n\mathbf{E}\left[\left(b_{k,i,s} - x_{(k,i,s)}^T \Sigma_s^{-1}\gamma_s\right) x_{(k,i,s)}^T \Pi_s^T \beta_{(s)}\right]^2 \\
& +2n\mathbf{E}\left[\left(x_{(k,i,s)}^T \Sigma_s^{-1}\gamma_s\right) x_{(k,i,s)}^T \beta_{(s)}\right]^2 \\
\leq & O(n\sigma_n^2) + O\left(n\right). \qquad (34)
\end{aligned}$$

Combining (27), (33) and (34), one has

$$\begin{aligned}
& \mathbf{Var}\left[\boldsymbol{\mu}_k^T P_{k,s}\boldsymbol{\mu}_k\right] \\
= & \mathbf{E}\Big[\beta_{(s)}^T X_{k,s}^T X_{k,s}\left(X_{k,s}^T X_{k,s}\right)^{-1} X_{k,s}^T X_{k,s}\beta_{(s)} - \mathbf{E}\left\{\beta_{(s)}^T X_{k,s}^T X_{k,s}\left(X_{k,s}^T X_{k,s}\right)^{-1} X_{k,s}^T X_{k,s}\beta_{(s)}\right\} \\
& +\boldsymbol{b}_{k,s}^T X_{k,s}\left(X_{k,s}^T X_{k,s}\right)^{-1} X_{k,s}^T \boldsymbol{b}_{k,s} - \mathbf{E}\left\{\boldsymbol{b}_{k,s}^T X_{k,s}\left(X_{k,s}^T X_{k,s}\right)^{-1} X_{k,s}^T \boldsymbol{b}_{k,s}\right\} \\
& +2\boldsymbol{b}_{k,s}^T X_{k,s}\left(X_{k,s}^T X_{k,s}\right)^{-1} X_{k,s}^T X_{k,s}\beta_{(s)} - 2\mathbf{E}[\boldsymbol{b}_{k,s}^T X_{k,s}\left(X_{k,s}^T X_{k,s}\right)^{-1} X_{k,s}^T X_{k,s}\beta_{(s)}]\Big]^2 \\
\leq & 3\Bigg(\mathbf{E}\left[\beta_{(s)}^T X_{k,s}^T X_{k,s}\beta_{(s)} - \mathbf{E}\left(\beta_{(s)}^T X_{k,s}^T X_{k,s}\beta_{(s)}\right)\right]^2 \\
& +\mathbf{Var}\left[\boldsymbol{b}_{k,s}^T X_{k,s}\left(X_{k,s}^T X_{k,s}\right)^{-1} X_{k,s}^T \boldsymbol{b}_{k,s}\right] + 4\mathbf{Var}\left[\boldsymbol{b}_{k,s}^T X_{k,s}\beta_{(s)}\right]\Bigg) \\
= & O\left(np_s\sigma_n^2\right) + O\left(n\right), \qquad (35)
\end{aligned}$$

which also holds uniformly for $1 \leq s \leq S$ and $1 \leq k \leq K$. On the other hand, it is clear that $P_{k,s}$ is an idempotent matrix. According to the assumption that all candidate models are nested, we see that $P_{k,i}P_{k,j} = P_{k,j}P_{k,i} = P_{k,\min_{i,j}}$ holds. Thus,

$$\begin{aligned}
& \left(P_{k,s} - P_{k,S}\right)\left\{P_k\left(w^*\right) - I_n\right\} \\
= & \left(P_{k,s} - P_{k,S}\right)\left\{\sum_{j=1}^{S} w_j^* P_{k,j} - \left(\sum_{j=1}^{S} w_j^*\right) I_n\right\} \\
= & \sum_{j=1}^{S} w_j^*\left(P_{k,s} - P_{k,S}\right)\left(P_{k,j} - I_n\right) \\
= & \left(\sum_{j=1}^{S-1} w_j^*\right) P_{k,S} - \left(\sum_{j=1}^{s} w_j^*\right) P_{k,s} - \sum_{j=s+1}^{S-1} w_j^* P_{k,j}, \qquad (36)
\end{aligned}$$

which together with (35) implies

$$
\begin{aligned}
& \mathbf{E}\left\|A-EA\right\|^2 \\
= & \mathbf{E}\left\{\sum_{s=1}^{S-1}\left[\left\{\left(\sum_{j=1}^{S-1} w_j^*\right)\boldsymbol{\mu}_k^T P_{k,S}\boldsymbol{\mu}_k-\left(\sum_{j=1}^{S-1} w_j^*\right)\mathbf{E}\left(\boldsymbol{\mu}_k^T P_{k,S}\boldsymbol{\mu}_k\right)\right\}\right.\right. \\
& -\left(\sum_{j=1}^{s} w_j^*\right)\left\{\boldsymbol{\mu}_k^T P_{k,s}\boldsymbol{\mu}_k-\mathbf{E}\left(\boldsymbol{\mu}_k^T P_{k,s}\boldsymbol{\mu}_k\right)\right\} \\
& \left.\left.-\sum_{j=s+1}^{S-1} w_j^*\left\{\boldsymbol{\mu}_k^T P_{k,j}\boldsymbol{\mu}_k-\mathbf{E}\left(\boldsymbol{\mu}_k^T P_{k,j}\boldsymbol{\mu}_k\right)\right\}\right]^2\right\} \\
\leq & \mathbf{E}\left\{3\sum_{s=1}^{S-1}\left[\left(\sum_{j=1}^{S-1} w_j^*\right)^2\left\{\boldsymbol{\mu}_k^T P_{k,S}\boldsymbol{\mu}_k-\mathbf{E}\left(\boldsymbol{\mu}_k^T P_{k,S}\boldsymbol{\mu}_k\right)\right\}^2\right.\right. \\
& +\left(\sum_{j=1}^{s} w_j^*\right)^2\left\{\boldsymbol{\mu}_k^T P_{k,s}\boldsymbol{\mu}_k-\mathbf{E}\left(\boldsymbol{\mu}_k^T P_{k,s}\boldsymbol{\mu}_k\right)\right\}^2 \\
& \left.\left.+\left(\sum_{j=s+1}^{S-1} w_j^*\left\{\boldsymbol{\mu}_k^T P_{k,j}\boldsymbol{\mu}_k-\mathbf{E}\left(\boldsymbol{\mu}_k^T P_{k,j}\boldsymbol{\mu}_k\right)\right\}\right)^2\right]\right\} \\
\leq & 3\sum_{s=1}^{S-1}\left(\mathbf{Var}\left[\boldsymbol{\mu}_k^T P_{k,S}\boldsymbol{\mu}_k\right]+\mathbf{Var}\left[\boldsymbol{\mu}_k^T P_{k,s}\boldsymbol{\mu}_k\right]+\max_{s+1\leq j\leq S-1}\mathbf{Var}\left[\boldsymbol{\mu}_k^T P_{k,j}\boldsymbol{\mu}_k\right]\right) \\
= & O\left(nSp_S\sigma_n^2\right)+O\left(nS\right). \qquad (37)
\end{aligned}
$$

Next, we will bound $\mathbf{E}\left\|B_1\right\|^2$. With (35) and (36), it is clear that

$$
\begin{aligned}
\mathbf{E}\left\|B_1\right\|^2 &= \mathbf{E}\left[\sum_{s=1}^{S-1}\left\{e_{(k)}^T\left(P_{k,s}-P_{k,S}\right)P_k\left(w^*\right)\boldsymbol{\mu}_k\right\}^2\right] \\
&\leq S\max_{1\leq s\leq S}\mathbf{E}\left[e_{(k)}^T P_{k,s}\boldsymbol{\mu}_k\right]^2 \\
&= S\max_{1\leq s\leq S}\mathbf{E}\left[\boldsymbol{\mu}_k^T P_{k,s}\boldsymbol{\mu}_k\right]^2\sigma^2 \\
&= O\left(nSp_S\right). \qquad (38)
\end{aligned}
$$

For $\mathbf{E}\left\|C_1-EC_1\right\|^2$, with Theorem 2 of Whittle (1960), we have

$$
\begin{aligned}
& \mathbf{Var}\left[e_{(k)}^T\left(P_{k,s}-P_{k,S}\right)P\left(w^*\right)e_{(k)}\middle| X_{(k)}\right] \\
= & O\left(\left\|(P_{k,s}-P_{k,S})P\left(w^*\right)\right\|_F^2\mathbf{E}e_{k,i}^4\right) \\
= & O\left(p_S\right),
\end{aligned}
$$

and then

$$\begin{aligned}
\mathbf{E}\left\|C_1 - EC_1\right\|^2 &= \sum_{s=1}^{S-1} \mathbf{Var}\left[e_{(k)}^T \left(P_{k,s} - P_{k,S}\right) P\left(w^*\right) e_{(k)}\right] \\
&= O\left(Sp_S\right). \qquad (39)
\end{aligned}$$

Similar to (39), we see that

$$\mathbf{E}\left\|D_2\right\|^2 = O\left(Sp_S^2\right). \qquad (40)$$

Combining (37)-(40), we get (23).

To prove (24), we calculate

$$\begin{aligned}
\nabla^2 C_k\left(w_0^*\right) &= \frac{2}{n}\left\{\left(X_{k,s_1}\hat{\beta}_{k,s_1} - X_{k,S}\hat{\beta}_{k,S}\right)^T \left(X_{k,s_2}\hat{\beta}_{k,s_2} - X_{k,S}\hat{\beta}_{k,S}\right)\right\}_{1\le s_1,s_2\le S-1} \\
&= \frac{2}{n}\left\{Y_k^T \left(P_{k,s_1} - P_{k,S}\right)^T \left(P_{k,s_2} - P_{k,S}\right) Y_k\right\}_{1\le s_1,s_2\le S-1} \\
&= \frac{2}{n}\left\{\left(\boldsymbol{\mu}_k + e_{(k)}\right)^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right)\left(\boldsymbol{\mu}_k + e_{(k)}\right)\right\}_{1\le s_1,s_2\le S-1}.
\end{aligned}$$

With the help of Theorem 2 of Whittle (1960), it can be claimed that

$$\begin{aligned}
& \mathbf{Var}\left[e_{(k)}^T \left(P_{k,s_1} - P_{k,S}\right)^T \left(P_{k,s_2} - P_{k,S}\right) e_{(k)}\right] \\
= \; & \mathbf{E}\left(\mathbf{Var}\left[e_{(k)}^T \left(P_{k,s_1} - P_{k,S}\right)^T \left(P_{k,s_2} - P_{k,S}\right) e_{(k)}\middle| X_{(k)}\right]\right) \\
& + \mathbf{Var}\left(\mathbf{E}\left[e_{(k)}^T \left(P_{k,s_1} - P_{k,S}\right)^T \left(P_{k,s_2} - P_{k,S}\right) e_{(k)}\middle| X_{(k)}\right]\right) \\
= \; & O\left(tr\left[\left(P_{k,s_1} - P_{k,S}\right)^2 \left(P_{k,s_2} - P_{k,S}\right)^2\right]\right) + \mathbf{Var}\left\{\sigma^2 tr\left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right)\right\} \\
= \; & O\left(tr\left[\left(P_{k,s_1} - P_{k,S}\right)^2 \left(P_{k,s_2} - P_{k,S}\right)^2\right]\right) + \mathbf{Var}\left\{\sigma^2\left(p_S - p_{\max\{s_1,s_2\}}\right)\right\} \\
= \; & O\left(p_S\right)
\end{aligned}$$

and

$$\begin{aligned}
& \mathbf{E}\left[e_{(k)}^T \left(P_{k,s_1} - P_{k,S}\right)^T \left(P_{k,s_2} - P_{k,S}\right)\boldsymbol{\mu}_k\right]^2 \\
= \; & \mathbf{E}\left[\mathbf{E}\left\{\left|e_{(k)}^T \left(P_{k,\max\{s_1,s_2\}} - P_{k,S}\right)\boldsymbol{\mu}_k\right|^2 \middle| X_{(k)}\right\}\right] = O\left(np_S\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
& \mathbf{Var}\left[\left(\boldsymbol{\mu}_k + e_{(k)}\right)^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right)\left(\boldsymbol{\mu}_k + e_{(k)}\right)\right] \\
= \ & \mathbf{E}\left[\boldsymbol{\mu}_k^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right)\boldsymbol{\mu}_k - \mathbf{E}\left\{\boldsymbol{\mu}_k^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right)\boldsymbol{\mu}_k\right\}\right. \\
& \quad + e_{(k)}^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right) e_{(k)} - \mathbf{E}\left\{e_{(k)}^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right) e_{(k)}\right\} \\
& \quad \left. + 2\boldsymbol{\mu}_k^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right) e_{(k)}\right]^2 \\
\leq \ & 3\left(\mathbf{Var}\left[\boldsymbol{\mu}_k^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right)\boldsymbol{\mu}_k\right] + \mathbf{Var}\left[e_{(k)}^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right) e_{(k)}\right]\right. \\
& \quad \left. + 4\mathbf{E}\left\{\boldsymbol{\mu}_k^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right) e_{(k)}\right\}^2\right) \\
= \ & O\left(np_S\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
& \mathbf{E}\left\|\nabla^2 C_k\left(w_0^*\right) - \nabla^2 \mathbf{E}\left\{C_k\left(w_0^*\right)\right\}\right\|_2^2 \leq \mathbf{E}\left\|\nabla^2 C_k\left(w_0^*\right) - \nabla^2 \mathbf{E}\left\{C_k\left(w_0^*\right)\right\}\right\|_F^2 \\
= \ & \frac{4}{n^2}\sum_{s_1=1}^{S-1}\sum_{s_2=1}^{S-1} \mathbf{Var}\left[\left(\boldsymbol{\mu}_k + e_{(k)}\right)^T \left(P_{k,S} - P_{k,\max\{s_1,s_2\}}\right)\left(\boldsymbol{\mu}_k + e_{(k)}\right)\right] = O\left(n^{-1}S^2 p_S\right),
\end{aligned}$$

which completes the proof of (24). ∎

Lemma 3 *Under Conditions 1-4, we have*

$$\mathbf{E}\left\|\hat{W}_{k,0} - w_0^*\right\|^2 = O\left(\frac{Sp_S(S+\sigma_n^2)}{\lambda_n^2 n}\right),$$

and then

$$\mathbf{E}\left\|\hat{W}_k - w^*\right\|^2 = O\left(\frac{S^2 p_S(S+\sigma_n^2)}{\lambda_n^2 n}\right).$$

Proof Recalling the events $\mathcal{E}_1$ and $\mathcal{E}_2$, we define the event $\mathcal{E} \triangleq \mathcal{E}_1 \cap \mathcal{E}_2$. In view of Lemma 1, we get

$$\begin{aligned}
& \mathbf{E}\left\|\hat{W}_{k,0} - w_0^*\right\|^2 = \mathbf{E}\left[1_{(\mathcal{E})}\left\|\hat{W}_{k,0} - w_0^*\right\|^2\right] + \mathbf{E}\left[1_{(\mathcal{E}^c)}\left\|\hat{W}_{k,0} - w_0^*\right\|^2\right] \\
\leq \ & \frac{4\mathbf{E}\left[1_{(\mathcal{E})}\left\|\nabla C_k\left(w_0^*\right)\right\|^2\right]}{\left(1-\rho\right)^2\lambda_n^2} + 2\mathbf{P}\left(\mathcal{E}^c\right) \leq \frac{4\mathbf{E}\left\|\nabla C_k\left(w_0^*\right)\right\|^2}{\left(1-\rho\right)^2\lambda_n^2} + 2\mathbf{P}\left(\mathcal{E}^c\right).
\end{aligned} \tag{41}$$

From Lemma 2 and some direct calculations, we obtain

$$\begin{aligned}
\mathbf{P}(\mathcal{E}^c) &= \mathbf{P}(\mathcal{E}_1^c \cup \mathcal{E}_2^c) \leq \mathbf{P}(\mathcal{E}_1^c) + \mathbf{P}(\mathcal{E}_2^c) \\
&\leq \frac{\mathbf{E}\left\|\nabla^2 C_k(w_0^*) - \nabla^2 R_0\left(w_0^*\right)\right\|_2^2}{\rho^2\lambda_n^2} + \frac{4\mathbf{E}\left\|\nabla C_k\left(w_0^*\right)\right\|^2}{\left(1-\rho\right)^2\lambda_n^2\delta_\rho^2} \\
&= O\left(Sp_S\lambda_n^{-2}n^{-1}(S+\sigma_n^2)\right),
\end{aligned}$$

which together with (41) leads to

$$\mathbf{E}\left\|\hat{W}_{k,0}-w_0^*\right\|^2=O\left(Sp_S\lambda_n^{-2}n^{-1}(S+\sigma_n^2)\right).$$

This completes the proof of Lemma 3. ■

Lemma 4 *Under Condition 2, for any random variable a with $\|a\|^2\leq 2$, we have*

$$\mathbf{E}\left[\max_{1\leq s\leq S}\left(x_{(i)}^T\Pi_s^T\beta_{\star,s}\right)^2\|a\|^2\right]=O\left(S^{2/(\eta+2)}\left(\mathbf{E}\|a\|^2\right)^{\eta/(\eta+2)}\right).$$

Proof Define the event

$$\mathcal{E}_3=\left\{\max_{1\leq s\leq S}\left(x_{(i)}^T\Pi_s^T\beta_{\star,s}\right)^2\leq\left(\mathbf{E}\|a\|^2/S\right)^{-2/(\eta+2)}\right\},$$

then

$$\begin{aligned}
&\mathbf{E}\left[\max_{1\leq s\leq S}\left(x_{(i)}^T\Pi_s^T\beta_{\star,s}\right)^2\|a\|^2\right]\\
\leq\;&\mathbf{E}\left[1_{(\mathcal{E}_3)}\max_{1\leq s\leq S}\left(x_{(i)}^T\Pi_s^T\beta_{\star,s}\right)^2\|a\|^2\right]+2S\mathbf{E}\left[1_{(\mathcal{E}_3^c)}\left(x_{(i)}^T\Pi_s^T\beta_{\star,s}\right)^2\right]\\
\leq\;&\left(\mathbf{E}\|a\|^2/S\right)^{-2/(\eta+2)}\mathbf{E}\|a\|^2+2S\left(\mathbf{E}\|a\|^2/S\right)^{\eta/(\eta+2)}\\
=\;&O\left(S^{2/(\eta+2)}\left(\mathbf{E}\|a\|^2\right)^{\eta/(\eta+2)}\right),
\end{aligned}$$

which leads to Lemma 4. ■

Lemma 5 *Under Condition 3, for $s=1,2,\ldots,S$, we have*

$$\mathbf{E}\left[\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}^T\Pi_sX_k^TX_k\Pi_s^T\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right]=O\left(p_s\sigma_n^2\right),$$

and

$$\max_{1\leq s\leq S}\mathbf{E}\left[\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}^T\Pi_sX_k^TX_k\Pi_s^T\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right]=O\left(p_S\sigma_n^2\right).$$

Proof Since

$$\hat{\beta}_{k,s}-\beta_{\star,s}\;=\;\left(X_{k,s}^TX_{k,s}\right)^{-1}X_{k,s}^T\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s+e_{(k)}\right),$$

we obtain

$$
\begin{aligned}
& \mathbf{E}\left[\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}^T \Pi_s X_k^T X_k \Pi_s^T \left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right] \\
= \; & \mathbf{E}\left[\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s+e_{(k)}\right)^T X_{k,s}\left(X_{k,s}^T X_{k,s}\right)^{-1} X_{k,s}^T \left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s+e_{(k)}\right)\right] \\
= \; & \mathbf{E}\left[\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s\right)^T P_{k,s}\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s\right)\right]+\mathbf{E}\left[e_{(k)}^T P_{k,s} e_{(k)}\right] \\
\leq \; & p_s \mathbf{E}\left[\lambda_{\max}\left(\mathbf{E}\left(\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s\right)\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s\right)^T | X_{k,s}\right)\right)\right]+\sigma^2 p_s \\
= \; & O\left(p_s\sigma_n^2\right)=O\left(p_S\sigma_n^2\right),
\end{aligned}
$$

which is uniformly true for all $1 \leq s \leq S$. Then Lemma 5 follows. ■

Lemma 6 *Under Conditions 3 and 6, for $s=1,2,\ldots,S$, we have*

$$
\mathbf{E}\left[\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}^T \Pi_s \Sigma_S \Pi_s^T \left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right]=O\left(\frac{p_s\sigma_n^2}{n}\right),
$$

and

$$
\max_{1\leq s\leq S}\left[\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}^T \Pi_s \Sigma_S \Pi_s^T \left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right]=O\left(\frac{p_S\sigma_n^2}{n}\right).
$$

Proof Denote $\overline{P}_{k,s}=X_{k,s}(X_{k,s}^T X_{k,s})^{-1}\Sigma_S(X_{k,s}^T X_{k,s})^{-1}X_{k,s}^T$, then it follows that $tr[\overline{P}_{k,s}]=tr[(X_{k,s}^T X_{k,s})^{-1}\Sigma_s]$, and hence

$$
\begin{aligned}
& \mathbf{E}\left[\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}^T \Pi_s \Sigma_S \Pi_s^T \left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right] \\
= \; & \mathbf{E}\left[\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s+e_{(k)}\right)^T \overline{P}_{k,s}\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s+e_{(k)}\right)\right] \\
= \; & \mathbf{E}\left[\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s\right)^T \overline{P}_{k,s}\left(\boldsymbol{b}_{k,s}-X_{k,s}\Sigma_s^{-1}\gamma_s\right)\right]+\mathbf{E}\left[e_{(k)}^T \overline{P}_{k,s} e_{(k)}\right] \\
\leq \; & \mathbf{E}\left\{tr\left[\left(X_{k,s}^T X_{k,s}\right)^{-1}\Sigma_s\right]\lambda_{\max}\left(\Sigma_{\infty|s}\right)\right\}+\sigma^2\mathbf{E}\left\{tr\left[\left(X_{k,s}^T X_{k,s}\right)^{-1}\Sigma_s\right]\right\} \\
= \; & O\left(\frac{p_s\sigma_n^2}{n}\right)=O\left(\frac{p_S\sigma_n^2}{n}\right),
\end{aligned}
$$

which uniformly holds for $1 \leq s \leq S$. ■

Appendix B. Proofs of Theorems

Proof of Theorem 1

To obtain the bound of $\mathbf{E}\|\overline{w}_0-w_0^*\|^2$, we first show that the function $C_k(w_0)$ behaves similarly to the risk function $R_0(w_0)$ in the neighborhood of the point w_0^* under the two events $\mathcal{E}_1$ and $\mathcal{E}_2$. Intuitively, $R_0(w_0)$ is locally strongly convex, so the minimizer $\hat{W}_{k,0}$ of $C_k(w_0)$ will be close to w_0^*. Hence our idea is to show that the events $\mathcal{E}_1$ and $\mathcal{E}_2$ hold with high probability, which will guarantee the closeness of $\hat{W}_{k,0}$ and w_0^*.

From the definition of $\bar{w}_0$, it is seen that

$$\begin{aligned}
\mathbf{E}\left\|\overline{w}_0 - w_0^*\right\|^2 &= \mathbf{E}\left\|\frac{1}{K}\sum_{k=1}^{K}\hat{W}_{k,0} - w_0^*\right\|^2 \\
&= \frac{1}{K^2}\mathbf{E}\left\{\sum_{k=1}^{K}\left\|\hat{W}_{k,0} - w_0^*\right\|^2 + \sum_{k\neq j}\left\langle \hat{W}_{k,0} - w_0^*, \hat{W}_{j,0} - w_0^*\right\rangle\right\} \\
&= \frac{1}{K^2}\sum_{k=1}^{K}\mathbf{E}\left\|\hat{W}_{k,0} - w_0^*\right\|^2 + \frac{1}{K^2}\sum_{k\neq j}\left\langle \mathbf{E}\left(\hat{W}_{k,0} - w_0^*\right), \mathbf{E}\left(\hat{W}_{j,0} - w_0^*\right)\right\rangle \\
&= \frac{1}{K}\mathbf{E}\left\|\hat{W}_{1,0} - w_0^*\right\|^2 + \frac{K(K-1)}{K^2}\left\|\mathbf{E}\left(\hat{W}_{1,0} - w_0^*\right)\right\|^2 \\
&\leq \frac{1}{K}\mathbf{E}\left\|\hat{W}_{1,0} - w_0^*\right\|^2 + \left\|\mathbf{E}\left(\hat{W}_{1,0} - w_0^*\right)\right\|^2, \qquad (42)
\end{aligned}$$

where the third equality is from the fact that the weights $\hat{W}_{k,0}$ and $\hat{W}_{j,0}$ are independent. The upper bound in (42) illuminates the path for the remainder of our proof: We only need to bound $\mathbf{E}\left\|\hat{W}_{1,0} - w_0^*\right\|^2$ and $\left\|\mathbf{E}\left(\hat{W}_{1,0} - w_0^*\right)\right\|^2$.

Noting that Lemma 3 gives the bound on $\mathbf{E}\left\|\hat{W}_{1,0} - w_0^*\right\|^2$, we derive the bound on $\left\|\mathbf{E}\left(\hat{W}_{1,0} - w_0^*\right)\right\|^2$ below. With the fact that $\triangledown C_1(\hat{W}_{1,0}) = 0$, and the Taylor series expansion of $\triangledown C_1(\hat{W}_{1,0})$ at w_0^*, we have

$$0 = \triangledown C_1\left(\hat{W}_{1,0}\right) = \triangledown C_1\left(w_0^*\right) + \triangledown^2 C_1\left(w_0'\right)\left(\hat{W}_{1,0} - w_0^*\right),$$

where $w_0' = \kappa w_0^* + (1-\kappa)\hat{W}_{1,0}$ for some $\kappa \in [0,1]$. Clearly, this is equivalent to

$$0 = \triangledown C_1\left(w_0^*\right) + \left[\triangledown^2 C_1\left(w_0^*\right) - \triangledown^2 R_0\left(w_0^*\right)\right]\left(\hat{W}_{1,0} - w_0^*\right) + \triangledown^2 R_0\left(w_0^*\right)\left(\hat{W}_{1,0} - w_0^*\right). \qquad (43)$$

By Condition 1, we can set $\Sigma = \triangledown^2 R_0\left(w_0^*\right)$ and $\Sigma^{-1} = \left[\triangledown^2 R_0\left(w_0^*\right)\right]^{-1}$. Multiplying both sides of (43) by Σ^{-1}, we obtain

$$\hat{W}_{1,0} - w_0^* = -\Sigma^{-1}\triangledown C_1\left(w_0^*\right) + \Sigma^{-1}\left[\triangledown^2 R_0\left(w_0^*\right) - \triangledown^2 C_1\left(w_0^*\right)\right](\hat{W}_{1,0} - w_0^*).$$

Therefore, by Lemmas 2 and 3, it is seen that

$$\begin{aligned}
\left\|\mathbf{E}\left(\hat{W}_{1,0} - w_0^*\right)\right\| &= \left\|\mathbf{E}\left\{\Sigma^{-1}\left(\triangledown^2 R_0\left(w_0^*\right) - \triangledown^2 C_1\left(w_0^*\right)\right)\left(\hat{W}_{1,0} - w_0^*\right)\right\}\right\| \\
&\leq \mathbf{E}\left\|\Sigma^{-1}\left(\triangledown^2 R_0\left(w_0^*\right) - \triangledown^2 C_1\left(w_0^*\right)\right)\left(\hat{W}_{1,0} - w_0^*\right)\right\| \\
&\leq \left(\mathbf{E}\left\|\Sigma^{-1}\left(\triangledown^2 R_0\left(w_0^*\right) - \triangledown^2 C_1\left(w_0^*\right)\right)\right\|_2^2\right)^{1/2}\left(\mathbf{E}\left\|\left(\hat{W}_{1,0} - w_0^*\right)\right\|^2\right)^{1/2} \\
&= O\left(S^{3/2}(S+\sigma_n^2)^{1/2}p_S\lambda_n^{-2}n^{-1}\right). \qquad (44)
\end{aligned}$$

By combining Lemma 5, Condition 4 and

$$|\lambda_n - \overline{\lambda}_S| \leq O\left(Sn^{-1/2}\max_{1\leq s\leq S}\mathbf{E}^{1/2}\left[\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}^T \Pi_s X_k^T X_k \Pi_s^T \left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right]\right),$$

it follows that $\lambda_n = \overline{\lambda}_S + o(\overline{\lambda}_S)$, which together with (44) leads to

$$\begin{aligned}\mathbf{E}\left\|\overline{w}_0 - w_0^*\right\|^2 &\leq \frac{1}{K}\mathbf{E}\left\|\hat{W}_{1,0}-w_0^*\right\|^2 + \left\|\mathbf{E}\left(\hat{W}_{1,0}-w_0^*\right)\right\|^2 \\ &= O\left(\frac{Sp_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2}\right) + O\left(\frac{S^3p_S^2(S+\sigma_n^2)}{n^2\overline{\lambda}_S^4}\right).\end{aligned}$$

Theorem 1 is proved. ■

Proof of Theorem 2

Noting that

$$\begin{aligned} & NL_N(w) \\ = & \sum_{k=1}^K \left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w) + X_k\beta_\star(w) - \boldsymbol{\mu}_k\right\|^2 \\ = & \sum_{k=1}^K \left\{\left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w)\right\|^2 + \left\|X_k\beta_\star(w)-\boldsymbol{\mu}_k\right\|^2 + 2\left\langle\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w), X_k\beta_\star(w)-\boldsymbol{\mu}_k\right\rangle\right\} \\ = & \sum_{k=1}^K \left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w)\right\|^2 + L_{N,\star}(w) + 2\sum_{k=1}^K\left\langle\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w), X_k\beta_\star(w)-\mu_k\right\rangle,\end{aligned}$$

we have

$$\begin{aligned} & NR_N(w) \\ = & NR_N^*(w) + \sum_{k=1}^K\mathbf{E}\left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w)\right\|^2 + 2\sum_{k=1}^K\mathbf{E}\left\langle\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w), X_k\beta_\star(w)-\boldsymbol{\mu}_k\right\rangle \\ \leq & NR_N^*(w) + \sum_{k=1}^K\mathbf{E}\left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w)\right\|^2 + 2\sqrt{NR_N^*(w)\sum_{k=1}^K\mathbf{E}\left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w)\right\|^2},\end{aligned}$$

and

$$\left|\frac{R_N(w)-R_N^*(w)}{R_N^*(w)}\right| \leq \frac{\sum_{k=1}^K\mathbf{E}\left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w)\right\|^2}{NR_N^*(w)} + 2\sqrt{\frac{\sum_{k=1}^K\mathbf{E}\left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w)\right\|^2}{NR_N^*(w)}}.$$

So we need only to prove

$$\sup_{w\in Q}\frac{\sum_{k=1}^K\mathbf{E}\left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w)\right\|^2}{NR_N^*(w)} = o(1).$$

Since

$$\begin{aligned}
\mathbf{E}\left\|\hat{\boldsymbol{\mu}}_k - X_k\beta_\star(w)\right\|^2 &= \mathbf{E}\left\|\sum_{s=1}^{S} w_s X_k \Pi_s^T\left(\hat{\beta}_{k,s} - \beta_{\star,s}\right)\right\|^2 \\
&\le \max_{1\le s\le S}\mathbf{E}\left\|X_k\Pi_s^T\left(\hat{\beta}_{k,s} - \beta_{\star,s}\right)\right\|^2,
\end{aligned}$$

it is sufficient to prove that

$$\frac{1}{n}\max_{1\le s\le S}\mathbf{E}\left\|X_k\Pi_s^T\left(\hat{\beta}_{k,s} - \beta_{\star,s}\right)\right\|^2 = o\left(\inf_{w\in Q} R_N^*(w)\right). \tag{45}$$

By the definitions of $\hat{\beta}_{k,s}$ and $\beta_{\star,s}$ and Lemma 5, it can be seen that

$$\max_{1\le s\le S}\mathbf{E}\left\|X_k\Pi_s^T\left(\hat{\beta}_{k,s} - \beta_{\star,s}\right)\right\|^2 \le p_S(\sigma_n^2 + \sigma^2).$$

So with the help of Condition 5, (45) holds.

Now from (7), we have

$$R_N(w^*) = R_N^*(w^*) + o\left(R_N^*(w^*)\right) = \xi_{\star,N} + o(\xi_{\star,N}).$$

This completes the proof of Theorem 2. ∎

Proof of Theorem 3

By applying Lemmas 3 and 6, (44) and Theorem 1, we obtain

$$\begin{aligned}
&\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\overline{\beta} - \beta_\star(w^*)\right\}\right\|^2 \\
=\ &\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k\left(\hat{W}_k\right) - \beta_\star(w^*)\right\}\right\|^2 \\
\le\ &\frac{1}{K}\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\hat{\beta}_1\left(\hat{W}_1\right) - \beta_\star(w^*)\right\}\right\|^2 + \frac{K(K-1)}{K^2}\left\|\mathbf{E}\left[\Sigma_S^{1/2}\left\{\hat{\beta}_1\left(\hat{W}_1\right) - \beta_\star(w^*)\right\}\right]\right\|^2 \\
\le\ &2\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\hat{\beta}_1\left(\hat{W}_1\right) - \beta_\star\left(\hat{W}_1\right)\right\}\right\|^2 + \frac{2}{K}\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\beta_\star\left(\hat{W}_1\right) - \beta_\star(w^*)\right\}\right\|^2 \\
&+\frac{2(K-1)}{K}\left\|\mathbf{E}\left[\Sigma_S^{1/2}\left\{\beta_\star\left(\hat{W}_1\right) - \beta_\star(w^*)\right\}\right]\right\|^2 \\
\le\ &2\max_{1\le s\le S}\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\hat{\beta}_{1,s} - \beta_{\star,s}\right\}\right\|^2 + \frac{2}{K}\left(\sum_{s=1}^{S}\beta_{\star,s}^T\Pi_s\Sigma_S\Pi_s^T\beta_{\star,s}\right)\mathbf{E}\left\|\hat{W}_1 - w^*\right\|^2 \\
&+\frac{2(K-1)}{K}\left(\sum_{s=1}^{S}\beta_{\star,s}^T\Pi_s\Sigma_S\Pi_s^T\beta_{\star,s}\right)\left\|\mathbf{E}\left[\hat{W}_1 - w^*\right]\right\|^2 \\
=\ &O\left(\frac{p_S\left(\sigma_n^2+\sigma^2\right)}{n} + \frac{S^3 p_S(S+\sigma_n^2)}{K\overline{\lambda}_S^2 n} + \frac{S^5 p_S^2\left(S+\sigma_n^2\right)}{n^2\overline{\lambda}_S^4}\right),
\end{aligned}$$

and

$$
\begin{aligned}
& \mathbf{E}\left\|\Sigma_S^{1/2}\left\{\tilde{\overline{\beta}}-\beta_*\left(w^*\right)\right\}\right\|^2 \\
= & \mathbf{E}\left\|\Sigma_S^{1/2}\left\{\sum_{s=1}^S \overline{w}_s \Pi_s^T \tilde{\beta}_s-\sum_{s=1}^S w_s^* \Pi_s^T \beta_{\star,s}\right\}\right\|^2 \\
= & \mathbf{E}\left\|\Sigma_S^{1/2}\left\{\sum_{s=1}^S \overline{w}_s \Pi_s^T\left(\tilde{\beta}_s-\beta_{\star,s}\right)+\sum_{s=1}^S\left(\overline{w}_s-w_s^*\right) \Pi_s^T \beta_{\star,s}\right\}\right\|^2 \\
\leq & 2\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\sum_{s=1}^S \overline{w}_s \Pi_s^T\left(\tilde{\beta}_s-\beta_{\star,s}\right)\right\}\right\|^2+2\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\sum_{s=1}^S\left(\overline{w}_s-w_s^*\right) \Pi_s^T \beta_{\star,s}\right\}\right\|^2 \\
\leq & 2\max_{1\leq s\leq S}\left(\frac{1}{K}\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right\|^2+\left\|\mathbf{E}\left[\Sigma_S^{1/2}\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right]\right\|^2\right) \\
& +2\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\sum_{s=1}^S\left(\overline{w}_s-w_s^*\right) \Pi_s^T \beta_{\star,s}\right\}\right\|^2 \\
= & O\left(m_S^2\right)+O\left(\frac{S^3 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2}+\frac{p_S\left(\sigma_n^2+\sigma^2\right)}{Kn}\right)+O\left(\frac{S^5 p_S^2(S+\sigma_n^2)}{n^2\overline{\lambda}_S^4}\right).
\end{aligned}
$$

With Condition 2 and the fact that the eigenvalues of a matrix are not greater than its maximum column sum, we have $\overline{\lambda}_S^2 \leq 4S^2 C_b$, which leads to that

$$
O\left(\frac{S^3 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2}+\frac{p_S\left(\sigma_n^2+\sigma^2\right)}{Kn}\right)=O\left(\frac{S^3 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2}\right).
$$

Thus, Theorem 3 is proved. ∎

Proof of Theorem 4

By Theorem 3 and Condition 5, we have

$$
\begin{aligned}
& \mathbf{E}\left(\overline{\mu}_v-\mu_v\right)^2 \\
= & \xi_{\star,N}+2\mathbf{E}\left[\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}^T x_{v,S}\left(x_{v,S}^T \beta_\star\left(w^*\right)-\mu_v\right)\right] \\
& \quad +\mathbf{E}\left[\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}^T x_{v,S} x_{v,S}^T\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}\right] \\
\leq & \xi_{\star,N}\left(1+2\sqrt{\xi_{\star,N}^{-1}\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}\right\|^2}+\xi_{\star,N}^{-1}\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}\right\|^2\right) \\
= & \xi_{\star,N}\left\{1+O\left(\sqrt{\frac{p_S\left(\sigma_n^2+\sigma^2\right)}{n\xi_{\star,N}}+\frac{S^3 p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2\xi_{\star,N}}+\frac{S^5 p_S^2(S+\sigma_n^2)}{n^2\overline{\lambda}_S^4\xi_{\star,N}}}\right)\right\}^2,
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathbf{E}\left(\widetilde{\overline{\mu}}_v-\mu_v\right)^2 \\
\leq\ & \xi_{\star,N}\left(1+2\sqrt{\xi_{\star,N}^{-1}\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\widetilde{\overline{\beta}}-\beta_\star\left(w^*\right)\right\}\right\|^2}+\xi_{\star,N}^{-1}\mathbf{E}\left\|\Sigma_S^{1/2}\left\{\widetilde{\overline{\beta}}-\beta_\star\left(w^*\right)\right\}\right\|^2\right) \\
=\ & \xi_{\star,N}\left\{1+O\left(\sqrt{\frac{m_S^2}{\xi_{\star,N}}+\frac{S^3p_S(S+\sigma_n^2)}{Kn\overline{\lambda}_S^2\xi_{\star,N}}+\frac{S^5p_S^2(S+\sigma_n^2)}{n^2\overline{\lambda}_S^4\xi_{\star,N}}}\right)\right\}^2.
\end{aligned}
$$

Hence, Theorem 4 holds. ■

Proof of Theorem 5

We first show (9). By Lemmas 4 and 5, Theorem 1, and noting that $\lambda_n=\overline{\lambda}_S+o(\overline{\lambda}_S)$, it is seen that

$$
\begin{aligned}
& \mathbf{E}\left\|X_k\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
=\ & \frac{1}{K^2}\mathbf{E}\left\|X_k\sum_{j=1}^K\left\{\hat{\beta}_j\left(\hat{W}_j\right)-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
\leq\ & \frac{1}{K}\sum_{j=1}^K\mathbf{E}\left\|X_k\left\{\hat{\beta}_j\left(\hat{W}_j\right)-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
=\ & \frac{1}{K}\mathbf{E}\left\|X_k\left\{\hat{\beta}_k\left(\hat{W}_k\right)-\beta_\star\left(w^*\right)\right\}\right\|^2+\frac{1}{K}\sum_{j\neq k}\mathbf{E}\left\|X_k\left\{\hat{\beta}_j\left(\hat{W}_j\right)-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
=\ & \frac{1}{K}\mathbf{E}\left\|X_k\left\{\hat{\beta}_k\left(\hat{W}_k\right)-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
& +\frac{K-1}{K}\mathbf{E}\left\|n^{1/2}\Sigma_S^{1/2}\left\{\hat{\beta}_j\left(\hat{W}_j\right)-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
\leq\ & \frac{2}{K}\mathbf{E}\left\|X_k\left\{\hat{\beta}_k\left(\hat{W}_k\right)-\beta_\star\left(\hat{W}_k\right)\right\}\right\|^2+\frac{2}{K}\mathbf{E}\left\|X_k\left\{\beta_\star\left(\hat{W}_k\right)-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
& +\frac{2(K-1)}{K}\mathbf{E}\left\|n^{1/2}\Sigma_S^{1/2}\left\{\hat{\beta}_j\left(\hat{W}_j\right)-\beta_\star\left(\hat{W}_j\right)\right\}\right\|^2 \\
& +\frac{2(K-1)}{K}\mathbf{E}\left\|n^{1/2}\Sigma_S^{1/2}\left\{\beta_\star\left(\hat{W}_j\right)-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
\leq\ & \frac{2}{K}\max_{1\leq s\leq S}\mathbf{E}\left\|X_{k,s}\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right\|^2+\frac{2S}{K}\mathbf{E}\left\{\sum_{s=1}^S\left\|X_{k,s}\beta_{\star,s}\right\|^2\left(\hat{w}_{k,s}-w_s^*\right)^2\right\} \\
& +\frac{2n(K-1)}{K}\max_{1\leq s\leq S}\mathbf{E}\left\|\Sigma_S^{1/2}\Pi_s^T\left\{\hat{\beta}_{k,s}-\beta_{\star,s}\right\}\right\|^2 \\
& +\frac{2n(K-1)}{K}\mathbf{E}\left(\left\{\sum_{s=1}^S\beta_{\star,s}^T\Pi_s\Sigma_S\Pi_s^T\beta_{\star,s}\right\}\left\|\hat{W}_k-w^*\right\|^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&= O\left(\frac{p_S\left(\sigma_n^2+\sigma^2\right)}{K}\right)+\frac{2Sn}{K}\mathbf{E}\left[\max_{1\le s\le S}\left(x_{(k,1)}^T\Pi_s^T\beta_{\star,s}\right)^2\left\|\hat{W}_k-w^*\right\|^2\right] \\
&\quad +O\left(p_S\left(\sigma_n^2+\sigma^2\right)\right)+O\left(S^3p_S\lambda_n^{-2}\left(S+\sigma_n^2\right)\right), \\
&= O\left(S^{\eta+4/(\eta+2)}nK^{-1}\left(S^2p_Sn^{-1}\overline{\lambda}_S^{-2}\left(S+\sigma_n^2\right)\right)^{\eta/(\eta+2)}+S^3p_S\overline{\lambda}_S^{-2}\left(S+\sigma_n^2\right)\right),
\end{aligned}
$$

where $x_{(k,1)} = X_k^T\epsilon$ with $\epsilon = (1,0,\ldots,0)^T$ being an n dimensional column vector. Therefore, we obtain

$$
\begin{aligned}
&\frac{1}{N}\sum_{k=1}^K\mathbf{E}\left\|X_k\overline{\beta}-\mu_k\right\|^2 \\
=\ &\xi_{\star,N}+\frac{2}{N}\sum_{k=1}^K\mathbf{E}\left[\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}^TX_k^T\left(X_k\beta_\star\left(w^*\right)-\mu_k\right)\right] \\
&+\frac{1}{N}\sum_{k=1}^K\mathbf{E}\left\|X_k\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
\le\ &\xi_{\star,N}\left(1+2\sqrt{\frac{\sum_{k=1}^K\mathbf{E}\left\|X_k\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}\right\|^2}{N\xi_{\star,N}}}+\frac{\sum_{k=1}^K\mathbf{E}\left\|X_k\left\{\overline{\beta}-\beta_\star\left(w^*\right)\right\}\right\|^2}{N\xi_{\star,N}}\right) \\
\le\ &\xi_{\star,N}\left\{1+O\left(\sqrt{\xi_{\star,N}^{-1}\cdot\left(\frac{S^3p_S\left(S+\sigma_n^2\right)}{n\overline{\lambda}_S^2}+\frac{S^{\frac{\eta+4}{\eta+2}}}{K}\left(\frac{S^2p_S\left(S+\sigma_n^2\right)}{n\overline{\lambda}_S^2}\right)^{\frac{\eta}{\eta+2}}\right)}\right)\right\}^2.
\end{aligned}
$$

This completes the proof of (9). Similarly, to prove (10), we can derive

$$
\begin{aligned}
&\mathbf{E}\left\|X_k\left\{\overline{\tilde{\beta}}-\beta_\star\left(w^*\right)\right\}\right\|^2 \\
\le\ &2\mathbf{E}\left\{\left\|X_k\left\{\sum_{s=1}^S\overline{w}_s\Pi_s^T\left(\tilde{\beta}_s-\beta_{\star,s}\right)\right\}\right\|^2+\left\|X_k\left\{\sum_{s=1}^S\left(\overline{w}_s-w_s^*\right)\Pi_s^T\beta_{\star,s}\right\}\right\|^2\right\} \\
=\ &2\mathbf{E}\left\|X_k\left\{\sum_{s=1}^S\left(\frac{1}{K}\sum_{k=1}^K\hat{w}_{k,s}\right)\Pi_s^T\left(\frac{1}{K}\sum_{k=1}^K\hat{\beta}_{k,s}-\beta_{\star,s}\right)\right\}\right\|^2 \\
&+2\mathbf{E}\left\|X_k\left\{\sum_{s=1}^S\left(\frac{1}{K}\sum_{k=1}^K\hat{w}_{k,s}-w_s^*\right)\Pi_s^T\beta_{\star,s}\right\}\right\|^2 \\
\le\ &2\max_{1\le s\le S}\mathbf{E}\left\|X_k\left\{\Pi_s^T\left(\frac{1}{K}\sum_{k=1}^K\hat{\beta}_{k,s}-\beta_{\star,s}\right)\right\}\right\|^2+2nS\mathbf{E}\left[\max_{1\le s\le S}\left(x_{(k,1)}^T\Pi_s^T\beta_{\star,s}\right)^2\left\|\overline{w}-w^*\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & 2\max_{1\leq s\leq S}\left\{\frac{1}{K^2}\sum_{q\neq k}\mathbf{E}\left\|X_k\left\{\Pi_s^T\left(\hat{\beta}_{q,s}-\beta_{\star,s}\right)\right\}\right\|^2+\frac{1}{K^2}\mathbf{E}\left\|X_k\left\{\Pi_s^T\left(\hat{\beta}_{k,s}-\beta_{\star,s}\right)\right\}\right\|^2\right\} \\
& +2\max_{1\leq s\leq S}\left\|\mathbf{E}\left[X_k\left\{\Pi_s^T\left(\hat{\beta}_{k,s}-\beta_{\star,s}\right)\right\}\right]\right\|^2+O\left(S^{\frac{\eta+4}{\eta+2}}n\left(\mathbf{E}\left\|\overline{w}-w^*\right\|^2\right)^{\frac{\eta}{\eta+2}}\right) \\
\leq\ & O\left(\frac{p_S\left(\sigma^2+\sigma_n^2\right)}{K}\right)+O(n\overline{m}_S^2)+O\left(S^{\frac{\eta+4}{\eta+2}}n\left(\frac{S^2p_S\left(S+\sigma_n^2\right)}{Kn\lambda_S^2}+\frac{S^4p_S^2\left(S+\sigma_n^2\right)}{n^2\lambda_S^4}\right)^{\frac{\eta}{\eta+2}}\right) \\
=\ & O\left(n\overline{m}_S^2+\frac{p_S\left(\sigma^2+\sigma_n^2\right)}{K}\right)+O\left(S^{\frac{\eta+4}{\eta+2}}n\left(\frac{S^2p_S\left(S+\sigma_n^2\right)}{Kn\overline{\lambda}_S^2}+\frac{S^4p_S^2\left(S+\sigma_n^2\right)}{n^2\overline{\lambda}_S^4}\right)^{\frac{\eta}{\eta+2}}\right) \\
=\ & O\left(n\overline{m}_S^2\right)+O\left(S^{\frac{\eta+4}{\eta+2}}n\left(\frac{S^2p_S\left(S+\sigma_n^2\right)}{Kn\overline{\lambda}_S^2}+\frac{S^4p_S^2\left(S+\sigma_n^2\right)}{n^2\overline{\lambda}_S^4}\right)^{\frac{\eta}{\eta+2}}\right).
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
& \frac{1}{N}\sum_{k=1}^{K}\mathbf{E}\left\|X_k\overline{\tilde{\beta}}-\mu_k\right\|^2 \\
=\ & \xi_{\star,N}+\frac{2}{N}\sum_{k=1}^{K}\mathbf{E}\left[\left\{\overline{\tilde{\beta}}-\beta_\star\left(w^*\right)\right\}^TX_k^T\left(X_k\beta_\star\left(w^*\right)-\mu_k\right)\right] \\
& +\frac{1}{N}\sum_{k=1}^{K}\mathbf{E}\left[\left\{\overline{\tilde{\beta}}-\beta_\star\left(w^*\right)\right\}^TX_k^TX_k\left\{\overline{\tilde{\beta}}-\beta_\star\left(w^*\right)\right\}\right] \\
\leq\ & \xi_{\star,N}\left(1+2\sqrt{\frac{\sum_{k=1}^K\mathbf{E}\left\|X_k\left\{\overline{\tilde{\beta}}-\beta_\star\left(w^*\right)\right\}\right\|^2}{N\xi_{\star,N}}}+\frac{\sum_{k=1}^K\mathbf{E}\left\|X_k\left\{\overline{\tilde{\beta}}-\beta_\star\left(w^*\right)\right\}\right\|^2}{N\xi_{\star,N}}\right) \\
\leq\ & \xi_{\star,N}\left\{1+O\left(\xi_{\star,N}^{-\frac{1}{2}}\cdot\left(\overline{m}_S^2+S^{\frac{\eta+4}{\eta+2}}\left(\frac{S^2p_S\left(S+\sigma_n^2\right)}{Kn\overline{\lambda}_S^2}+\frac{S^4p_S^2\left(S+\sigma_n^2\right)}{n^2\overline{\lambda}_S^4}\right)^{\frac{\eta}{\eta+2}}\right)^{-\frac{1}{2}}\right)\right\},
\end{aligned}
$$

that is, (10) holds. Thus, we conclude the proof of Theorem 5. ■

Proof of Theorem 7

We first consider (15). Without loss of generality, we assume that Θ is compact, then there is a $\theta^* \in \Theta$ such that

$$\theta^* \triangleq \operatorname*{argmax}_{\theta\in\Theta}\overline{MSE},$$

and the corresponding $\overline{MSE}$, $\bar{\lambda}_S$, and λ_n are denoted by $\overline{MSE}_{\theta^*}$, $\bar{\lambda}_S(\theta^*)$, and $\lambda_n(\theta^*)$, respectively. For the model with parameter θ^*, by $\sigma_n^2 = o(n)$ and the definition of Θ, it is easy to check that Conditions 1 and 4 hold, and by C_r inequality, $\sup_{j\geq 1}\mathbf{E}|x_{k,i,j}|^q<\infty$ and $\|\theta^*\|\leq\varepsilon_3$ can deduce that Condition 2 holds. Thus, Conditions 1 - 4 and 6 are all true for the model with parameter θ^*, and so by Theorem 5,

$$\frac{\overline{MSE}_{\theta^*}}{\inf_{w\in Q}R_N^\star(w)}=1+O\left(\frac{\sigma_n^2}{n\bar{\lambda}_S^2(\theta^*)}+\frac{1}{K}\left(\frac{\sigma_n^2}{n\bar{\lambda}_S^2(\theta^*)}\right)^{\frac{q-2}{q}}\right). \tag{46}$$

Noting that $\lambda_n(\theta^*)=\bar{\lambda}_S(\theta^*)+o(\bar{\lambda}_S(\theta^*))$, this together with (46) leads to (15). In a similar manner, we can show (16).

Next, we focus on (17) and (18). By Lemma 6, it can be verified that

$$\begin{aligned}&\sup_{W_1\in Q,W_2\in Q,\ldots,W_K\in Q}\sup_{\theta\in\Theta}\left(\overline{Mse}\left(W_1,W_2,\ldots,W_K\right)-\frac{1}{N}\sum_{k=1}^K\mathbf{E}\left\|X_k\frac{1}{K}\sum_{k=1}^K\beta_\star(W_k)-\boldsymbol{\mu}_k\right\|^2\right)\\&=O\left(\frac{\sigma_n^2}{n}\right)\end{aligned}$$

and

$$\begin{aligned}&\sup_{W_1\in Q,W_2\in Q,\ldots,W_K\in Q}\sup_{\theta\in\Theta}\left(\widetilde{Mse}\left(W_1,W_2,\ldots,W_K\right)-\frac{1}{N}\sum_{k=1}^K\mathbf{E}\left\|X_k\frac{1}{K}\sum_{k=1}^K\beta_\star(W_k)-\boldsymbol{\mu}_k\right\|^2\right)\\&=O\left(\frac{\sigma_n^2}{n}\right).\end{aligned}$$

Note that

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^K\mathbf{E}\left\|X_k\frac{1}{K}\sum_{k=1}^K\beta_\star(W_k)-\boldsymbol{\mu}_k\right\|^2&=\frac{1}{N}\sum_{k=1}^K\mathbf{E}\left\|X_k\sum_{s=1}^S\left(\frac{1}{K}\sum_{k=1}^K w_{k,s}\right)\Pi_s^T\beta_{\star,s}-\boldsymbol{\mu}_k\right\|^2\\&=\frac{1}{N}\sum_{k=1}^K\mathbf{E}\left\|X_k\beta_\star\left(\frac{1}{K}\sum_{k=1}^K W_k\right)-\boldsymbol{\mu}_k\right\|^2.\end{aligned}$$

Then

$$\begin{aligned}&\inf_{W_1\in Q,W_2\in Q,\ldots,W_K\in Q}\sup_{\theta\in\Theta}\overline{Mse}\left(W_1,W_2,\ldots,W_K\right)\\
\leq\;&\inf_{W_1\in Q,W_2\in Q,\ldots,W_K\in Q}\sup_{\theta\in\Theta}\frac{1}{N}\sum_{k=1}^K\mathbf{E}\left\|X_k\frac{1}{K}\sum_{k=1}^K\beta_\star(W_k)-\boldsymbol{\mu}_k\right\|^2+O\left(\frac{\sigma_n^2}{n}\right)\\
=\;&\inf_{W_1\in Q,W_2\in Q,\ldots,W_K\in Q}\sup_{\theta\in\Theta}\frac{1}{N}\sum_{k=1}^K\mathbf{E}\left\|X_k\beta_\star\left(\frac{1}{K}\sum_{k=1}^K W_k\right)-\boldsymbol{\mu}_k\right\|^2+O\left(\frac{\sigma_n^2}{n}\right)\\
=\;&\inf_{w\in Q}\sup_{\theta\in\Theta}R_N^\star(w)+O\left(\frac{\sigma_n^2}{n}\right)\\
\leq\;&\left(1+O\left(\frac{\sigma_n^2}{n}\right)\right)\inf_{w\in Q}\sup_{\theta\in\Theta}R_N^\star(w),\end{aligned}$$

where the last inequality is obtained from the definition of S_2, which confirms (17). Similarly, we can imitate the above process to prove (18). This completes the proof of Theorem 7. ■

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Tomohiro Ando and Ker-Chau Li. A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109(505):254–265, 2014.

Maria Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity and privacy. *Conference on Learning Theory*, 26(1):1–22, 2012.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3): 1352–1382, 2018.

James O Berger and Dipak K Dey. Combining coordinates in simultaneous estimation of normal means. *Journal of Statistical Planning and Inference*, 8(2):143–160, 1983.

Steven T Buckland, Kenneth P Burnham, and Nicole H Augustin. Model selection: an integral part of inference. *Biometrics*, 53(2):603–618, 1997.

Ali Charkhi, Gerda Claeskens, and Bruce E Hansen. Minimum mean squared error model averaging in likelihood models. *Statistica Sinica*, 26(2):809–840, 2016.

Jia Chen, Degui Li, Oliver Linton, and Zudi Lu. Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, 113(522):919–932, 2018.

Xi Chen, Weidong Liu, and Yichen Zhang. Quantile regression under memory constraint. *The Annals of Statistics*, 47(6):3244–3273, 2019.

Xi Chen, Weidong Liu, Xiaojun Mao, and Zhuoyi Yang. Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research*, 21 (182):1–43, 2020.

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

Gerda Claeskens and Raymond J Carroll. An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika*, 94(2):249–265, 2007.

Dipak K Dey and James O Berger. On truncation of shrinkage estimators in simultaneous estimation of normal means. *Journal of the American Statistical Association*, 78(384): 865–869, 1983.

Francis X Diebold and Robert S Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 20(1):134–144, 2002.

Bradley Efron and Carl Morris. Combining possibly related estimation problems. *Journal of the Royal Statistical Society: Series B*, 35(3):379–402, 1973.

Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National Science Review*, 1(2):293–314, 2014.

Fang Fang, Xiangju Yin, and Qiang Zhang. Divide and conquer algorithms for model averaging with massive data. *Journal of Systems Science and Mathematics*, 38(7):764–776, 2018.

Yan Gao, Xinyu Zhang, Shouyang Wang, and Guohua Zou. Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics*, 192(1):139–151, 2016.

Yan Gao, Xinyu Zhang, Shouyang Wang, Terence Tai-leung Chong, and Guohua Zou. Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics*, 71(2):275–306, 2019.

Dan Garber, Ohad Shamir, and Nathan Srebro. Communication-efficient algorithms for distributed stochastic principal component analysis. *Proceedings of the 34th International Conference on Machine Learning*, 70:1203–1212, 2017.

Edward I George. Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, 81(394):437–445, 1986a.

Edward I George. Minimax multiple shrinkage estimation. *The Annals of Statistics*, 14(1): 188–205, 1986b.

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

Bruce E Hansen. Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics*, 5(3):495–530, 2014.

Bruce E Hansen and Jeffrey S Racine. Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46, 2012.

Nils L Hjort and Gerda Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

Degui Li, Oliver Linton, and Zudi Lu. A flexible semiparametric forecasting model for time series. *Journal of Econometrics*, 187(1):345–357, 2015.

Hua Liang, Guohua Zou, Alan TK Wan, and Xinyu Zhang. Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*, 106(495):1053–1066, 2011.

Jun Liao and Guohua Zou. Corrected mallows criterion for model averaging. *Computational Statistics and Data Analysis*, 144:106902, 2020.

Jun Liao, Xianpeng Zong, Xinyu Zhang, and Guohua Zou. Model averaging based on leave-subject-out cross-validation for vector autoregressions. *Journal of Econometrics*, 209(1): 35–60, 2019.

Nan Lin and Ruibin Xi. Aggregated estimating equation estimation. *Statistics and Its Interface*, 4(1):73–83, 2011.

Chu-An Liu. Distribution theory of the least squares averaging estimator. *Journal of Econometrics*, 186(1):142–159, 2015.

Qingfeng Liu and Ryo Okui. Heteroskedasticity-robust C_p model averaging. *The Econometrics Journal*, 16(3):463–472, 2013.

Karim Lounici, Massimiliano Pontil, Sara van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4): 2164–2204, 2011.

Xun Lu and Liangjun Su. Jackknife model averaging for quantile regressions. *Journal of Econometrics*, 188(1):40–58, 2015.

Mauro Maggioni and James M Murphy. Learning by unsupervised nonlinear diffusion. *Journal of Machine Learning Research*, 20(160):1–56, 2019.

Takeru Matsuda, Masatoshi Uehara, and Aapo Hyvarinen. Information criteria for non-normalized models. *Journal of Machine Learning Research*, 22(158):1–33, 2021.

Beniamino Murgante Sanjay Misra, Ana Maria AC Rocha Carmelo Torre, Jorge Gustavo Rocha Maria Irene Falcão, David Taniar Bernady O Apduhan, and Osvaldo Gervasi. *Computational Science and Its Applications–ICCSA 2019*. Springer, 2019.

Mathilde Mougeot, Dominique Picard, and Karine Tribouley. Grouping strategies and thresholding for high dimensional linear models. *Journal of Statistical Planning and Inference*, 143(9):1417–1438, 2013.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

Behrooz Safarinejadian, Mohammad B. Menhaj, and Mehdi Karrari. Distributed unsupervised gaussian mixture learning for density estimation in sensor networks. *IEEE Transactions on Instrumentation and Measurement*, 59(9):2250–2260, 2010.

Michael Schomaker and Christian Heumann. When and when not to use optimal model averaging. *Statistical Papers*, 61(5):2221–2240, 2020.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

Ela Sienkiewicz, Dong Song, F. Jay Breidt, and Haonan Wang. Sparse functional dynamical models–a big data approach. *Journal of Computational and Graphical Statistics*, 26(2): 319–329, 2017.

Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–27, 2014.

Alan T.K Wan, Xinyu Zhang, and Guohua Zou. Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156(2):277–283, 2010.

Chun Wang, Ming-Hui Chen, Elizabeth Schifano, Jing Wu, and Jun Yan. Statistical methods and computing for big data. *Statistics and Its Interface*, 9(4):399–411, 2016.

Chun Wang, Ming-Hui Chen, Jing Wu, Jun Yan, Yuping Zhang, and Elizabeth Schifano. Online updating method with new variables for big data streams. *Canadian Journal of Statistics*, 46(1):123–146, 2018.

Xiaozhou Wang, Zhuoyi Yang, Xi Chen, and Weidong Liu. Distributed inference for linear support vector machine. *Journal of Machine Learning Research*, 20(113):1–41, 2019.

Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12):3571–3594, 2010.

Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14(24):867–897, 2013.

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

Ruibin Xi, Nan Lin, and Yixin Chen. Compression and aggregation for logistic regression analysis in data cubes. *IEEE Transactions on Knowledge and Data Engineering*, 21(4): 479–492, 2009.

Ganggang Xu, Suojin Wang, and Jianhua Z Huang. Focused information criterion and model averaging based on weighted composite quantile regression. *Scandinavian Journal of Statistics*, 41(2):365–381, 2014.

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

Yuhong Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588, 2001.

Zheng Yuan and Yuhong Yang. Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100(472):1202–1214, 2005.

Haili Zhang and Guohua Zou. Cross-validation model averaging for generalized functional linear model. *Econometrics*, 8(1):7, 2020.

Xinyu Zhang. A new study on asymptotic optimality of least squares model averaging. *Econometric Theory*, 37(2):388–407, 2021.

Xinyu Zhang and Hua Liang. Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics*, 39(1):174–200, 2011.

Xinyu Zhang, Alan TK Wan, and Zhou Sherry Z. Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. *Journal of Business and Economic Statistics*, 30(1):132–142, 2012.

Xinyu Zhang, Alan TK Wan, and Guohua Zou. Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, 174(2):82–94, 2013a.

Xinyu Zhang, Guohua Zou, Hua Liang, and Raymond J Carroll. Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, 115(530):972–984, 2020.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1):3321–3363, 2013b.

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

Rong Zhu, Alan TK Wan, Xinyu Zhang, and Guohua Zou. A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association*, 114(526):882–892, 2019.

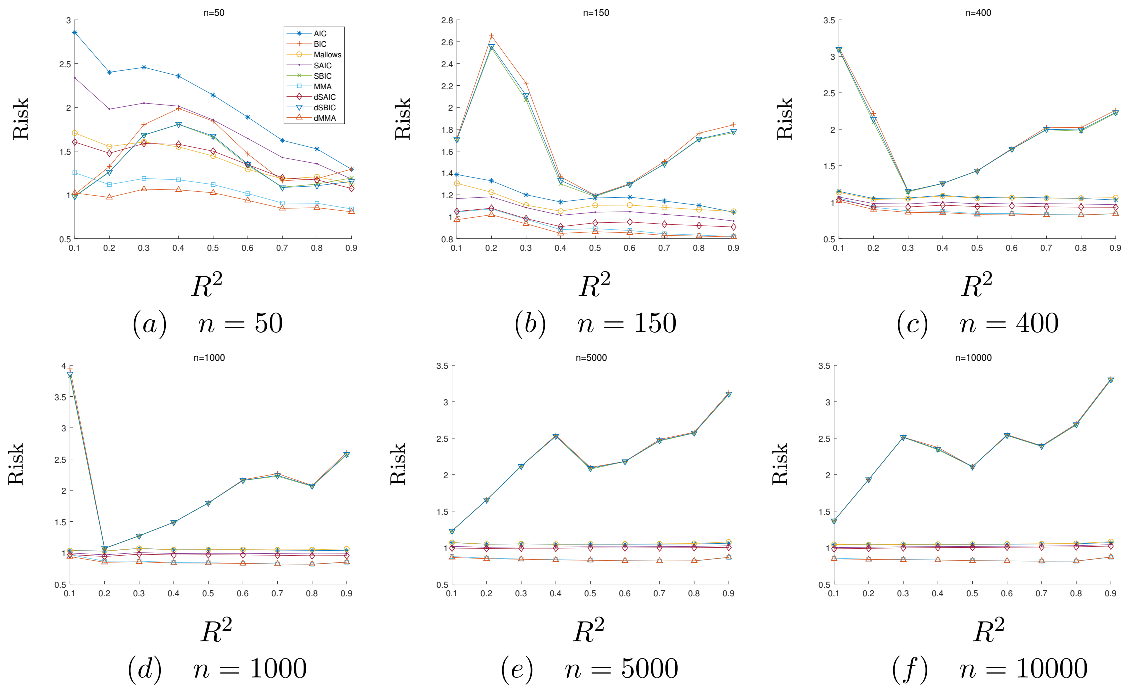


Figure 1: In-sample risk results with $\alpha = 0.5$ and $K = 2$ in Section 4.2.

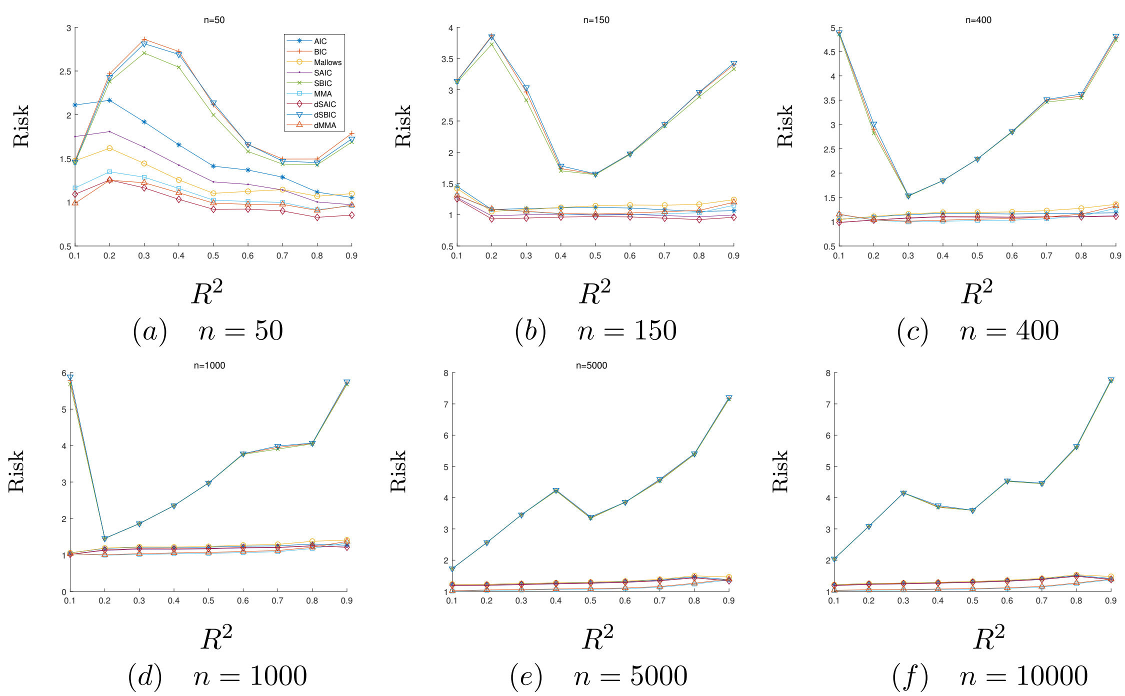
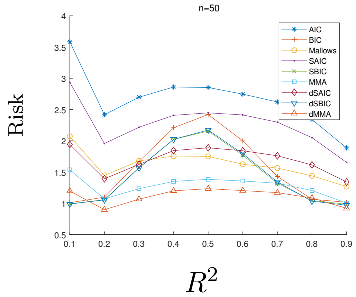
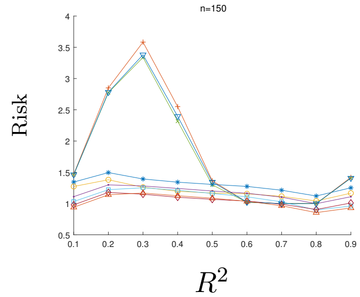


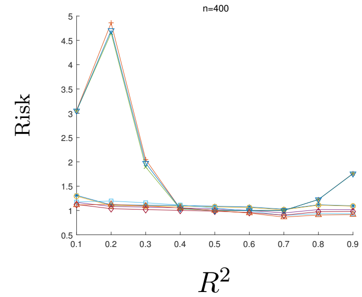
Figure 2: In-sample risk results with $\alpha = 0.5$ and $K = 5$ in Section 4.2.



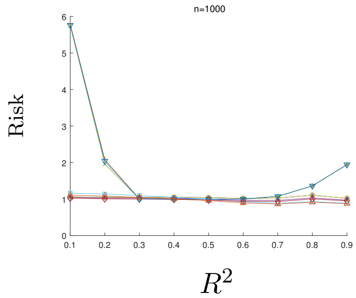
$(a)\quad n = 50$



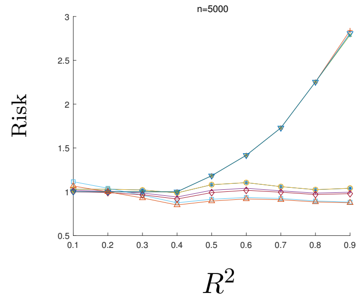
$(b)\quad n = 150$



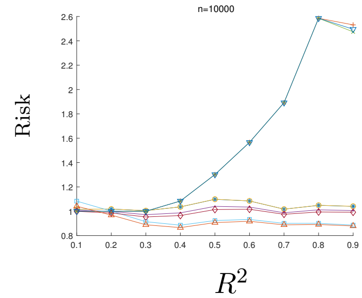
$(c)\quad n = 400$



$(d)\quad n = 1000$

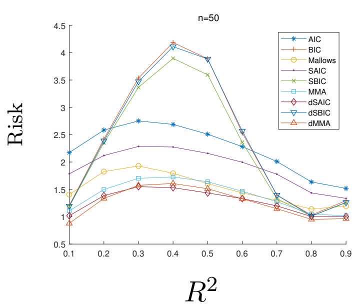


$(e)\quad n = 5000$

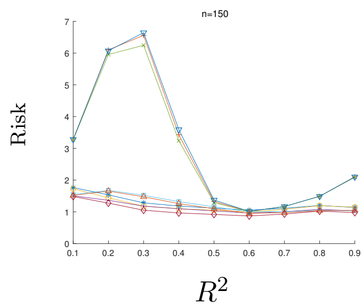


$(f)\quad n = 10000$

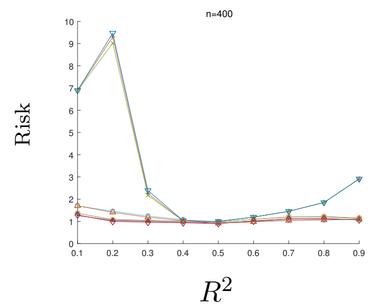
Figure 3: In-sample risk results with $\alpha = 1$ and $K = 2$ in Section 4.2.



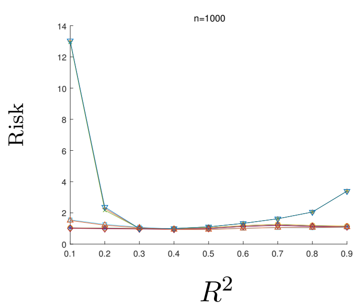
$(a)\quad n = 50$



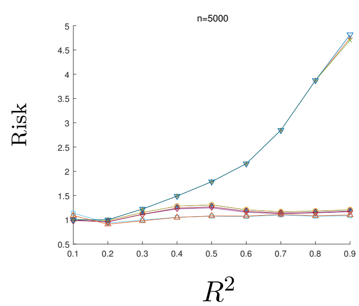
$(b)\quad n = 150$



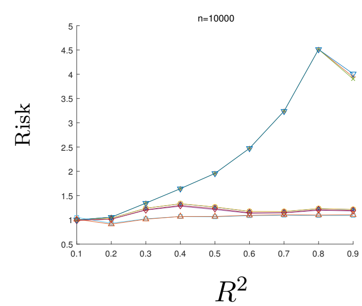
$(c)\quad n = 400$



$(d)\quad n = 1000$



$(e)\quad n = 5000$



$(f)\quad n = 10000$

Figure 4: In-sample risk results with $\alpha = 1$ and $K = 5$ in Section 4.2.

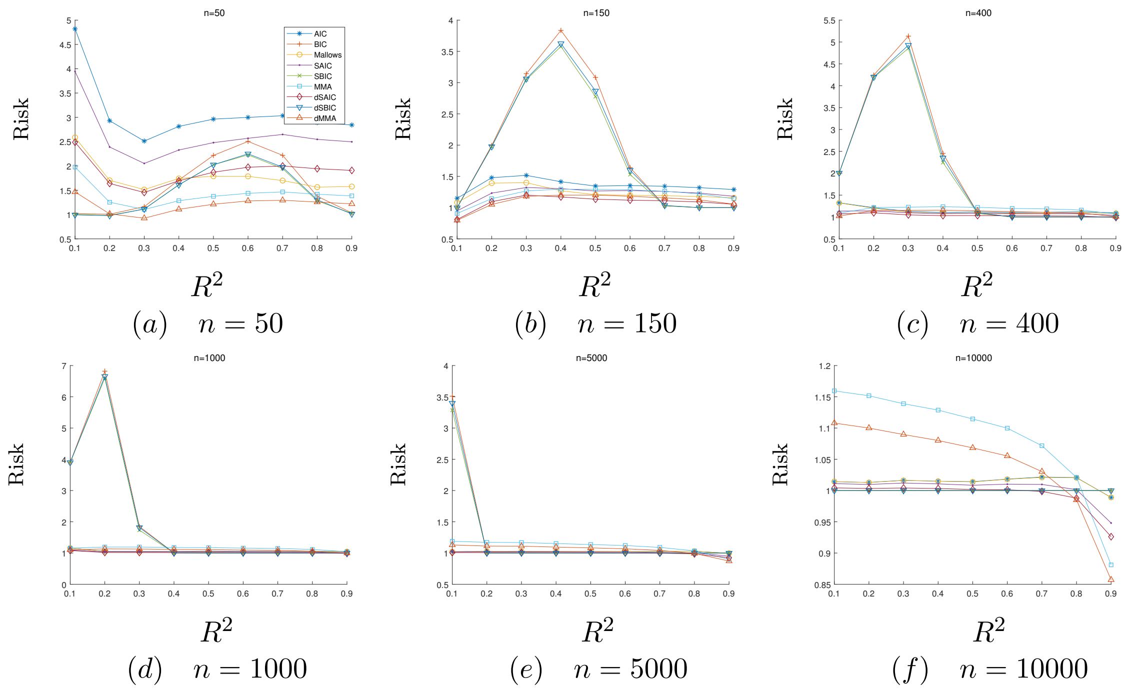


Figure 5: In-sample risk results with $\alpha = 1.5$ and $K = 2$ in Section 4.2.

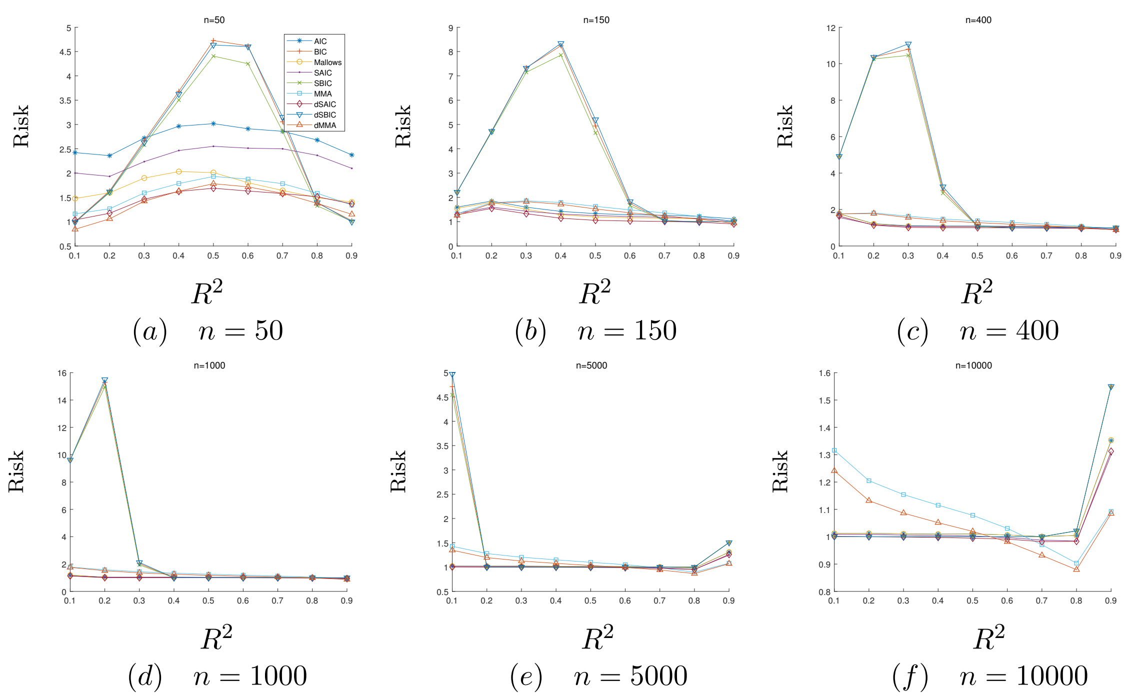
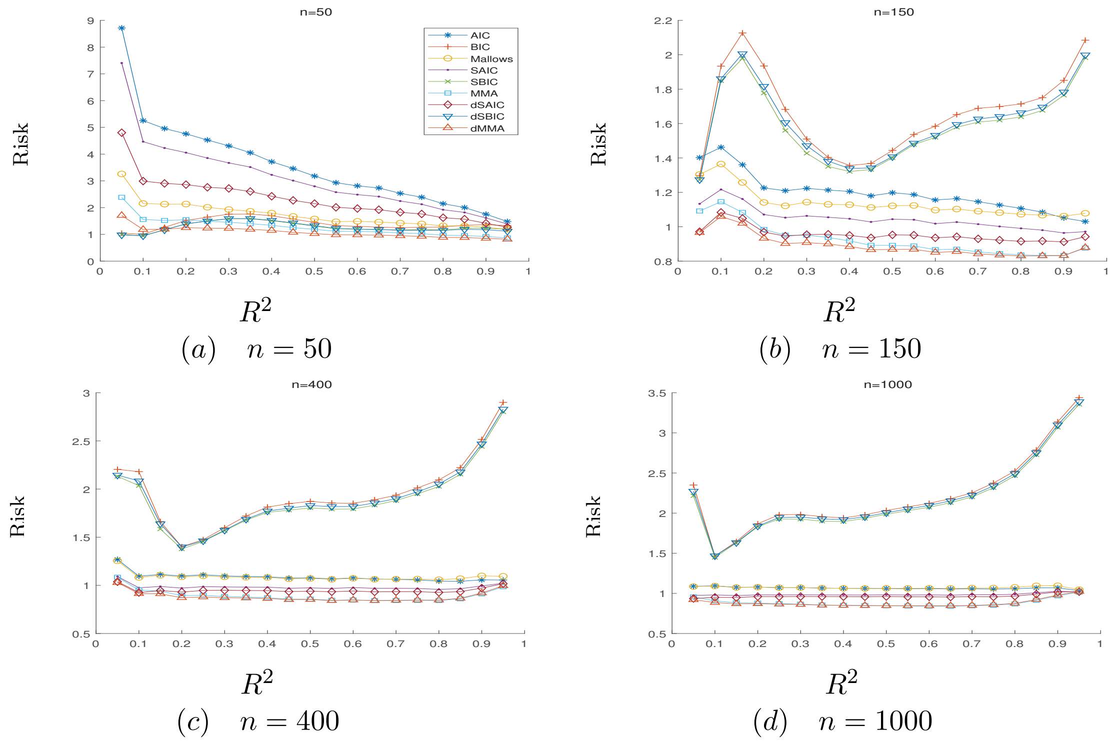


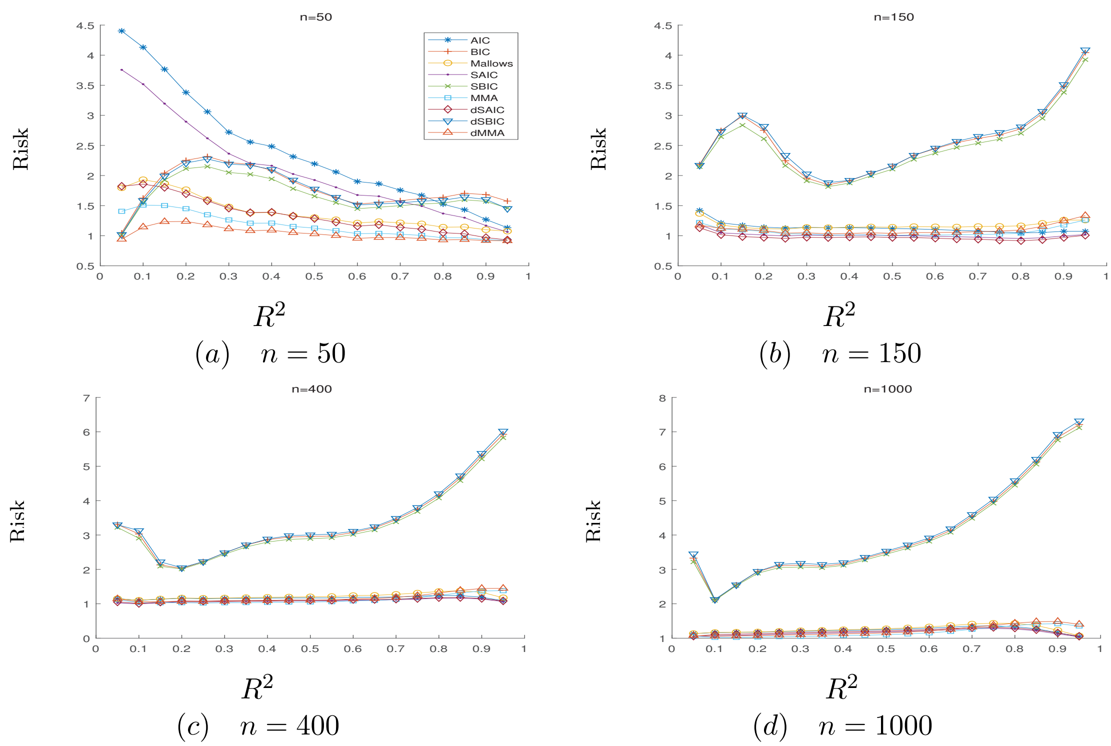
Figure 6: In-sample risk results with $\alpha = 1.5$ and $K = 5$ in Section 4.2.



$(a)\quad n = 50$ $(b)\quad n = 150$

$(c)\quad n = 400$ $(d)\quad n = 1000$

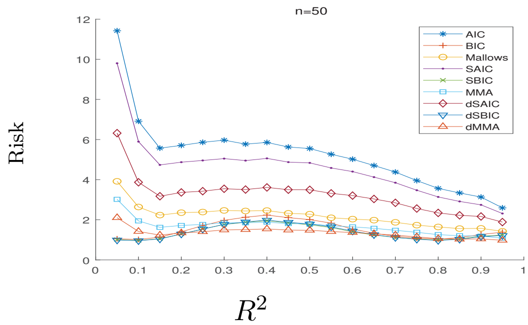
Figure 7: Out-of-sample risk results with $\alpha = 0.5$ and $K = 2$ in Section 4.3.



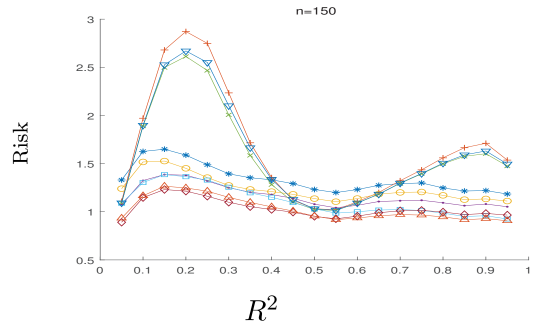
$(a)\quad n = 50$ $(b)\quad n = 150$

$(c)\quad n = 400$ $(d)\quad n = 1000$

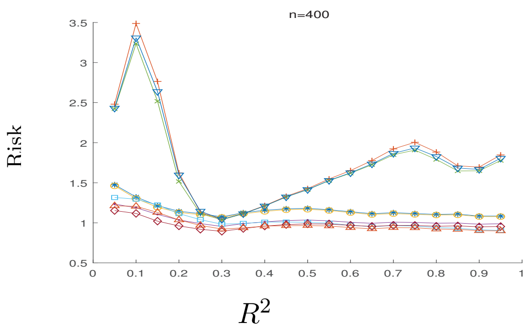
Figure 8: Out-of-sample risk results with $\alpha = 0.5$ and $K = 5$ in Section 4.3.



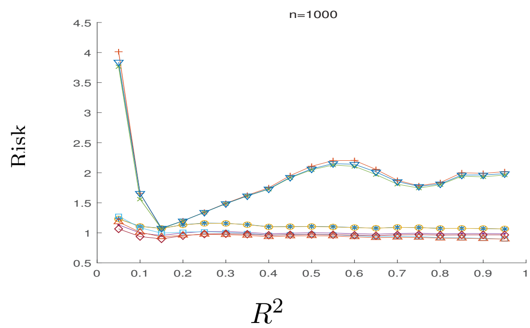
$(a)\quad n = 50$



$(b)\quad n = 150$

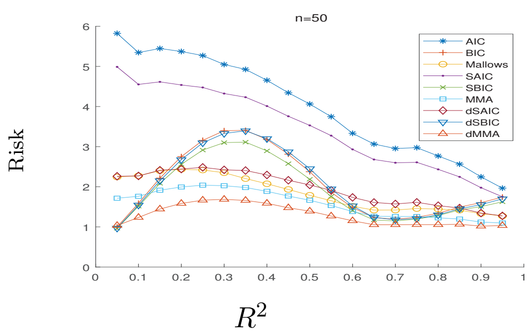


$(c)\quad n = 400$

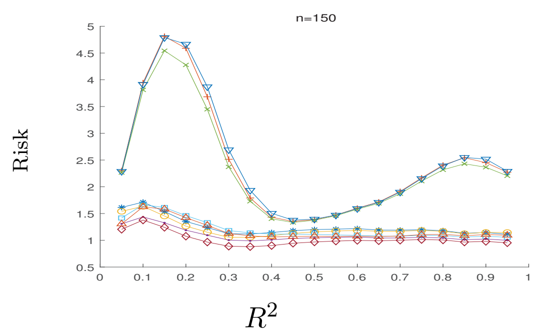


$(d)\quad n = 1000$

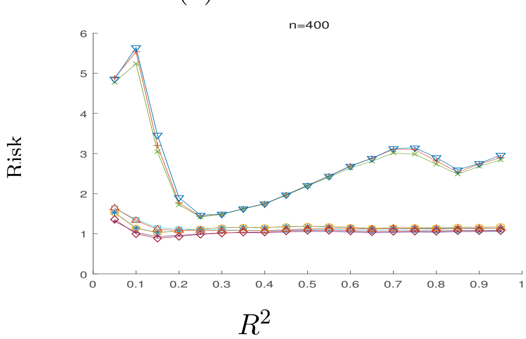
Figure 9: Out-of-sample risk results with $\alpha = 1$ and $K = 2$ in Section 4.3.



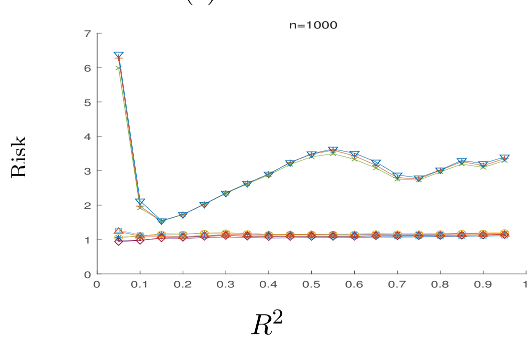
$(a)\quad n = 50$



$(b)\quad n = 150$



$(c)\quad n = 400$



$(d)\quad n = 1000$

Figure 10: Out-of-sample risk results with $\alpha = 1$ and $K = 5$ in Section 4.3.

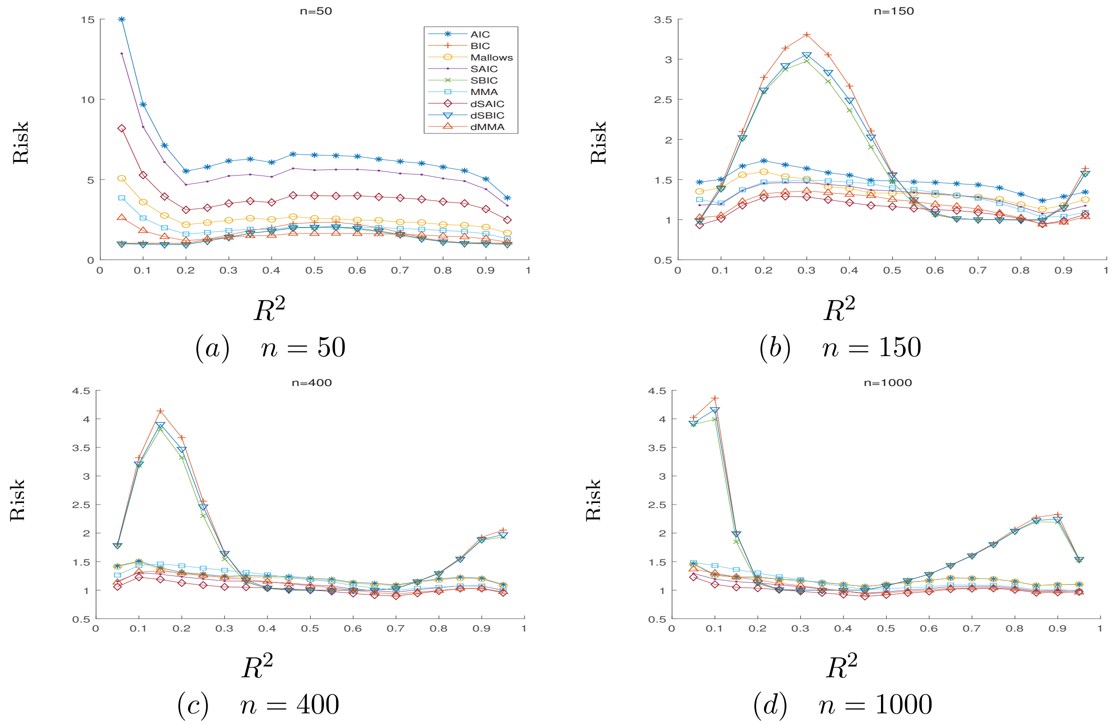


Figure 11: Out-of-sample risk results with $\alpha = 1.5$ and $K = 2$ in Section 4.3.

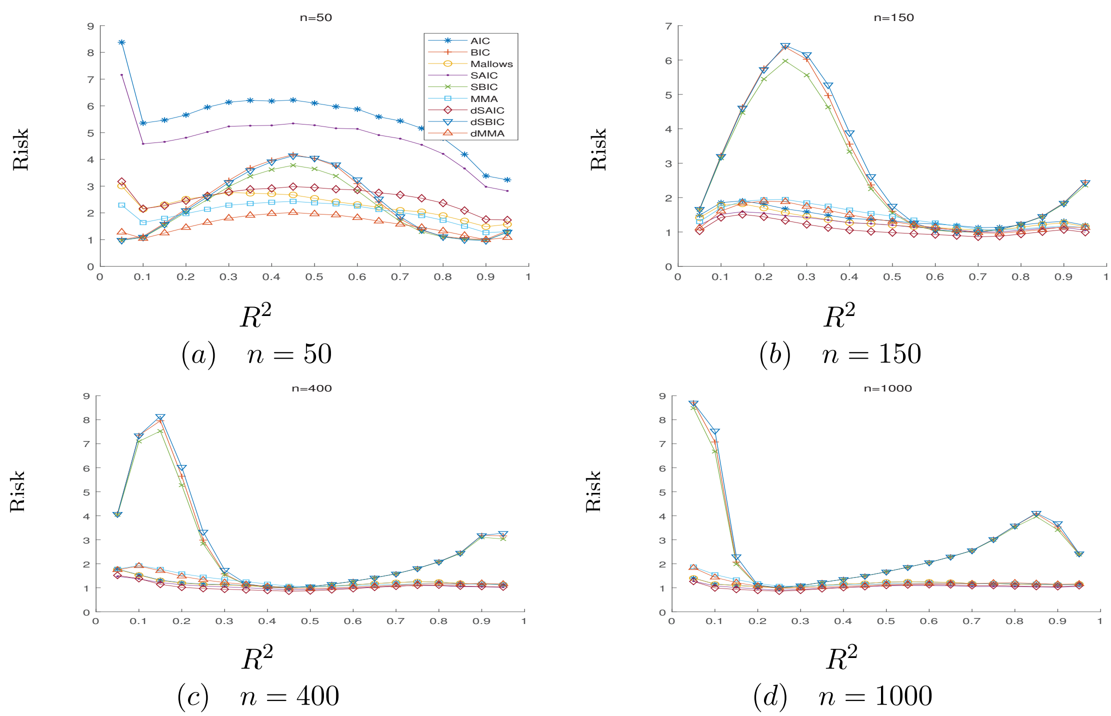


Figure 12: Out-of-sample risk results with $\alpha = 1.5$ and $K = 5$ in Section 4.3.