Journal of Machine Learning Research 25 (2024) 1-59 Submitted 7/23; Revised 10/24; Published 10/24

Stability and L_2-penalty in Model Averaging

Hengkun Zhu HKZHU@CNU.EDU.CN
Guohua Zou GHZOU@AMSS.AC.CN
School of Mathematical Sciences
Capital Normal University
Beijing 100048, China

Editor: Aryeh Kontorovich

Abstract

Model averaging has received much attention in the past two decades, which integrates available information by averaging over potential models. Although various model averaging methods have been developed, there is little literature on the theoretical properties of model averaging from the perspective of stability, and the majority of these methods constrain model weights to a simplex. The aim of this paper is to introduce stability from statistical learning theory into model averaging. Thus, we define the stability, asymptotic empirical risk minimization, generalization and consistency of model averaging, and study the relationship among them. Similar to the existing results in literature, we find that stability can ensure that the model averaging estimator has good generalization performance and consistency under reasonable conditions, where consistency means that the model averaging estimator can asymptotically minimize the mean squared prediction error. We also propose an L_2-penalty model averaging method without limiting model weights, and prove that it has stability and consistency. In order to overcome selection uncertainty of the L_2-penalty parameter, we use cross-validation to select a candidate set of L_2-penalty parameters, and then perform a weighted average of the estimators of model weights based on cross-validation errors. We demonstrate the usefulness of the proposed method with a Monte Carlo simulation and application to a prediction task on the wage1 dataset.

Keywords: Model averaging, Stability, Mean squared prediction error, L_2-penalty

1. Introduction

In practical applications, data analysts usually determine multiple models based on exploratory analysis for data and empirical knowledge to describe the relationship between variables of interest and related variables. However, how to use these models to produce good results is a more important problem. It is very common to select one model using some data-driven criteria, such as AIC (Akaike, 1973), BIC (Schwarz, 1978), C_p (Mallows, 1973) and FIC (Hjort and Claeskens, 2003). An alternative to model selection is to make a compromise across a set of competing models. Statisticians find that they can usually obtain better and more stable results by combining information from different models. This process of combining multiple model results is known as model averaging. The problem of Bayesian and frequentist model averaging (BMA and FMA) has been well studied. Fragoso et al. (2018) reviewed the relevant literature on BMA. In this paper, we focus on FMA. In the past decades, the model averaging method has been applied in various fields. Wan and

Zhang (2009) examined the applications of model averaging in tourism research. Zhang and Zou (2011) applied the model averaging method to grain production forecasting in China. Moral-Benito (2015) reviewed the literature on model averaging with special emphasis on its applications in economics. The key to FMA lies in how to select model weights. The common weight selection methods include: 1) methods based on information criteria, such as smoothed AIC and smoothed BIC in Buckland et al. (1997), and smoothed FIC (Hjort and Claeskens, 2003); 2) Mallows model averaging (MMA), proposed by Hansen (2007) (see also Wan et al., 2010), modified by Liu and Okui (2013) to make it applicable to heteroscedasticity, and improved by Liao and Zou (2020) in small sample sizes; 3) adaptive methods (Yang, 2001; Yuan and Yang, 2005); 4) OPT method (Liang et al., 2011); 5) cross validation methods, such as jackknife model averaging (JMA; Hansen and Racine, 2012; Zhang et al., 2013) and leave-subject-out cross validation model averaging (Gao et al., 2016; Liao et al., 2019).

In computational learning theory, stability is used to measure an algorithm's sensitivity to perturbation in the training set and is an important tool for analyzing generalization and learnability. Bousquet and Elisseeff (2002) introduced four notions of stability (hypothesis stability, pointwise hypothesis stability, error stability and uniform stability), and showed how to use them to derive generalization error bounds based on the empirical error and the leave-one-out error. Kutin and Niyogi (2002) introduced several weaker variants of stability and showed how they are sufficient to obtain generalization bounds for algorithms. Rakhlin et al. (2005) and Mukherjee et al. (2006) discussed the necessity of stability for learnability under the assumption that uniform convergence is equivalent to learnability. As commented by Shalev-Shwartz et al. (2010), it was recognized that a fundamental and long-standing answer about how to characterize learnability, at least for the cases of supervised classification and regression, was that learnability is equivalent to uniform convergence of the empirical risk to the population risk, and that if a problem is learnable, it is learnable via empirical risk minimization. However, Shalev-Shwartz et al. (2010) found that in the general learning setting which includes most statistical learning problems as special cases, there are non-trivial learning problems where uniform convergence does not hold, and so empirical risk minimization fails, yet these problems are learnable using alternative mechanisms. Further, instead of uniform convergence, Shalev-Shwartz et al. (2010) identified stability as the key necessary and sufficient condition for learnability.

Although various model averaging methods have been proposed, there is little literature on their theoretical properties from the perspective of stability and the majority of these methods are concerned only with whether the resultant estimator obtains a good approximation to the minimum of a given target when the model weights are constrained to a simplex. Thus, our aim in this paper is to study stability in model averaging and to answer whether the resultant estimator can get a good approximation for the minimum of target function when the model weights are unrestricted.

Our first contribution is to introduce the concept of stability from statistical learning theory into model averaging. Stability discusses how much the algorithm's output varies when the sample changes a little. Shalev-Shwartz et al. (2010) discussed the relationship among asymptotic empirical risk minimization (AERM), stability, generalization and consistency. However, the relevant conclusions cannot be directly applied to model averaging. Therefore, we explore the relevant definitions and conclusions of Shalev-Shwartz et al. (2010) under

the model averaging framework. Similar to the existing results from Bousquet and Elisseeff (2002) and Shalev-Shwartz et al. (2010), we find that stability can ensure that model averaging has good generalization performance and consistency under reasonable conditions, where consistency means that the model averaging estimator can asymptotically minimize the mean squared prediction error (MSPE).

Further, we find that for MMA and JMA, extreme weights tend to appear due to the influence of correlation between candidate models when the model weights are unrestricted. This results in poor performance of the model averaging estimator. Therefore, we should not simply remove the weight constraint and directly use the existing model averaging methods. Similar to ridge regression in Hoerl and Kennard (1970), we introduce an L_2-penalty for the weight vector in MMA and JMA. We call them Ridge-Mallows model averaging (RMMA) and Ridge-jackknife model averaging (RJMA), respectively. This is our second contribution. Like Theorem 4.3 in Hoerl and Kennard (1970), we discuss the reasonability of introducing L_2-penalty. In this regard, there are some related works in literature. Skolkova (2023) proposed the ridge model averaging estimator (RMA). In order to reduce the instability of regression estimators of candidate models caused by high correlation among covariates, RMA uses ridge regression to replace ordinary least squares. Liu (2023) proposed penalized Mallows model averaging (pMMA) to avoid over-selection of candidate models in forming a combined estimator. Unlike RMA and pMMA, we introduce an L_2-penalty for the model weights in order to reduce the instability of the model weight estimator caused by high correlation among candidate models when the model weights are unrestricted. We also prove the stability and consistency of RMMA and RJMA.

In the context of shrinkage estimation, Schomaker (2012) discussed the impact of tuning parameter selection and pointed out that the weighted average of shrinkage estimators with different tuning parameters can improve overall stability, predictive performance and standard errors of shrinkage estimators. Hence, like Schomaker (2012), we use cross-validation to select a candidate set of L_2-penalty parameters, and then perform a weighted average of the estimators of model weights based on cross-validation errors. Moreover, we provide empirical support for the usefulness of the proposed method with a Monte Carlo simulation and application to a prediction task on the wage1 dataset in which our approach outperforms MMA and JMA, as well as some commonly used model selection methods.

The remainder of this paper is organized as follows. In Section 2, we give the definitions of consistency and stability, and discuss their relationship. In Section 3, we propose RMMA and RJMA methods and prove that they are stable and consistent. Section 4 conducts the Monte Carlo simulation experiment. Section 5 applies the proposed method to a real data set. Section 6 concludes. The proofs of theorems are provided in the Appendix B.

2. Consistency and Stability for Model Averaging

We assume that $S = \{z_i = (y_i, x_i^{'})^{'} \in \mathcal{Z}, i = 1, 2, \cdots, n\}$ is a simple random sample from distribution $\mathcal{D}$, where y_i is the i-th observation of the response variable and x_i is the i-th observation of covariates. Let $z^* = (y^*, x^{*'})^{'}$ be an observation that is from distribution $\mathcal{D}$ and independent of S.

2.1 Model Averaging

In model averaging, M candidate models are selected first in order to describe the relationship between response variable and covariates. We assume that the hypothesis spaces of M candidate models are

$$\mathcal{H}_m = \{h_m(x_m^*), h_m \in \mathcal{F}_m\}, m = 1, 2, \cdots, M,$$

where x_m^* consists of some elements of x^* and $\mathcal{F}_m$ is a given function set. For example, in MMA, we take

$$\mathcal{H}_m = \{x_m^{*'}\theta_m, \theta_m \in R^{dim(x_m^*)}\}, m = 1, 2, \cdots, M$$

in order to estimate $E(y^*|x^*)$, where $dim(\cdot)$ represents the dimension of vector. For the m-th candidate model, a proper estimation method A_m is selected, and then $\hat{h}_m$, the estimator of h_m, is obtained based on S and A_m. Then, the hypothesis space of model averaging is defined as follows:

$$\mathcal{H} = \{\hat{h}(x^*, w) = H[w, \hat{h}_1(x_1^*), \hat{h}_2(x_2^*), \cdots, \hat{h}_M(x_M^*)], w \in W\},$$

where W is a given weight space and $H(\cdot)$ is a given function of weight vector and estimators of M candidate models. In MMA, we take

$$H[w, \hat{h}_1(x_1^*), \hat{h}_2(x_2^*), \cdots, \hat{h}_M(x_M^*)] = \sum_{m=1}^{M} w_m \hat{h}_m(x_m^*)$$

as the model averaging estimator of $E(y^*|x^*)$. An important problem for model averaging is the choice of model weights. Here, the estimator $\hat{w}$ of the weight vector is obtained based on S and a proper weight selection criterion $A(w)$ that makes $\hat{w}$ be optimal in a certain sense.

The selection of $\{A_m, m = 1, 2, \cdots, M\}$ and $A(w)$ is closely related to the definition of the loss function. Let $L[\hat{h}(x^*, w), y^*]$ be a real value loss function which is defined in $\mathcal{H} \times \mathcal{Y}$, where $\mathcal{Y}$ is the value space of y^*. Then, the risk function is defined as follows:

$$F(w, S) = E_{z^*}\{L[\hat{h}(x^*, w), y^*]\},$$

which is MSPE given the sample S and weight vector w.

2.2 Related Concepts

In this paper, we mainly discuss whether $F(\hat{w}, S)$ can approximate the smallest possible risk $inf_{w \in W} F(w, S)$. If $A(w)$ has such a property, we say that $A(w)$ is consistent. For fixed m, related concepts from Shalev-Shwartz et al. (2010) can be used to discuss the stability and consistency of A_m. Obviously, for model averaging, we need to pay more attention to the stability and consistency of weight selection criterion. We note that the relevant conclusions of Shalev-Shwartz et al. (2010) cannot be directly applied to model averaging because $\mathcal{H}$ depends on S. Therefore, we extend the relevant definitions and conclusions to model averaging. The following is the definition of consistency:

Definition 1 (Consistency) *If there is a sequence of constants* $\{\epsilon_n, n \in N_+\}$ *such that* $\epsilon_n = o(1)$ *and* $\hat{w}$ *satisfies*

$$E_S\big[F(\hat{w}, S) - inf_{w\in W} F(w, S)\big] = O(\epsilon_n),$$

then $A(w)$ *is said to be consistent with rate* ϵ_n*.*

In statistical learning theory, stability concerns how much the algorithm's output varies when S changes a little. "Leave-one-out (Loo)" and "Replace-one (Ro)" are two common tools used to evaluate stability. Loo considers the change in the algorithm's output after removing an observation from S and Ro considers such a change after replacing an observation in S with an observation that is independent of S. Accordingly, the stability is called Loo stability and Ro stability, respectively. Here, we will give the formal definitions of Loo stability and Ro stability. To this end, we first give the definition of algorithm symmetry:

Definition 2 (Symmetry) *If the algorithm's output is not affected by the order of the observations in* S*, then the algorithm is symmetric.*

Now let S^{-i} be the sample after removing the i-th observation from S, $\hat{h}_m^{-i}$ be the estimator of h_m based on S^{-i} and A_m, $\hat{w}^{-i}$ be the estimator of weight vector based on S^{-i} and $A(w)$ and $F(w, S^{-i}) = E_{z^*}\big\{L[\hat{h}^{-i}(x^*, w), y^*]\big\}$, where

$$\hat{h}^{-i}(x^*, w) = H[w, \hat{h}_1^{-i}(x_1^*), \hat{h}_2^{-i}(x_2^*), \cdots, \hat{h}_M^{-i}(x_M^*)].$$

We define Loo stability as follows:

Definition 3 (PLoo Stability) *If there is a sequence of constants* $\{\epsilon_n, n \in N_+\}$ *such that* $\epsilon_n = o(1)$ *and* $A(w)$ *satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} E_S\big[F(\hat{w}, S) - F(\hat{w}^{-i}, S^{-i})\big] = O(\epsilon_n),$$

then $A(w)$ *is Predicted-Loo (PLoo) stable with rate* ϵ_n*; If* $\{A_m, m = 1, 2, \cdots, M\}$ *and* $A(w)$ *are symmetric, then a PLoo stable* $A(w)$ *needs only to satisfy*

$$E_S\big[F(\hat{w}, S) - F(\hat{w}^{-n}, S^{-n})\big] = O(\epsilon_n).$$

Definition 4 (FLoo Stability) *If there is a sequence of constants* $\{\epsilon_n, n \in N_+\}$ *such that* $\epsilon_n = o(1)$ *and* $A(w)$ *satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} E_S\big\{L[\hat{h}(x_i, \hat{w}), y_i] - L[\hat{h}^{-i}(x_i, \hat{w}^{-i}), y_i]\big\} = O(\epsilon_n),$$

then $A(w)$ *is Fitted-Loo (FLoo) stable with rate* ϵ_n*; If* $\{A_m, m = 1, 2, \cdots, M\}$ *and* $A(w)$ *are symmetric, then a FLoo stable* $A(w)$ *needs only to satisfy*

$$E_S\big\{L[\hat{h}(x_n, \hat{w}), y_n] - L[\hat{h}^{-n}(x_n, \hat{w}^{-n}), y_n]\big\} = O(\epsilon_n).$$

Let S^i be the sample S with the i-th observation replaced by $z_i^* = (y_i^*, x_i^{*'})'$, $\hat{h}_m^i$ be the estimator of h_m based on S^i and A_m, and $\hat{w}^i$ be the estimator of the weight vector based on S^i and $A(w)$, where z_i^* is from distribution $\mathcal{D}$ and independent of S. Let $F(w, S^i) = E_{z^*}\big\{L[\hat{h}^i(x^*, w), y^*]\big\}$, where $\hat{h}^i(x^*, w) = H[w, \hat{h}_1^i(x_1^*), \hat{h}_2^i(x_2^*), \cdots, \hat{h}_M^i(x_M^*)]$. Note that

$$\frac{1}{n}\sum_{i=1}^{n} E_{S,z_i^*}\big[F(\hat{w}, S) - F(\hat{w}^i, S^i)\big] = 0.$$

Therefore, we define Ro stability as follows:

Definition 5 (Ro Stability) *If there is a sequence of constants $\{\epsilon_n, n \in N_+\}$ such that $\epsilon_n = o(1)$ and $A(w)$ satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} E_{S,z_i^*}\big\{L[\hat{h}(x_i, \hat{w}), y_i] - L[\hat{h}^i(x_i, \hat{w}^i), y_i]\big\} = O(\epsilon_n),$$

then $A(w)$ is Ro stable with rate ϵ_n; If $\{A_m, m = 1, 2, \cdots, M\}$ and $A(w)$ are symmetric, then an Ro stable $A(w)$ needs only to satisfy

$$E_{S,z_n^*}\big\{L[\hat{h}(x_n, \hat{w}), y_n] - L[\hat{h}^n(x_n, \hat{w}^n), y_n]\big\} = O(\epsilon_n).$$

Before discussing the relationship between stability and consistency, we give the definitions of AERM and generalization. The empirical risk function is defined as follows:

$$\hat{F}(w, S) = \frac{1}{n}\sum_{i=1}^{n} L[\hat{h}(x_i, w), y_i].$$

Definition 6 (AERM) *If there is a sequence of constants $\{\epsilon_n, n \in N_+\}$ such that $\epsilon_n = o(1)$ and $A(w)$ satisfies*

$$E_S\big[\hat{F}(\hat{w}, S) - inf_{w \in W}\hat{F}(w, S)\big] = O(\epsilon_n),$$

then $A(w)$ is an AERM with rate ϵ_n.

Vapnik (1998) proved some theoretical properties of the empirical risk minimization principle. However, when the sample size is small, the empirical risk minimizer tends to produce over-fitting phenomenon. Therefore, the structural risk minimization principle is proposed in Vapnik (1998). Shalev-Shwartz et al. (2010) also discussed the deficiency of the empirical risk minimization principle and the importance of AERM.

Definition 7 (Generalization) *If there is a sequence of constants $\{\epsilon_n, n \in N_+\}$ such that $\epsilon_n = o(1)$ and $A(w)$ satisfies*

$$E_S\big[\hat{F}(\hat{w}, S) - F(\hat{w}, S)\big] = O(\epsilon_n),$$

then $A(w)$ generalizes with rate ϵ_n.

In statistical learning theory, generalization refers to the performance of the concept learned by models on unknown sample. It can be seen from Definition 7 that the generalization of $A(w)$ describes the difference between using $\hat{w}$ to fit the training set S and predict the unknown sample.

2.3 Relationship between Different Concepts

In this subsection, we study the relationship between different concepts based on the findings from Bousquet and Elisseeff (2002) and Shalev-Shwartz et al. (2010). Bousquet and Elisseeff (2002) uses triangle inequality to illustrate that Loo stability implies Ro stability. However, we note that for any $i \in \{1, 2, \cdots, n\}$,

$$\begin{aligned}
&E_{S,z_i^*}\Big\{L[\hat{h}(x_i,\hat{w}),y_i] - L[\hat{h}^i(x_i,\hat{w}^i),y_i]\Big\} \\
&= E_{S,z_i^*}\Big\{L[\hat{h}(x_i,\hat{w}),y_i] - L[\hat{h}^{-i}(x_i,\hat{w}^{-i}),y_i] + L[\hat{h}^{-i}(x_i,\hat{w}^{-i}),y_i] - L[\hat{h}^i(x_i,\hat{w}^i),y_i]\Big\} \\
&= E_S\Big\{L[\hat{h}(x_i,\hat{w}),y_i] - L[\hat{h}^{-i}(x_i,\hat{w}^{-i}),y_i]\Big\} + E_S[F(\hat{w}^{-i},S^{-i}) - F(\hat{w},S)].
\end{aligned}$$

From this, we give the following theorem to illustrate the relationship between Loo stability and Ro stability:

Theorem 8 *If $A(w)$ has two of FLoo stability, PLoo stability and Ro stability with rate ϵ_n, then it has all three stabilities with rate ϵ_n.*

Shalev-Shwartz et al. (2010) emphasized that Ro stability and Loo stability are in general incomparable notions, but Theorem 8 shows that they are closely related. By definitions of generalization and Ro stability, we have

$$\begin{aligned}
&E_S[\hat{F}(\hat{w},S) - F(\hat{w},S)] \\
&= E_{S,z_1^*,z_2^*,\cdots,z_n^*}\Big\{\frac{1}{n}\sum_{i=1}^n L[\hat{h}(x_i,\hat{w}),y_i] - \frac{1}{n}\sum_{i=1}^n L[\hat{h}(x_i^*,\hat{w}),y_i^*]\Big\} \\
&= \frac{1}{n}\sum_{i=1}^n E_{S,z_i^*}\Big\{L[\hat{h}(x_i,\hat{w}),y_i] - L[\hat{h}^i(x_i,\hat{w}^i),y_i]\Big\}.
\end{aligned}$$

From this, we can give the following theorem to illustrate the equivalence of Ro stability and generalization:

Theorem 9 *$A(w)$ has Ro stability with rate ϵ_n if and only if $A(w)$ generalizes with rate ϵ_n.*

For the symmetric algorithm, this result has been given by Lemma 7 of Bousquet and Elisseeff (2002), and extending Lemma 7 of Bousquet and Elisseeff (2002) to the asymmetric case is straightforward. For the theoretical completeness of this section, we still present this result here as a theorem. Theorem 9 shows that stability is an important property of weight selection criteria, which can ensure that the corresponding estimator has good generalization performance. Let $\hat{w}^* \in W$ satisfy $F(\hat{w}^*,S) = inf_{w\in W}F(w,S)$. Note that

$$\begin{aligned}
&E_S[F(\hat{w},S) - F(\hat{w}^*,S)] \\
&= E_S[F(\hat{w},S) - \hat{F}(\hat{w},S) + \hat{F}(\hat{w},S) - \hat{F}(\hat{w}^*,S) + \hat{F}(\hat{w}^*,S) - F(\hat{w}^*,S)] \\
&\le E_S[F(\hat{w},S) - \hat{F}(\hat{w},S) + \hat{F}(\hat{w},S) - inf_{w\in W}\hat{F}(w,S) + \hat{F}(\hat{w}^*,S) - F(\hat{w}^*,S)].
\end{aligned}$$

We give the following theorem to illustrate the relationship between stability and consistency:

Theorem 10 *If $A(w)$ is an AERM and has Ro stability with rate ϵ_n, and $\hat{w}^*$ satisfies*

$$E_S\big[\hat{F}(\hat{w}^*, S) - F(\hat{w}^*, S)\big] = O(\epsilon_n),$$

then $A(w)$ is consistent with rate ϵ_n.

Remark 11 *For general learning algorithms, Bousquet and Elisseeff (2002) studied how to use stability to derive generalization error bounds based on the empirical error and the leave-one-out error, while we study the relationships among AERM, stability, generalization and consistency in model averaging. Although the relationships among Loo stability, Ro stability and generalization that are demonstrated by Bousquet and Elisseeff (2002) are applicable to model averaging, no relationship between stability and consistency is explored. From Theorem 10, we can see the close relationship between stability and consistency in model averaging.*

Remark 12 *If for any $v > 0$, there exists a w_v independent of S such that $F(w_v, S) \leq F(\hat{w}^*, S) + v$, then Lemma 15 of Shalev-Shwartz et al. (2010) can be applied to model averaging. However, since $\hat{w}^*$ and $\mathcal{H}$ depend on S, it is difficult to guarantee that such a w_v always exists, and so we could not immediately obtain Theorem 10 from the proof of Lemma 15 in Shalev-Shwartz et al. (2010). Further, on the basis of Lemma 15 of Shalev-Shwartz et al. (2010), we find that this requirement for the existence of $\{w_v : v > 0\}$ can be replaced by "$F(w, S)$ generalizes with rate ϵ_n".*

In the next section, we will propose an L_2-penalty model averaging method and prove that it has stability and consistency under certain reasonable conditions.

3. L_2-penalty Model Averaging

In most existing literature on model averaging, the theoretical properties are explored under the weight set $W^0 = \{w \in [0,1]^M : \sum_{m=1}^M w_m = 1\}$. From Definition 1, it is seen that even if the corresponding weight selection criterion is consistent, such consistency holds only under the subspace of R^M. Therefore, a natural question is whether it is possible to make the weight space unrestricted. What will happen when we do so? The unrestricted Granger-Ramanathan method obtains the estimator of the weight vector under R^M by minimizing the sum of squared forecast errors from the combination forecast. However, its poor performance is observed when it is compared with some other methods (Hansen, 2008). One possible reason for this is that it merely removes weight constraints without addressing the resulting weight instability. In Section 3.2, we will provide relevant explanations for the causes of this instability. On the other hand, in the prediction task, we are more concerned about whether the resulting estimator can predict better. Intuitively, the estimator that minimizes MSPE in the full space would most likely outperform the estimator that minimizes MSPE in the subspace. To demonstrate this point, some new ideas are needed.

3.1 Model Framework and Estimators

We assume that the response variable y_i and covariates $x_{1i}, x_{2i}, \cdots$ satisfy the following data generating process:

$$y_i = \mu_i + e_i = \sum_{k=1}^{\infty} x_{ki}\theta_k + e_i,$$

and M candidate models are given by

$$y_i = \sum_{k=1}^{k_m} x_{m(k)i}\theta_{m,(k)} + e_i, m = 1, 2, \cdots, M.$$

We assume that the M-th candidate model contains all of the considered covariates and define that $b_{mi} = \mu_i - \sum_{k=1}^{k_m} x_{m(k)i}\theta_{m,(k)}$ is the approximating error of the m-th candidate model. Let $x_i = (x_{(1)i}, x_{(2)i}, \cdots, x_{(k_M)i})$ and $S = \{z_i = (y_i, x_i^{'})^{'}, i = 1, 2, \cdots, n\}$ be a simple random sample from distribution $\mathcal{D}$ throughout the rest of this article, where $x_{(k)i} = x_{M(k)i}$.

Let $y = (y_1, y_2, \cdots, y_n)^{'}$, $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^{'}$, $e = (e_1, e_2, \cdots, e_n)^{'}$ and $b_m = (b_{m1}, b_{m2}, \cdots, b_{mn})^{'}$. Then, the corresponding matrix form of the true model is $y = X_m\theta_m + b_m + e$, where $\theta_m = (\theta_{m,(1)}, \theta_{m,(2)}, \cdots, \theta_{m,(k_m)})^{'}$ and X_m is the design matrix of the m-th candidate model. When there is an candidate model such that $b_m = \mathbf{0}$, the true model is included in the m-th candidate model, i.e. the model is correctly specified. Unlike Hansen (2007) , Wan et al. (2010) and Hansen and Racine (2012), we do not require that $\inf_{w\in W^0} R_n(w)$ (this is defined in Section 3.2) tends to infinity. Therefore, we allow that the model is correctly specified.

Let $\pi_m \in R^{K\times k_m}$ be the variable selection matrix satisfying $X_M\pi_m = X_m$ and $\pi_m^{'}\pi_m = I_{k_m}$, $m = 1, 2, \cdots, M$. Then, the hypothesis spaces of M candidate models are

$$\mathcal{H}_m = \left\{x^{*'}\pi_m\theta_m, \theta_m \in R^{k_m}\right\}, m = 1, 2, \cdots, M.$$

The least squares estimator of θ_m is $\hat{\theta}_m = (X_m^{'}X_m)^{-1}X_m^{'}y$, $m = 1, 2, \cdots, M$.

3.2 Weight Selection Criterion

Let $P_m = X_m(X_m^{'}X_m)^{-1}X_m^{'}$, $P(w) = \sum_{m=1}^{M} w_m P_m$, $L_n(w) = \|\mu - P(w)y\|_2^2$ and $R_n(w) = E_e[L_n(w)]$. When $\sigma_i^2 \equiv \sigma^2$, Hansen (2007) and Wan et al. (2010) used Mallows criterion $C_n(w) = \|y - \hat{\Omega}w\|_2^2 + 2\sigma^2 w^{'}\kappa$ to select a model weight vector from W^0 and proved that the estimator of the weight vector asymptotically minimizes $L_n(w)$, where $\hat{\Omega} = (P_1y, P_2y, \cdots, P_My)$ and $\kappa = (k_1, k_2, \cdots, k_M)^{'}$. Hansen and Racine (2012) used Jackknife criterion $J_n(w) = \|y - \bar{\Omega}w\|_2^2$ to select a model weight vector from W^0 and proved that the estimator of the weight vector asymptotically minimizes $L_n(w)$ and $R_n(w)$, where $\bar{\Omega} = \left[y - D_1(I - P_1)y, y - D_2(I - P_2)y, \cdots, y - D_M(I - P_M)y\right]$ with $D_m = diag[(1 - h_{ii}^m)^{-1}]$ and $h_{ii}^m = x_i^{'}\pi_m(X_m^{'}X_m)^{-1}\pi_m^{'}x_i, i = 1, 2, \cdots, n$.

Different from Hansen (2007), Wan et al. (2010) and Hansen and Racine (2012), we focus on whether the model averaging estimator can asymptotically minimize MSPE when the model weights are not restricted. Let $\hat{\gamma} = (x^{*'}\pi_1\hat{\theta}_1, x^{*'}\pi_2\hat{\theta}_2, \cdots, x^{*'}\pi_M\hat{\theta}_M)$. Then, the risk function and the empirical risk function are defined as:

$$F(w, S) = E_{z^*}[y^* - x^{*'}\hat{\theta}(w)]^2 = E_{z^*}[(y^* - \hat{\gamma}w)^2]$$

and

$$\hat{F}(w,S)=\frac{1}{n}\sum_{i=1}^{n}[y_i-x_i^{'}\hat{\theta}(w)]^2=\frac{1}{n}\|y-\hat{\Omega}w\|_2^2,$$

respectively, where $\hat{\theta}(w)=\sum_{m=1}^{M}w_m\pi_m\hat{\theta}_m$. Since Hansen (2007), Wan et al. (2010) and Hansen and Racine (2012) restricted $w\in W^0$, the corresponding estimators of the weight vector do not necessarily asymptotically minimize $F(w,S)$ on R^M. An intuitive way that makes the estimator of the weight vector asymptotically minimize $F(w,S)$ on R^M is to remove the restriction $w\in W^0$ directly.

Let $\hat{P}$ and $\bar{P}$ be the orthogonal matrices satisfying $\hat{P}^{'}\hat{\Omega}^{'}\hat{\Omega}\hat{P}=diag(\hat{\zeta}_1,\hat{\zeta}_2,\cdots,\hat{\zeta}_M)$ and $\bar{P}^{'}\bar{\Omega}^{'}\bar{\Omega}\bar{P}=diag(\bar{\zeta}_1,\bar{\zeta}_2,\cdots,\bar{\zeta}_M)$, where $\hat{\zeta}_1\leq\hat{\zeta}_2\leq\cdots\leq\hat{\zeta}_M$ and $\bar{\zeta}_1\leq\bar{\zeta}_2\leq\cdots\leq\bar{\zeta}_M$ are the eigenvalues of $\hat{\Omega}^{'}\hat{\Omega}$ and $\bar{\Omega}^{'}\bar{\Omega}$, respectively. We assume that $E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})$, $\hat{\Omega}^{'}\hat{\Omega}$ and $\bar{\Omega}^{'}\bar{\Omega}$ are invertible (this is reasonable under Assumption 3), then

$$\hat{w}^0=argmin_{w\in R^M}C_n(w)=(\hat{\Omega}^{'}\hat{\Omega})^{-1}(\hat{\Omega}^{'}y-\sigma^2\kappa),$$

$$\bar{w}^0=argmin_{w\in R^M}J_n(w)=(\bar{\Omega}^{'}\bar{\Omega})^{-1}\bar{\Omega}^{'}y,$$

$$\tilde{w}=argmin_{w\in R^M}\hat{F}(w,S)=(\hat{\Omega}^{'}\hat{\Omega})^{-1}\hat{\Omega}^{'}y$$

and

$$\hat{w}^*=argmin_{w\in R^M}F(w,S)=[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})]^{-1}E_{z^*}(\hat{\gamma}^{'}y^*).$$

In order to satisfy the consistency, $\hat{w}^0$ and $\bar{w}^0$ should be good estimators of $\hat{w}^*$. However, when candidate models are highly correlated, the minimum eigenvalues of $\hat{\Omega}^{'}\hat{\Omega}$ and $\bar{\Omega}^{'}\bar{\Omega}$ may be small so that $\|\hat{w}^0\|_2^2=\sum_{m=1}^{M}\frac{a_m^2}{\hat{\zeta}_m^2}\geq\frac{a_1^2}{\hat{\zeta}_1^2}$ and $\|\bar{w}^0\|_2^2=\sum_{m=1}^{M}\frac{b_m^2}{\bar{\zeta}_m^2}\geq\frac{b_1^2}{\bar{\zeta}_1^2}$ are too large, which usually result in extreme weights, where $(a_1,a_2,\cdots,a_M)^{'}=\hat{P}^{'}(\hat{\Omega}^{'}y-\sigma^2\kappa)$ and $(b_1,b_2,\cdots,b_M)^{'}=\bar{P}^{'}\bar{\Omega}y$. Therefore, similar to ridge regression in Hoerl and Kennard (1970), we make the following correction to $C_n(w)$ and $J_n(w)$:

$$C(w,S)=C_n(w)+\lambda_n w^{'}w$$

and

$$J(w,S)=J_n(w)+\lambda_n w^{'}w,$$

where $\lambda_n\geq 0$ is a tuning parameter. The above corrections are actually L_2-penalty for weight vector. Let $\hat{Z}=(\hat{\Omega}^{'}\hat{\Omega}+\lambda_n I)^{-1}\hat{\Omega}^{'}\hat{\Omega}$ and $\bar{Z}=(\bar{\Omega}^{'}\bar{\Omega}+\lambda_n I)^{-1}\bar{\Omega}^{'}\bar{\Omega}$. Then

$$\hat{w}=argmin_{w\in R^M}C(w,S)=(\hat{\Omega}^{'}\hat{\Omega}+\lambda_n I)^{-1}(\hat{\Omega}^{'}y-\sigma^2\kappa)=\hat{Z}\hat{w}^0$$

and

$$\bar{w}=argmin_{w\in R^M}J(w,S)=(\bar{\Omega}^{'}\bar{\Omega}+\lambda_n I)^{-1}\bar{\Omega}^{'}y=\bar{Z}\bar{w}^0.$$

In the next subsection, we discuss the theoretical properties of $C(w,S)$ and $J(w,S)$.

3.3 Stability and Consistency

Let $\lambda_{min}(\cdot)$ be the minimum eigenvalue of a square matrix, X_m^{-n} be the matrix after removing the n-th row of X_m, $X_m^{-(n-1,n)}$ be the matrix after removing the n-th and $(n-1)$-th rows of X_m, y^{-n} be the vector after removing the n-th element of y and χ be the value space of K covariates, where $K = k_M$. Then, we define

$$\hat{\Omega}^{-n} = (X_1^{-n}\hat{\theta}_1^{-n}, X_2^{-n}\hat{\theta}_2^{-n}, \cdots, X_M^{-n}\hat{\theta}_M^{-n}),$$

$$\bar{\Omega}^{-n} = \left[(I_{n-1} - D_m^{-n})y^{-n} + D_m^{-n}X_m^{-n}\hat{\theta}_m^{-n} : m = 1, 2\cdots, M\right]$$

and

$$\hat{\gamma}^{-n} = (x^{*'}\pi_1\hat{\theta}_1^{-n}, x^{*'}\pi_2\hat{\theta}_2^{-n}, \cdots, x^{*'}\pi_M\hat{\theta}_M^{-n}),$$

where

$$\hat{\theta}_m^{-n} = (X_m^{-n'}X_m^{-n})^{-1}X_m^{-n'}y^{-n}$$

and

$$D_m^{-n} = diag[(1 - h_{ii}^{m,-n})^{-1} : i \neq n]$$

with $h_{ii}^{m,-n} = x_i^{'}\pi_m(X_m^{-n'}X_m^{-n})^{-1}\pi_m^{'}x_i$. In order to discuss the stability and consistency of the proposed method, we need the following assumptions:

Assumption 1 *There is a constant $C_1 > 0$ such that $\lambda_{min}(n^{-1}X_M^{-(n-1,n)'}X_M^{-(n-1,n)}) \geq C_1$, a.s..*

Assumption 2 *There is a constant $C_2 > 0$ such that $n^{-1}y^{'}y \leq C_2$, a.s.; There is a constant $C_3 > 0$ such that $\chi \subset B(\mathbf{0_K}, C_3)$ a.s., where $B(\mathbf{0_K}, C_3)$ is a sphere with the center $\mathbf{0_K}$ and radius C_3, and $\mathbf{0_K}$ is the K-dimensional vector of zeros.*

Assumption 3 *There is a constant $C_4 > 0$ such that*

$$\min[\lambda_{min}(n^{-1}\hat{\Omega}^{'}\hat{\Omega}), \lambda_{min}(n^{-1}\hat{\Omega}^{-n'}\hat{\Omega}^{-n})] \geq C_4, \ a.s.,$$

$$\min[\lambda_{min}(n^{-1}\bar{\Omega}^{'}\bar{\Omega}), \lambda_{min}(n^{-1}\bar{\Omega}^{-n'}\bar{\Omega}^{-n})] \geq C_4, \ a.s.$$

and

$$\min\{\lambda_{min}[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})], \lambda_{min}[E_{z^*}(\hat{\gamma}^{-n'}\hat{\gamma}^{-n})]\} \geq C_4, \ a.s..$$

Assumption 4 *There is a constant $C_5 > 0$ such that $\max(\|\hat{w}^*\|_2^2, \|\hat{w}^{-n*}\|_2^2) \leq C_5$, a.s., where $\hat{w}^{-n*} = argmin_{w \in R^M} F(w, S^{-n})$ and $F(w, S^{-n}) = E_{z^*}\{[y^* - x^{*'}\hat{\theta}^{-n}(w)]^2\}$.*

Assumption 5 *$K^3M^2 = o(n)$ and $\lambda_n = O(K^2M)$.*

Assumption 1 is weak as $n^{-1}X_M^{'}X_M$ is often positive definite. The similar assumption is often made in literature. For example, Condition (b) of Zou (2006) and Condition (C.1) of Zhang and Liu (2019) require $n^{-1}X_M^{'}X_M$ to converge to a positive definite matrix. In particular, Zhang and Liu (2019) points out that if y_i is a stationary and ergodic martingale difference sequence with finite fourth moments, then their Condition (C.1) is true. Here we

provide an example where Assumption 1 holds. Let $x_1, \cdots, x_n$ be i.i.d. Gaussian random vectors with zero mean and unit covariance matrix and $K/n \to B \in (0,1)$. Then by using Theorem 2 of Bai and Yin (1993), we see that

$$lim_{n\to\infty}\lambda_{min}(n^{-1}X_M^{'}X_M) = (1-\sqrt{B})^2, \ a.s.$$

which shows that Assumption 1 is true. Shalev-Shwartz et al. (2010) assumed that the loss function is bounded, which is usually not satisfied in traditional regression analysis. We replace this assumption with Assumption 2. Tong and Wu (2017) assumed that $\chi \times \mathcal{Y}$ is a compact subset of R^{K+1}, under which Assumption 2 is obviously true. Assumption 3 requires that matrices $n^{-1}\hat{\Omega}^{'}\hat{\Omega}$ and $n^{-1}\hat{\Omega}^{-n'}\hat{\Omega}^{-n}$ are almost always positive definite, which is similar to condition (C.4) of Liao and Zou (2020). Lemmas 22 and 23 justify the rationality of the assumptions regarding the eigenvalue of $n^{-1}\hat{\Omega}^{'}\hat{\Omega}$. Lemma 24 guarantees the rationality of this assumption about the eigenvalues of $\bar{\Omega}^{'}\bar{\Omega}$, $\bar{\Omega}^{-n'}\bar{\Omega}^{-n}$, $E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})$ and $E_{z^*}(\hat{\gamma}^{-n'}\hat{\gamma}^{-n})$. Assumption 4 requires that the L_2-norm of the optimal weight $\hat{w}^*$ and $\hat{w}^{-n*}$ is bounded. Lemma 25 shows that Assumption 4 has a certain rationality. Further, Lemma 29 provides a case where Assumption 4 holds. Assumption 5 limits the growth rate of the numbers of covariates and models, and also makes a mild assumption about λ_n to avoid excessive penalty.

Let $\hat{V}(\lambda_n) = \|\hat{Z}\hat{w}^0 - \hat{Z}\hat{w}^*\|_2^2$, $\hat{B}(\lambda_n) = \|\hat{Z}\hat{w}^* - \hat{w}^*\|_2^2$, $\bar{V}(\lambda_n) = \|\bar{Z}\bar{w}^0 - \bar{Z}\hat{w}^*\|_2^2$ and $\bar{B}(\lambda_n) = \|\bar{Z}\hat{w}^* - \hat{w}^*\|_2^2$. We define

$$\hat{M}(\lambda_n) = \|\hat{Z}\hat{w}^0 - \hat{w}^*\|_2^2 = \hat{V}(\lambda_n) + \hat{B}(\lambda_n) + 2(\hat{Z}\hat{w}^0 - \hat{Z}\hat{w}^*)^{'}(\hat{Z}\hat{w}^* - \hat{w}^*)$$

and

$$\bar{M}(\lambda_n) = \|\bar{Z}\bar{w}^0 - \hat{w}^*\|_2^2 = \bar{V}(\lambda_n) + \bar{B}(\lambda_n) + 2(\bar{Z}\bar{w}^0 - \bar{Z}\hat{w}^*)^{'}(\bar{Z}\hat{w}^* - \hat{w}^*).$$

In order to make $F(\hat{Z}\hat{w}^0, S)$ and $F(\bar{Z}\bar{w}^0, S)$ better approximate $F(\hat{w}^*, S)$, we naturally hope $E_S[\hat{M}(\lambda_n)]$ and $E_S[\bar{M}(\lambda_n)]$ to be as small as possible. In the following discussion, we refer to $E_S[\hat{M}(\lambda_n)]$ and $E_S[\bar{M}(\lambda_n)]$, $E_S[\hat{V}(\lambda_n)]$ and $E_S[\bar{V}(\lambda_n)]$, $E_S[\hat{B}(\lambda_n)]$ and $E_S[\bar{B}(\lambda_n)]$ as the corresponding mean squared errors, estimation variances and estimation biases of model weight estimator, respectively. Obviously, when $\lambda_n = 0$, $\hat{Z} = \bar{Z} = I_M$ which means that the estimation bias is equal to zero. From Assumption 3 and Lemma 25, we see that under Assumptions 1-5, $\hat{B}(\lambda_n)$ and $\bar{B}(\lambda_n)$ are $O(n^{-2}\lambda_n^2)$ a.s.. On the other hand, the existence of extreme weights may make the performance of $\hat{w}^0$ and $\bar{w}^0$ extremely unstable. So the purpose of introducing L_2-penalty is to reduce estimation variance by introducing estimation bias, which results in the stable performance of the model averaging estimator. Further, we define

$$\hat{M}_1(\lambda_n) = \hat{V}(\lambda_n) + \hat{B}(\lambda_n)$$

and

$$\bar{M}_1(\lambda_n) = \bar{V}(\lambda_n) + \bar{B}(\lambda_n).$$

Like Theorem 4.3 in Hoerl and Kennard (1970), we give the following theorem to illustrate the reasonability of introducing L_2-penalty:

Theorem 13 *Let* $\hat{\lambda}_n = \min\{\lambda_n : \frac{d}{d\lambda_n}\hat{M}_1(\lambda_n) = 0\}$ *and* $\bar{\lambda}_n = \min\{\lambda_n : \frac{d}{d\lambda_n}\bar{M}_1(\lambda_n) = 0\}$. *Then, 1) when* $\hat{w}^0 \neq \hat{w}^*$, $\hat{\lambda}_n > 0$ *and* $\hat{M}_1(\hat{\lambda}_n) < \hat{M}_1(0)$*; 2) when* $\bar{w}^0 \neq \hat{w}^*$, $\bar{\lambda}_n > 0$ *and* $\bar{M}_1(\bar{\lambda}_n) < \bar{M}_1(0)$.

Theorem 13 shows that the use of L_2-penalty reduces estimation variance by introducing estimation bias. However, since $\hat{w}^*$ is unknown, $\hat{\lambda}_n$ and $\bar{\lambda}_n$ are also unknown. In Section 3.4, we use cross validation to select the tuning parameter λ_n. The following theorem shows that $C(w,S)$ and $J(w,S)$ are AERM.

Theorem 14 *Under Assumptions 1-5, both $C(w,S)$ and $J(w,S)$ are AERMs with rate $n^{-1}K^2M$.*

The following theorem shows that $C(w,S)$, $J(w,S)$ and $F(w,S)$ have FLoo stability and PLoo stability.

Theorem 15 *Under Assumptions 1-5, $C(w,S)$, $J(w,S)$ and $F(w,S)$ all have FLoo stability and PLoo stability with rate $n^{-1}K^3M^2$.*

It can be seen from Theorems 8, 9 and 15 that $C(w,S)$, $J(w,S)$ and $F(w,S)$ have Ro stability and generalization. The following theorem shows that $C(w,S)$ and $J(w,S)$ have consistency, which is a direct consequence of Theorems 8-10 and 14-15.

Theorem 16 *Under Assumptions 1-5, both $C(w,S)$ and $J(w,S)$ have consistency with rate $n^{-1}K^3M^2$.*

From Theorem 3 and Proposition 3 of Mourtada (2022), we see that the excess quadratic risk of the largest candidate model is of order $O(K/n)$ in the linear regression with random-design. However, Mourtada (2022) required to assume $E(\|x^*\|_2^2) < \infty$, which is usually not true when K diverges, e.g. $x^*_{(1)}, x^*_{(2)}, \cdots, x^*_{(K)}$ are independently identically distributed with $U(-1,1)$. Under the assumption that the functions to be aggregated are bounded and independent of the current data, Theorem 4 of Tsybakov (2003) showed that the excess quadratic risk of linear aggregation is of order $O(M/n)$. This boundedness assumption is often not met in regression analysis. For $C(w,S)$, in order to get a faster rate $O(KM/n)$, we need the following additional assumptions:

Assumption 6 $E_S\{\Lambda^2_{max}[E_{z^*}(\hat{\gamma}'\hat{\gamma}) - \hat{\Omega}'\hat{\Omega}/n]\} = O(n^{-2}K^4M^2)$, *where $\Lambda_{max}(\cdot)$ represents the maximum singular value of the corresponding matrix.*

Assumption 7 $\max_{1\le k\le K} var(x^*_{(k)}y^*) \le C_6$.

Theorem 17 *Under Assumptions 1-3 and 5-7, $C(w,S)$ has consistency with rate $O(n^{-1}KM)$.*

Remark 18 *From the proof of Lemma 24 and Gershgorin's Theorem, we have*

$$\Lambda^2_{max}\{E_S[E_{z^*}(\hat{\gamma}'\hat{\gamma}) - \hat{\Omega}'\hat{\Omega}/n]\} = O(n^{-2}K^4M^2),$$

hence Assumption 6 possesses a certain degree of rationality. Assumption 7 is a common constraint on second-order moment.

Remark 19 *Let $\hat{\theta} = (\pi_1\hat{\theta}_1, \pi_2\hat{\theta}_2, \cdots, \pi_M\hat{\theta}_M)$. When $\lambda_{min}(\hat{\theta}\hat{\theta}^{'}) \geq C_7$, a.s. (this requires that the matrix $\hat{\theta}\hat{\theta}^{'}$ is positive definite. Lemma 22 provides a case where this requirement holds), from Marcinkiewicz-Zygmund-Burkholder inequality in Lin and Bai (2010), we have*

$$\begin{aligned}
&E_S\{\|\hat{\theta}^{'}[X_M^{'}y/n - E_{z^*}(x^*y^*)]\|_2^2\} \\
&\geq C_7 E_S\{[X_M^{'}y/n - E_{z^*}(x^*y^*)]^{'}[X_M^{'}y/n - E_{z^*}(x^*y^*)]\} \\
&\geq C_7 \sum_{k=1}^{K} E_S\Big\{\Big[\frac{1}{n}\sum_{i=1}^{n}[x_{(k)i}y_i - E_{z^*}(x^*_{(k)}y^*)]\Big]^2\Big\} \\
&\geq 4^{-1}n^{-2}KC_7 \min_{1\leq k\leq K} E_S\Big\{\sum_{i=1}^{n}[x_{(k)i}y_i - E_{z^*}(x^*_{(k)}y^*)]^2\Big\} \\
&= 4^{-1}n^{-1}KC_7 \min_{1\leq k\leq K} E_S\Big\{\frac{1}{n}\sum_{i=1}^{n}[x_{(k)i}y_i - E_{z^*}(x^*_{(k)}y^*)]^2\Big\} \\
&= 4^{-1}n^{-1}KC_7 \min_{1\leq k\leq K} var(x_{(k)i}y_i).
\end{aligned}$$

Thus, from the proof of Theorem 17 and

$$E_S[\|E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})(\hat{w} - \hat{w}^*)\|_2^2] \leq E_S\{\lambda_{max}[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})](\hat{w} - \hat{w}^*)^{'}E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})(\hat{w} - \hat{w}^*)\},$$

we see that when $\min_{1\leq k\leq K} var(x^*_{(k)}y^*)$ has a non-zero lower bound and $\lambda_{max}[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})] = O(1)$, a.s., then the rate of consistency of $C(w, S)$ is not lower than $O(n^{-1}K)$. Thus, when M is bounded, Theorem 17 indicates that the rate $O(n^{-1}K)$ is optimal. Further, from the proof of Theorem 17, we see that when $\lambda_{max}(\hat{\theta}\hat{\theta}^{'}) = O(1)$, a.s., the rate $O(n^{-1}K)$ is also optimal even if M diverges. For $J(w, S)$, we can obtain a similar conclusion by using (9) and Lemma 25.

3.4 Optimal Weighting Based on Cross Validation

Although Theorem 13 shows that there are $\hat{\lambda}_n$ and $\bar{\lambda}_n$ such that $\hat{w}$ and $\bar{w}$ are better than $\hat{w}^0$ and $\bar{w}^0$, $\hat{\lambda}_n$ and $\bar{\lambda}_n$ cannot be obtained. Therefore, like Schomaker (2012), we propose an algorithm based on cross validation to obtain the estimator of the weight vector, which is a weighted average of the weight estimators for different tuning parameters. That is, we first select $\mathcal{L}$ segmentation points on $[0, M\log n]$ with equal intervals as the candidates of λ_n. Then, we calculate the estimation error for each candidate of λ_n using cross validation. Based on this, we remove those candidates with large estimation error. Lastly, for the remaining candidates, the estimation errors are used to perform a weighted average of the estimators of the weight vector. We summarize our algorithm for RMMA below, and a similar algorithm can be given for RJMA.

Algorithm 1 Optimal weighting based on cross validation

Require: S, $\mathcal{L} \geq 1$, $\mathcal{B} \geq 2$, $l \in [1, \mathcal{L}]$, $b \in [1, \mathcal{B} - 1]$;
Ensure: $\hat{w}$;

1: $\hat{E}_L = 0, L = 1, 2, \cdots, \mathcal{L}$;
2: The sample S is randomly divided into $\mathcal{B}$ equal-sized subsets, and the sample index set belonging to the B-th part is denoted as $S_B, B = 1, 2, \cdots, \mathcal{B}$;

for each $B \in \{1, 2, \cdots, \mathcal{B}\}$ **do**
  if $B + b \leq \mathcal{B}$ **then**
    Let $B_{idx} = (B, B+1, \cdots, B+b)$;
  else
    Let $B_{idx} = (B, B+1, \cdots, \mathcal{B}, 1, 2, \cdots, B+b-\mathcal{B})$;
  Let $S_{train} = \{S_i, i \in B_{idx}\}$ and $S_{test} = \{S_i, i \notin B_{idx}\}$;
  $\hat{\theta}_m^B$ is obtained based on S_{train}, $m = 1, 2, \cdots, M$;
  for each $L \in \{1, 2, \cdots, \mathcal{L}\}$ **do**
    $\hat{w}_{BL}$ is obtained based on $\lambda_n = \frac{(L-1)M \log n}{\mathcal{L}-1}$ and $C(w, S_{train})$;
    Let $\hat{\theta}^B(\hat{w}_{BL}) = \sum_{m=1}^{M} \hat{w}_{BLm} \pi_m \hat{\theta}_m^B$;
    The estimation error of $\hat{w}_{BL}$ on S_{test} is obtained as

$$\hat{E}(\hat{w}_{BL}) = \sum_{z_i \in S_{test}} [y_i - x_i' \hat{\theta}^B(\hat{w}_{BL})]^2;$$

    $\hat{E}_L = \hat{E}_L + \hat{E}(\hat{w}_{BL})$;
Let S_λ be the index set of the smallest l numbers in $\{\hat{E}_L, L = 1, 2, \cdots, \mathcal{L}\}$;
$\hat{w}_L$ is obtained based on $\lambda_n = \frac{(L-1)M \log n}{\mathcal{L}-1}$, S and $C(w, S)$, where $L \in S_\lambda$;
$\hat{w} = \sum_{L \in S_\lambda} \frac{exp(-0.5\hat{E}_L)}{\sum_{L \in S_\lambda} exp(-0.5\hat{E}_L)} \hat{w}_L$.

4. Simulation Study

In this section, we conduct simulation experiments to demonstrate the finite sample performance of the proposed method. Similar to Hansen (2007), we consider the following data generating process:

$$y_i = \mu_i + e_i = \sum_{k=1}^{\mathcal{K}} x_{ki} \theta_k + e_i \quad i = 1, 2, \cdots, n,$$

where $\theta_k, k = 1, 2, \cdots \mathcal{K}$ are the model parameters, $x_{1i} \equiv 1$, $(x_{2i}, x_{3i}, \cdots, x_{\mathcal{K}i}) \sim N(0, \Sigma)$ and $(e_1, e_2, \cdots, e_n) \sim N[0, diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)]$. We set $n = 100, 300, 500, 700$, $\Sigma = (\sigma_{kl})$ and $\sigma_{kl} = \rho^{|k-l|}$ with $\rho = 0.3, 0.6$, and $R^2 = 0.1, 0.2, \cdots, 0.9$, where the population $R^2 = \frac{var(\mu_i)}{var(\mu_i + e_i)}$. For the homoskedastic simulation, we set $\sigma_i^2 \equiv 1$, while for the heteroskedastic simulation, we set $\sigma_i^2 = x_{2i}^2$.

We compare the following model selection/averaging methods: 1) model selection with AIC (AI), model selection with BIC (BI) and model selection with C_p (Cp); 2) model averaging with smoothed AIC (SA) and model averaging with smoothed BIC (SB); 3) Mallows model averaging (MM), jackknife model averaging (JM) and least squares model averaging based on generalized cross validation (GM; Li et al., 2021); 4) Ridge-Mallows model averaging (RM) and Ridge-jackknife model averaging (RJ). To evaluate these methods, we generate a test set $\{(y_i^*, x_i^*), i = 1, 2, \cdots, n_t\}$ by the above data generating process, and

$$MSE = n_t^{-1} \sum_{i=1}^{n_t} [\mu_i^* - x_i^{*'} \hat{\theta}(\hat{w})]^2$$

is calculated as a measure of consistency, where $\mu_i^* = \sum_{k=1}^{\mathcal{K}} x_{ki}^* \theta_k$. In the simulation, we set $n_t = 1000$ and repeat 400 times. For each parameterization, we normalize the MSE by dividing by the infeasible MSE (the mean of the smallest MSEs of M candidate models in 400 simulations).

We consider two candidate model settings. In the first setting, like Hansen (2007), all candidate models are misspecified and the candidate models are strictly nested. In the second setting, the true model is one of the candidate models and the candidate models are non-nested. For Algorithm 1, we set $\mathcal{L} = 100$, $\mathcal{B} = 10$, $l = 50$ and $b = 5$.

4.1 Nested Setting and Results

We set $\mathcal{K} = 400$, $\theta_k = c\sqrt{2\alpha}k^{-\alpha-\frac{1}{2}}$ with $\alpha = 0.5, 1.0, 1.5$ and $K = \log_4^2 n$ (i.e. $K = 11, 17, 20, 22$), where R^2 is controlled by the parameter c. For $\rho = 0.3$, the mean of normalized MSEs in 400 simulations is shown in Figures 1-6. The results with $\rho = 0.6$ are similar and so omitted for saving space.

For the homoskedastic case, we can draw the following conclusions from Figures 1-3. When $\alpha = 0.5$ (Figure 1): 1) RM and RJ perform better than other methods if $R^2 \leq 0.5$ and $n = 300, 500$ and 700, and comparably with the best method in other cases; 2) RM performs better than RJ in most of cases if $n = 100$. When $\alpha = 1.0$ (Figure 2): 1) RM and RJ perform better than other methods if $R^2 \leq 0.8$, and comparably with the best method if $R^2 = 0.9$; 2) RM performs better than RJ in most of cases if $n = 100$. When $\alpha = 1.5$ (Figure 3): 1) RM and RJ always perform better than other methods; 2) RM performs better than RJ in most of cases if $n = 100$.

For the heteroskedastic case, we can draw the following conclusions from Figures 4-6. When $\alpha = 0.5$ (Figure 4): 1) RM and RJ perform better than other methods if $R^2 \leq 0.4$, and comparably with the best method in other cases; 2) RM performs better than RJ. When $\alpha = 1.0$ (Figure 5): 1) RM and RJ perform better than other methods if $R^2 \leq 0.7$, and comparably with the best method in other cases; 2) RM performs better than RJ. When $\alpha = 1.5$ (Figure 6): 1) RM and RJ always perform better than other methods; 2) RM performs better than RJ.

To sum up, the conclusions are as follows: 1) RM and RJ are the best in most cases, and even when they are not the best, their performance is close to that of the best method; 2) When α is small and R^2 is large, GM has the best performance, and RM and RJ are the best in other cases; 3) RM performs better than RJ.

4.2 Non-nested Setting and Results

We set $\mathcal{K} = 12$, and $\theta_k = c\sqrt{2\alpha}k^{-\alpha-\frac{1}{2}}$ with $\alpha = 0.5, 1.0, 1.5$ for $1 \leq k \leq 10$ and $\theta_k = 0$ for $k = 11, 12$, where R^2 is controlled by the parameter c. Each candidate model contains the first 6 covariates, and the last 6 covariates are combined to obtain 2^6 candidate models. For $\rho = 0.3$, the mean of normalized MSEs in 400 simulations is shown in Figures 7-12. Like the nested case, the results with $\rho = 0.6$ are similar and so omitted.

For this setting, we can draw the following conclusions from Figures 7-12. When $\alpha = 0.5$ (Figures 7 and 10): 1) RM and RJ perform better than other methods if $R^2 \leq 0.5$, and have performance close to the best method if $R^2 > 0.5$; 2) RM performs better than RJ in most of cases. When $\alpha = 1.0$ (Figures 8 and 11): 1) RM and RJ perform better than other

methods except for SB if $R^2 \leq 0.7$, but the performance of SB is very unstable; 2) RM and RJ have performance close to the best method if $R^2 > 0.7$; 3) RM performs better than RJ in most of cases. When $\alpha = 1.5$ (Figures 9 and 12): 1) RM and RJ perform better than other methods except for BI and SB, but the performance of SB and BI is very unstable; 2) RM performs better than RJ in most of cases.

To sum up, the conclusions are as follows: 1) RM and RJ are the best in most cases and have stable performance; 2) One of SB, SA, BI and AI may perform the best when R^2 is small or large, but their performance is unstable compared to RM and RJ; 3) On the whole, RM performs better than RJ.

5. Real Data Analysis

In this section, we apply the proposed method to the real "wage1" dataset in Wooldridge (2003) from the US Current Population Survey for the year 1976. There are 526 observations in this dataset. The response variable is the log of average hourly earning, while covariates include: 1) dummy variables—nonwhite, female, married, numdep, smsa, northcen, south, west, construc, ndurman, trcommpu, trade, services, profserv, profocc, clerocc and servocc; 2) non-dummy variables—educ, exper and tenure; 3) interaction variables—nonwhite × educ, nonwhite × exper, nonwhite × tenure, female × educ, female × exper, female × tenure, married × educ, married × exper and married × tenure.

We consider the following two cases: 1) We rank the covariates according to their linear correlations with the response variable, and then consider the strictly nested model averaging method (the intercept term is considered and ranked first); 2) 100 models are selected by the function "regsubsets" in "leaps" package of R language, where the parameters "nvmax" and "nbest" are taken to be 20 and 5, respectively, and other parameters use default values. For Algorithm 1, we set $b = 9$ and the rest of the settings are the same as in the simulation study.

We randomly divide the data into two parts: a training sample S of n observations for estimating the models and a test sample S_t of $n_t = 529 - n$ observations for validating the results. We consider $n = 110, 210, 320, 420$, and

$$MSE = n_t^{-1} \sum_{z_i \in S_t} [y_i - x_i^{'} \hat{\theta}(\hat{w})]^2$$

is calculated as a measure of consistency. We replicate the process 400 times. The box plots of MSEs in 400 simulations are shown in Figures 13-14. From these figures, we see that the performance of RM and RJ is good and stable. We also compute the mean and median of MSEs, as well as the best performance rate (BPR), which is the frequency of achieving the lowest risk across the replications. The results are shown in Tables 1-2. From these tables, we can draw the following conclusions: 1) RM and RJ are superior to other methods in terms of mean and median of MSEs, and BPR; 2) The performance of RM and RJ is basically the same in terms of mean and median of MSEs; 3) For BPR, on the whole, RM outperforms RJ.

6. Concluding Remarks

In this paper, we study the relationship among AERM, stability, generalization and consistency in model averaging. The results indicate that stability is an important property of model averaging, which can ensure that model averaging estimator has good generalization performance and consistency under reasonable conditions. When the model weights are not restricted, similar to ridge regression in Hoerl and Kennard (1970), extreme weights tend to appear due to the influence of correlation between candidate models. This results in poor performance of the corresponding model averaging estimator. Thus, we propose an L_2-penalty model averaging method. We prove that it has stability and consistency. In order to overcome selection uncertainty of tuning parameter, we use cross-validation to select a candidate set of tuning parameter, and then perform a weighted average of the estimators of model weights based on cross-validation errors. We show the superiority of the proposed method with a Monte Carlo simulation and application to a prediction task on the wage1 dataset.

Many issues deserve further investigation. We apply the methods of Section 2 to the generalization of MMA and JMA only for linear regression. It is worth investigating whether it is possible to extend the proposed method to more complex scenarios, such as generalized linear model, quantile regression and dependent data. Further, it is also interesting to develop a model averaging framework with stability and consistency under the online setting. In addition, with RMMA and RJMA, we see that the estimators of the weight vector are explicitly expressed. So, how to study their asymptotic behavior based on these explicit expressions is a meaningful but challenging topic.

7. Acknowledgements

The authors thank the editor, and two referees for their insightful suggestions and comments that have substantially improved an earlier version of this article. This work was partially supported by the National Natural Science Foundation of China (Grant nos. 12031016 and 11971323) and the Beijing Natural Science Foundation (Grant no. Z210003).

References

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory, Akademia Kiado, Budapest*, pages 267–281, 1973.

Donald W.K. Andrews. Generic uniform convergence. *Econometric Theory*, 8(2):241–257, 1992.

Zhi-Dong Bai and Yong-Qua Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

Jean-Marie Dufour. Recursive stability analysis of linear regression relationships: An exploratory methodology. *Journal of Econometrics*, 19(1):31–76, 1982.

Tiago M Fragoso, Wesley Bertoli, and Francisco Louzada. Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, 86(1): 1–28, 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.

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

Bruce E Hansen. Least-squares forecast averaging. *Journal of Econometrics*, 146(2):342–350, 2008.

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.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *In Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence*, page 275–282. 2002.

Xinmin Li, Guohua Zou, Xinyu Zhang, and Shangwei Zhao. Least squares model averaging based on generalized cross validation. *Acta Mathematicae Applicatae Sinica, English Series*, 37(3):495–509, 2021.

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 Econometric*, 209:35–60, 2019.

Zhengyan Lin and Zhidong Bai. *Probability Inequalities*. Springer, Berlin Heidelberg, 2010.

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

Yifan Liu. Penalized mallow's model averaging. *Communications in Statistics-Theory and Methods*, doi: 10.1080/03610926.2023.2264995, 2023.

Colin L Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

Enrique Moral-Benito. Model averaging in economics: An overview. *Journal of Economic Surveys*, 29(1):46–75, 2015.

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.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

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

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

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Alena Skolkova. Model averaging with ridge regularization. Cerge-ei working papers, The Center for Economic Research and Graduate Education - Economics Institute, Prague, 2023. URL `https://EconPapers.repec.org/RePEc:cer:papers:wp758`.

Hongzhi Tong and Qiang Wu. Learning performance of regularized moving least square regression. *Journal of Computational and Applied Mathematics*, 325:42–55, 2017.

Alexandre B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer Berlin Heidelberg, 2003.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

Alan TK Wan and Xinyu Zhang. On the use of model averaging in tourism research. *Annals of Tourism Research*, 36:525–532, 2009.

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

Jeffrey M Wooldridge. *Introductory Econometrics*. Thompson South-Western, Thompson, 2003.

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.

Xinyu Zhang and Chu-An Liu. Inference after model averaging in linear regression models. *Econometric Theory*, 35(4):816–841, 2019.

Xinyu Zhang and Guohua Zou. Model averaging method and its application in forecast. *Statistical Research*, 28(6):6, 2011.

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, 2013.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Appendix A. Lemmas and Their Proofs

Let $\hat{\theta}^{-(n-1,n)}(w)$ be the model averaging estimator based on $S^{-(n-1,n)}$, where $S^{-(n-1,n)}$ is the set of observations after removing the $(n-1)$-th and n-th observations from S. We give the following lemmas in order to prove Theorems 14-17.

The following Lemma 20 shows that the L_2-norms of the parameter estimators of M candidate models are uniformly bounded, which will be repeatedly used in subsequent proofs.

Lemma 20 *Under Assumptions 1 and 2, there exists a constant $B_1 > 0$ such that*

$$\max_{1\le m\le M} max\left\{\|\hat{\theta}_m\|_2^2, \|\hat{\theta}_m^{-n}\|_2^2, \|\hat{\theta}_m^{-(n-1,n)}\|_2^2\right\} \le B_1, a.s..$$

Proof It follows from Assumption 1 that

$$P\big[\lambda_{min}(X_M^{-(n-1,n)'}X_M^{-(n-1,n)}/n) \ge C_1\big] = 1.$$

Thus, we have

$$\begin{aligned}
&P\big[\lambda_{min}(X_M^{-n'}X_M^{-n}/n) \ge C_1\big] \\
&= P\big\{\lambda_{min}[(X_M^{-(n-1,n)'}X_M^{-(n-1,n)} + x_{n-1}x_{n-1}^{'})/n] \ge C_1\big\} \\
&\ge P\big[\lambda_{min}(X_M^{-(n-1,n)'}X_M^{-(n-1,n)}/n) \ge C_1\big] \\
&= 1,
\end{aligned}$$

which means that

$$\lambda_{min}(X_M^{-n'}X_M^{-n}/n) \ge C_1,\ a.s.. \tag{1}$$

Similarly, we have

$$\begin{aligned}
&P\big[\lambda_{min}(X_M^{'}X_M/n) \ge C_1\big] \\
&= P\big\{\lambda_{min}[(X_M^{-n'}X_M^{-n} + x_n x_n^{'})/n] \ge C_1\big\} \\
&\ge P\big[\lambda_{min}(X_M^{-n'}X_M^{-n}/n) \ge C_1\big] \\
&= 1,
\end{aligned}$$

which means that

$$\lambda_{min}(X_M^{'}X_M/n) \ge C_1,\ a.s.. \tag{2}$$

From $X_m^{'}X_m = \pi_m^{'}X_M^{'}X_M\pi_m$, the definition of π_m and (2), we have

$$\lambda_{min}(n^{-1}X_m^{'}X_m) \ge \lambda_{min}(n^{-1}X_M^{'}X_M) \ge C_1, a.s..$$

Note that

$$X_m(X_m^{'}X_m)^{-1}(X_m^{'}X_m)^{-1}X_m^{'} = X_m(X_m^{'}X_m)^{-1/2}(X_m^{'}X_m)^{-1}(X_m^{'}X_m)^{-1/2}X_m^{'}.$$

Hence, we have

$$\lambda_{max}[X_m(X_m^{'}X_m)^{-1}(X_m^{'}X_m)^{-1}X_m^{'}] \leq \frac{1}{\lambda_{min}(X_m^{'}X_m)} \leq \frac{1}{C_1 n}, a.s.. \tag{3}$$

Thus, it follows from Assumption 2 that

$$\begin{aligned}
\max_{1\leq m\leq M} \|\hat{\theta}_m\|_2^2 &= \max_{1\leq m\leq M} \|(X_m^{'}X_m)^{-1}X_m^{'}y\|_2^2 \\
&= \max_{1\leq m\leq M} y^{'}X_m(X_m^{'}X_m)^{-1}(X_m^{'}X_m)^{-1}X_m^{'}y \\
&\leq C_1^{-1}n^{-1}y^{'}y \\
&\leq C_1^{-1}C_2, a.s..
\end{aligned}$$

Note that from Assumption 2, we have

$$n^{-1}y^{-n^{'}}y^{-n} \leq n^{-1}y^{'}y \leq C_2,\ a.s. \tag{4}$$

and

$$n^{-1}y^{-(n-1,n)^{'}}y^{-(n-1,n)} \leq n^{-1}y^{'}y \leq C_2,\ a.s..$$

So in a similar way, we obtain

$$\max_{1\leq m\leq M} \|\hat{\theta}_m^{-n}\|_2^2 \leq C_1^{-1}C_2,\ a.s.$$

and

$$\max_{1\leq m\leq M} \|\hat{\theta}_m^{-(n-1,n)}\|_2^2 \leq C_1^{-1}C_2,\ a.s..$$

We complete the proof by taking $B_1 = C_1^{-1}C_2$. ■

The following lemma characterizes the degree of impact of removing an observation on the parameter estimators of the candidate models. From the proofs of Lemmas 24-28, we can see that they are crucial for proving the stability of our method.

Lemma 21 *Under Assumptions 1 and 2, we have*

$$E_S\Big(\max_{1\leq m\leq M} \|\hat{\theta}_m - \hat{\theta}_m^{-n}\|_2^2\Big) = O(n^{-2}K^2)$$

and

$$E_S\Big(\max_{1\leq m\leq M} \|\hat{\theta}_m^{-n} - \hat{\theta}_m^{-(n-1,n)}\|_2^2\Big) = O(n^{-2}K^2).$$

Proof By Dufour (1982), we see that for any $m \in \{1, 2, \cdots, M\}$, we have

$$\hat{\theta}_m = \hat{\theta}_m^{-n} + (X_m^{'}X_m)^{-1}\pi_m^{'}x_n(y_n - x_n^{'}\pi_m\hat{\theta}_m^{-n}).$$

It follows from (2) that

$$\begin{aligned}
&E_S\Big[\max_{1\le m\le M}\|\hat{\theta}_m-\hat{\theta}_m^{-n}\|_2^2\Big]\\
&=E_S\Big[\max_{1\le m\le M}x_n^{'}\pi_m(X_m^{'}X_m)^{-1}(X_m^{'}X_m)^{-1}\pi_m^{'}x_n(y_n-x_n^{'}\pi_m\hat{\theta}_m^{-n})^2\Big]\\
&\le E_S\Big[\max_{1\le m\le M}\lambda_{max}^2[(X_m^{'}X_m)^{-1}]\|\pi_m^{'}x_n(y_n-x_n^{'}\pi_m\hat{\theta}_m^{-n})\|_2^2\Big]\\
&\le E_S\Big[C_1^{-2}n^{-2}\max_{1\le m\le M}\|\pi_m^{'}x_n(y_n-x_n^{'}\pi_m\hat{\theta}_m^{-n})\|_2^2\Big]. \qquad (5)
\end{aligned}$$

From Assumption 2, we obtain

$$E[(y^*)^2]=E[n^{-1}y^{'}y]\le C_2$$

and

$$\begin{aligned}
&E_S\Big[\max_{1\le m\le M}\|\pi_m^{'}x_n(y_n-x_n^{'}\pi_m\hat{\theta}_m^{-n})\|_2^2\Big]\\
&\le E_S\Big[\max_{1\le m\le M}\sum_{k=1}^{K}x_{(k)n}^2(y_n-x_n^{'}\pi_m\hat{\theta}_m^{-n})^2\Big]\\
&\le C_3^2KE_S\Big[\max_{1\le m\le M}(y_n-x_n^{'}\pi_m\hat{\theta}_m^{-n})^2]\\
&\le C_3^2KE_S\Big[\max_{1\le m\le M}[2y_n^2+2(x_n^{'}\pi_m\hat{\theta}_m^{-n})^2]\Big]\\
&\le C_3^2K\Big\{2C_2+2E_S\Big[\max_{1\le m\le M}(x_n^{'}\pi_m\hat{\theta}_m^{-n})^2\Big]\Big\}. \qquad (6)
\end{aligned}$$

Further, from Assumption 2 and Lemma 20, we have

$$\begin{aligned}
&E_S\Big[\max_{1\le m\le M}(x_n^{'}\pi_m\hat{\theta}_m^{-n})^2\Big]\\
&\le E_S\Big(\max_{1\le m\le M}\|x_n\|_2^2\|\hat{\theta}_m^{-n}\|_2^2\Big)\\
&\le C_3^2KE_S\Big(\max_{1\le m\le M}\|\hat{\theta}_m^{-n}\|_2^2\Big)\\
&\le B_1C_3^2K. \qquad (7)
\end{aligned}$$

Combining (5)-(7), it is seen that

$$E_S\Big(\max_{1\le m\le M}\|\hat{\theta}_m-\hat{\theta}_m^{-n}\|_2^2\Big)=O(n^{-2}K^2).$$

In a similar way, it can be shown that

$$E_S\Big(\max_{1\le m\le M}\|\hat{\theta}_m^{-n}-\hat{\theta}_m^{-(n-1,n)}\|_2^2\Big)=O(n^{-2}K^2)$$

by using (1). ■

The following lemma shows that under certain conditions, $n^{-1}\hat{\Omega}'\hat{\Omega}$ and $\hat{\theta}\hat{\theta}'$ are positive definite when the m-th model contains the first m covariates (which is a common nested model setting in model averaging literature, such as Hansen (2007) and Zhang and Liu (2019). Let $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$, $\hat{y}_m = P_m y$ and $\hat{\sigma}_m^2 = n^{-1}\|y - \hat{y}_m\|_2^2$. Specifically, this lemma requires the following two conditions: 1) $n^{-1}\hat{y}_1'\hat{y}_1 > 0$, a.s.; 2) $\hat{\sigma}_1^2 > \hat{\sigma}_2^2 > \cdots > \hat{\sigma}_M^2$, a.s.. Assume the first model contains only the constant term, then $n^{-1}\hat{y}_1'\hat{y}_1 = \bar{y}^2$, and therefore the first condition only needs $\bar{y} \neq 0$, a.s.. It follows from

$$\hat{\sigma}_1^2 = \|y - X_M\pi_1\hat{\theta}_1\|_2^2/n = \min_{\theta:\theta_2=\theta_3=\cdots=\theta_M=0} \|y - X_M\theta\|_2^2/n,$$

$$\hat{\sigma}_2^2 = \|y - X_M\pi_2\hat{\theta}_2\|_2^2/n = \min_{\theta:\theta_3=\theta_4=\cdots=\theta_M=0} \|y - X_M\theta\|_2^2/n,$$

$$\vdots$$

$$\hat{\sigma}_M^2 = \|y - X_M\pi_M\hat{\theta}_M\|_2^2/n = \min_{\theta} \|y - X_M\theta\|_2^2/n$$

that

$$\hat{\sigma}_1^2 \geq \hat{\sigma}_2^2 \geq \cdots \geq \hat{\sigma}_M^2.$$

On the other hand, if $x_2, x_3, ..., x_M$ are all significant covariates, then the second condition is reasonable. Additionally, when $M = n$, we know that $\hat{\sigma}_M^2 = 0$. Therefore, even if $x_2, x_3, ..., x_M$ contain insignificant covariates, the second condition may still be satisfied due to the ever-expanding parameter space.

Lemma 22 *When the m-th model contains the first m covariates, if $n^{-1}\hat{y}_1'\hat{y}_1 > 0$ and $\hat{\sigma}_1^2 > \hat{\sigma}_2^2 > \cdots > \hat{\sigma}_M^2$, a.s., then $n^{-1}\hat{\Omega}'\hat{\Omega}$ and $\hat{\theta}\hat{\theta}'$ are positive definite, a.s..*

Proof From the proof of Lemma 2 of Hansen (2007), we have

$$\begin{aligned}
&\hat{\Omega}'\hat{\Omega} \\
&= (y'P_mP_ty)_{M\times M} \\
&= \begin{pmatrix}
y'P_1y & y'P_1y & y'P_1y & \cdots & y'P_1y \\
y'P_1y & y'P_2y & y'P_2y & \cdots & y'P_2y \\
y'P_1y & y'P_2y & y'P_3y & \cdots & y'P_3y \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
y'P_1y & y'P_2y & y'P_3y & \cdots & y'P_My
\end{pmatrix}.
\end{aligned}$$

After a series of elementary row transformations, we see that

$$\begin{aligned}
&\hat{\Omega}'\hat{\Omega} \\
&\to \begin{pmatrix}
y'P_1y & y'P_1y & y'P_1y & \cdots & y'P_1y \\
0 & y'(P_2-P_1)y & y'(P_2-P_1)y & \cdots & y'(P_2-P_1)y \\
0 & y'(P_2-P_1)y & y'(P_3-P_1)y & \cdots & y'(P_3-P_1)y \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & y'(P_2-P_1)y & y'(P_3-P_1)y & \cdots & y'(P_M-P_1)y
\end{pmatrix}
\end{aligned}$$

$$
\begin{aligned}
&\to \begin{pmatrix}
y^{'}P_1y & y^{'}P_1y & y^{'}P_1y & \cdots & y^{'}P_1y \\
0 & y^{'}(P_2-P_1)y & y^{'}(P_2-P_1)y & \cdots & y^{'}(P_2-P_1)y \\
0 & 0 & y^{'}(P_3-P_2)y & \cdots & y^{'}(P_3-P_2)y \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & y^{'}(P_3-P_2)y & \cdots & y^{'}(P_M-P_2)y
\end{pmatrix} \\
&\to \begin{pmatrix}
y^{'}P_1y & y^{'}P_1y & y^{'}P_1y & \cdots & y^{'}P_1y \\
0 & y^{'}(P_2-P_1)y & y^{'}(P_2-P_1)y & \cdots & y^{'}(P_2-P_1)y \\
0 & 0 & y^{'}(P_3-P_2)y & \cdots & y^{'}(P_3-P_2)y \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & y^{'}(P_M-P_{M-1})y
\end{pmatrix} \\
&\to \begin{pmatrix}
\hat{y}_1^{'}\hat{y}_1 & \hat{y}_1^{'}\hat{y}_1 & \hat{y}_1^{'}\hat{y}_1 & \cdots & \hat{y}_1^{'}\hat{y}_1 \\
0 & \hat{y}_2^{'}\hat{y}_2-\hat{y}_1^{'}\hat{y}_1 & \hat{y}_2^{'}\hat{y}_2-\hat{y}_1^{'}\hat{y}_1 & \cdots & \hat{y}_2^{'}\hat{y}_2-\hat{y}_1^{'}\hat{y}_1 \\
0 & 0 & \hat{y}_3^{'}\hat{y}_3-\hat{y}_2^{'}\hat{y}_2 & \cdots & \hat{y}_3^{'}\hat{y}_3-\hat{y}_2^{'}\hat{y}_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \hat{y}_M^{'}\hat{y}_M-\hat{y}_{M-1}^{'}\hat{y}_{M-1}
\end{pmatrix}.
\end{aligned}
$$

It follows from

$$
\begin{aligned}
&\hat{\sigma}_m^2/n \\
&=(y^{'}y+\hat{y}_m^{'}\hat{y}_m-2y^{'}\hat{y}_m)/n \\
&=(y^{'}y+\hat{y}_m^{'}\hat{y}_m-2y^{'}P_my)/n \\
&=(y^{'}y-\hat{y}_m^{'}\hat{y}_m)/n
\end{aligned}
$$

that $\hat{\sigma}_{m-1}^2-\hat{\sigma}_m^2>0$ is equivalent to $(\hat{y}_m^{'}\hat{y}_m-\hat{y}_{m-1}^{'}\hat{y}_{m-1})/n>0$. Thus, $n^{-1}\hat{\Omega}^{'}\hat{\Omega}$ is positive definite, a.s.. Further, it follows from $\hat{\Omega}^{'}\hat{\Omega}=\hat{\theta}^{'}X_M^{'}X_M\hat{\theta}$ and $M=K$ that $\hat{\theta}\hat{\theta}^{'}$ is positive definite a.s..

■

In the following, we provide an example to illustrate the rationality of Assumption 3 based on Lemma 22.

Lemma 23 *Consider the nested setting of Lemma 22 with*

- *Assumption 2 holds,* $x_{i1}\equiv 1$ *and* $|\bar{y}|\geq c_0>0$, *a.s.,*
- $E(e_i|x_i)=0$ *and* $E(|e_i||x_i)<\infty$,
- $b_M^{'}b_M=o(n)$ *and* $\mu^{'}\mu=O(n)$,
- $0<c_1\leq \min_{1\leq k\leq K}|\theta_{M,(k)}|\leq \max_{1\leq k\leq K}|\theta_{M,(k)}|\leq c_2<\infty$;
- *the orthogonal design, i.e.* $n^{-1}X_M^{'}X_M=I_K$,

then, there is a constant $c_3 > 0$ such that $|n^{-1}\hat{\Omega}'\hat{\Omega}| \sim c_3^M$, a.s., and when M is bounded, there is a constant $c_4 > 0$ such that $\lambda_{min}(n^{-1}\hat{\Omega}'\hat{\Omega}) \sim c_4$, a.s., where $a_n \sim b_n$ means $a_n/b_n \to 1$ and $b_n/a_n \to 1$.

Proof Denote X_m^c as a matrix consisting of columns of X_M not contained in X_m, and π_m^c as a selection matrix such that $X_m^c = X_M\pi_m^c$. Let $\theta_m^c = \pi_m^{c'}\theta_M$ and $\xi_i(w) = \sum_{w\in W} w_m x_{(m)i}e_i$, where $W = \{w \in [0,1]^M : \sum_{m=1}^M w_m = 1\}$. From

- W is a totally bounded metric space with metric L_1-norm,
- for any $w \in W$, $n^{-1}\sum_{i=1}^n \xi_i(w) = o(1)$, a.s.,
- for any $w^1, w^2 \in W$, $|\xi_i(w^1) - \xi_i(w^2)| \le \big(\max_{1\le m\le M}|x_{(m)i}e_i|\big)\|w^1 - w^2\|_1$,

and Theorem 3 of Andrews (1992), we have $sup_{w\in W}|n^{-1}\sum_{i=1}^n \xi_i(w)| = o(1)$, a.s., and then

$$\begin{aligned}
&n^{-1}e'(X_mX_m' - X_{m-1}X_{m-1}')e\\
&= n^{-1}\sum_{i=1}^n\sum_{j=1}^n x_{(m)i}x_{(m)j}e_ie_j\\
&= n\Big(n^{-1}\sum_{i=1}^n x_{(m)i}e_i\Big)^2\\
&= o(n),\ a.s.
\end{aligned}$$

uniformly holds for $m \in \{1, 2, \cdots, M\}$. So from $(P_m - P_{m-1})(P_m - P_{m-1}) = P_m - P_{m-1}$, we obtain

$$\begin{aligned}
|\Lambda_{m,m-1}| =&|b_M'(P_m - P_{m-1})b_M + 2b_M'(X_{m-1}^c\theta_{m-1}^c - X_m^c\theta_m^c)\\
&+ e'(P_m - P_{m-1})e + 2\mu'(P_m - P_{m-1})e|\\
=&|b_M'(P_m - P_{m-1})b_M + 2b_M'X_M(\pi_{m-1}^c\theta_{m-1}^c - \pi_m^c\theta_m^c)\\
&+ e'(P_m - P_{m-1})e + 2\mu'(P_m - P_{m-1})e|\\
\le& b_M'(P_m - P_{m-1})b_M + 2\sqrt{b_M'b_M}\sqrt{n\theta_{M,(m)}^2} + n^{-1}e'(X_mX_m' - X_{m-1}X_{m-1}')e\\
&+ 2\sqrt{\mu'\mu}\sqrt{e'(P_m - P_{m-1})e}\\
\le& b_M'b_M + 2\sqrt{b_M'b_M}\sqrt{nc_2^2} + n^{-1}e'(X_mX_m' - X_{m-1}X_{m-1}')e\\
&+ 2\sqrt{\mu'\mu}\sqrt{n^{-1}e'(X_mX_m' - X_{m-1}X_{m-1}')e}\\
=& o(n),\ a.s.
\end{aligned}$$

uniformly holds for $m \in \{2, 3, \cdots, M\}$. On the other hand, it is seen that

$$\begin{aligned}
n\hat{\sigma}_m^2 =& (X_m\theta_m + X_m^c\theta_m^c + b_M)'(I - P_m)(X_m\theta_m + X_m^c\theta_m^c + b_M)\\
&+ e'(I - P_m)e + 2\mu'(I - P_m)e\\
=& (X_m^c\theta_m^c + b_M)'(I - P_m)(X_m^c\theta_m^c + b_M) + e'(I - P_m)e + 2\mu'(I - P_m)e
\end{aligned}$$

$$= n\theta_m^{c'}\theta_m^c + b_M^{'}(I-P_m)b_M + 2b_M^{'}X_m^c\theta_m^c + e^{'}(I-P_m)e + 2\mu^{'}(I-P_m)e.$$

Therefore,

$$\begin{aligned} & n(\hat{\sigma}_{m-1}^2 - \hat{\sigma}_m^2) \\ & = n(\theta_{m-1}^{c'}\theta_{m-1}^c - \theta_m^{c'}\theta_m^c) + \Lambda_{m,m-1} \\ & = n\theta_{M,(m)}^2 + \Lambda_{m,m-1} \\ & \geq nc_1^2 + o(n),\ a.s.. \end{aligned}$$

Thus, for large n, there is a constant $c_3 > 0$ such that

$$\min\left[\bar{y}^2, \min_{2\leq m\leq M}(\hat{\sigma}_{m-1}^2 - \hat{\sigma}_m^2)\right] \geq c_3 + o(1),\ a.s..$$

From

$$n^{-1}\hat{\Omega}'\hat{\Omega} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \bar{y}^2 & \bar{y}^2 & \bar{y}^2 & \cdots & \bar{y}^2 \\ 0 & \hat{\sigma}_1^2 - \hat{\sigma}_2^2 & \hat{\sigma}_1^2 - \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_1^2 - \hat{\sigma}_2^2 \\ 0 & 0 & \hat{\sigma}_2^2 - \hat{\sigma}_3^2 & \cdots & \hat{\sigma}_2^2 - \hat{\sigma}_3^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \hat{\sigma}_{M-1}^2 - \hat{\sigma}_M^2 \end{pmatrix},$$

we obtain $|n^{-1}\hat{\Omega}'\hat{\Omega}| \geq [c_3 + o(1)]^M \sim c_3^M$, a.s., that is, $n^{-1}\hat{\Omega}'\hat{\Omega}$ is positive definite, a.s.. When M is bounded, there is obviously a constant $c_3 > 0$ such that $\lambda_{min}(n^{-1}\hat{\Omega}'\hat{\Omega}) \sim c_3$, a.s..

■

The following lemma is used to illustrate the rationality of Assumption 3. Understanding its proof process is helpful for us to understand the proofs of Lemma 27.

Lemma 24 *Under Assumptions 1 and 2, we have*

$$E_{S,z^*}(\hat{\gamma}^{'}\hat{\gamma}) = n^{-1}E_S(\hat{\Omega}'\hat{\Omega}) + O(n^{-1}K^2)$$

and

$$n^{-1}E_S(\bar{\Omega}'\bar{\Omega}) = n^{-1}E_S(\hat{\Omega}'\hat{\Omega}) + O[n(C_1 n - C_3^2 K)^{-2}K].$$

Proof Note that

$$\hat{\gamma}^{'}\hat{\gamma} = (x^{*'}\pi_m\hat{\theta}_m\hat{\theta}_t'\pi_t^{'}x^*)_{M\times M},$$

$$\bar{\Omega}'\bar{\Omega} = \left\{[y - D_m(y - P_m y)]^{'}[y - D_t(y - P_t y)]\right\}_{M\times M}$$

and

$$\hat{\Omega}'\hat{\Omega} = (y^{'}P_m^{'}P_t y)_{M\times M} = \Big(\sum_{i=1}^{n} x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_t'\pi_t^{'}x_i\Big)_{M\times M}.$$

It follows from Assumption 2, and Lemmas 20 and 21 that

$$
\begin{aligned}
&|E_S(x_n^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_n - x_n^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_n)| \\
&\le |E_S[x_n^{'}\pi_m(\hat{\theta}_m - \hat{\theta}_m^{-n})\hat{\theta}_t^{'}\pi_t^{'}x_n]| + |E_S[x_n^{'}\pi_m\hat{\theta}_m^{-n}(\hat{\theta}_t - \hat{\theta}_t^{-n})^{'}\pi_t^{'}x_n]| \\
&\le \sqrt{E_S(\|\hat{\theta}_m - \hat{\theta}_m^{-n}\|_2^2)}\sqrt{E_S(\|x_n^{'}\pi_t\hat{\theta}_t\pi_m^{'}x_n\|_2^2)} + \sqrt{E_S(\|x_n^{'}\pi_m\hat{\theta}_m^{-n}\pi_t^{'}x_n\|_2^2)}\sqrt{E_S(\|\hat{\theta}_t - \hat{\theta}_t^{-n}\|_2^2)} \\
&\le \sqrt{E_S(\|\hat{\theta}_m - \hat{\theta}_m^{-n}\|_2^2)}\sqrt{E_S(\|x_n\|_2^2\|\hat{\theta}_t\|_2^2\|x_n\|_2^2)} + \sqrt{E_S(\|x_n\|_2^2\|\hat{\theta}_m^{-n}\|_2^2\|x_n\|_2^2)}\sqrt{E_S(\|\hat{\theta}_t - \hat{\theta}_t^{-n}\|_2^2)} \\
&= O(n^{-1}K^2).
\end{aligned}
$$

In a similar way, we obtain

$$
|E_{S,z^*}(x^{*'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x^* - x^{*'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x^*)| = O(n^{-1}K^2).
$$

Further, it is readily seen that

$$
\begin{aligned}
&E_{S,z^*}(\frac{1}{n}\sum_{i=1}^{n} x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_i - x^{*'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x^*) \\
&= E_{S,z^*}(x_n^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_n - x^{*'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x^*) \\
&= E_{S,z^*}(x_n^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_n - x^{*'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x^*) + E_{S,z^*}(x^{*'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x^* - x^{*'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x^*) \\
&= E_S(x_n^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_n - x_n^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_n) + E_{S,z^*}(x^{*'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x^* - x^{*'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x^*),
\end{aligned}
$$

so we have

$$
E_{S,z^*}(\hat{\gamma}^{'}\hat{\gamma}) = n^{-1}E_S[\hat{\Omega}^{'}\hat{\Omega}] + O(n^{-1}K^2).
$$

On the other hand, it follows from (2) and Assumption 2 that

$$
\max_{1\le i\le n}\max_{1\le m\le M} h_{ii}^m = \max_{1\le i\le n}\max_{1\le m\le M} x_i^{'}\pi_m(X_m^{'}X_m)^{-1}\pi_m^{'}x_i \le \frac{C_3^2K}{nC_1}, a.s.. \tag{8}
$$

Hence, from Assumption 2, we have

$$
\begin{aligned}
&|y^{'}P_m^{'}P_ty - [y - D_m(y - P_my)]^{'}[y - D_t(y - P_ty)]| \\
&= |y^{'}P_m^{'}P_ty - [(I_n - D_m)y + D_mP_my]^{'}[(I_n - D_t)y + D_tP_ty]| \\
&\le |y^{'}P_m^{'}(D_mD_t - I_n)P_ty| + |y^{'}(I_n - D_m)(I_n - D_t)y| + |y^{'}(I_n - D_m)D_tP_ty| \\
&\quad + |y^{'}P_m^{'}D_m(I_n - D_t)y| \\
&\le \max_{1\le m\le M}\max_{1\le i\le n}[(1 - h_{ii}^m)^{-2} - 1]y^{'}P_m^{'}P_my + \max_{1\le m\le M}\max_{1\le i\le n}[1 - (1 - h_{ii}^m)^{-1}]^2y^{'}y \\
&\quad + 2\max_{1\le m\le M}\max_{1\le i\le n}\sqrt{(1 - h_{ii}^m)^{-2}[1 - (1 - h_{ii}^m)^{-1}]^2y^{'}yy^{'}P_m^{'}P_my} \\
&\le \max_{1\le m\le M}\max_{1\le i\le n}\{[(1 - h_{ii}^m)^{-2} - 1] + [(1 - h_{ii}^m)^{-1} - 1]^2 \\
&\quad + 2(1 - h_{ii}^m)^{-1}[(1 - h_{ii}^m)^{-1} - 1]\}y^{'}y \\
&\le C_2n\{[(1 - \frac{C_3^2K}{nC_1})^{-2} - 1] + [(1 - \frac{C_3^2K}{nC_1})^{-1} - 1]^2
\end{aligned}
$$

$$+ 2(1 - \frac{C_3^2 K}{nC_1})^{-1}[(1 - \frac{C_3^2 K}{nC_1})^{-1} - 1]\}$$
$$= \frac{4C_1 C_2 C_3^2 n^2 K}{(C_1 n - C_3^2 K)^2}, \ a.s.. \tag{9}$$

Thus, we obtain

$$n^{-1} E_S(\bar{\Omega}'\bar{\Omega}) = n^{-1} E_S(\hat{\Omega}'\hat{\Omega}) + O[n(C_1 n - C_3^2 K)^{-2} K].$$

■

Let $\tilde{w}^{-n} = argmin_{w \in R^M} \hat{F}(w, S^{-n})$, where

$$\hat{F}(w, S^{-n}) = \frac{1}{n} \sum_{i=1}^{n-1} [y_i - x_i' \hat{\theta}^{-n}(w)]^2.$$

The following lemma shows that L_2-norm of the estimator of model weight vector is bounded, which makes Assumption 4 reasonable. In subsequent proofs, we use this conclusion many times.

Lemma 25 *Under Assumptions 2-3, there is a constant $B_2 > 0$ such that*

$$\max(\|\hat{w}\|_2^2, \|\hat{w}^{-n}\|_2^2) \leq B_2(1 + n^{-2} K^2 M), \ a.s.,$$
$$\max(\|\bar{w}\|_2^2, \|\bar{w}^{-n}\|_2^2) \leq B_2, \ a.s.$$

and

$$\max(\|\tilde{w}\|_2^2, \|\tilde{w}^{-n}\|_2^2) \leq B_2, \ a.s..$$

Proof It is straightforward to check that

$$\lambda_{max}[\hat{\Omega}(\hat{\Omega}'\hat{\Omega} + \lambda_n I_n)^{-1}(\hat{\Omega}'\hat{\Omega} + \lambda_n I_n)^{-1}\hat{\Omega}'] \leq [\lambda_{min}(\hat{\Omega}'\hat{\Omega} + \lambda_n I_n)]^{-1}.$$

Similarly, we can prove that

$$\lambda_{max}[\bar{\Omega}(\bar{\Omega}'\bar{\Omega} + \lambda_n I_n)^{-1}(\bar{\Omega}'\bar{\Omega} + \lambda_n I_n)^{-1}\bar{\Omega}'] \leq [\lambda_{min}(\bar{\Omega}'\bar{\Omega} + \lambda_n I_n)]^{-1}.$$

It follows from Assumptions 2-3 that

$$\begin{aligned}
\|\hat{w}\|_2^2 &= \|(\hat{\Omega}'\hat{\Omega} + \lambda_n I_n)^{-1}\hat{\Omega}' y - \sigma^2(\hat{\Omega}'\hat{\Omega} + \lambda_n I_n)^{-1}\kappa\|_2^2 \\
&\leq 2y'\hat{\Omega}(\hat{\Omega}'\hat{\Omega} + \lambda_n I_n)^{-1}(\hat{\Omega}'\hat{\Omega} + \lambda_n I_n)^{-1}\hat{\Omega}' y + 2\sigma^4 \kappa'(\hat{\Omega}'\hat{\Omega} + \lambda_n I_n)^{-1}(\hat{\Omega}'\hat{\Omega} + \lambda_n I_n)^{-1}\kappa \\
&\leq 2C_4^{-1} n^{-1} y' y + 2\sigma^4 C_4^{-2} n^{-2} \kappa'\kappa \\
&\leq 2C_4^{-1} C_2 + 2\sigma^4 C_4^{-2} n^{-2} K^2 M, \ a.s.,
\end{aligned}$$

and

$$\|\bar{w}\|_2^2 = \|(\bar{\Omega}'\bar{\Omega} + \lambda_n I_n)^{-1}\bar{\Omega}' y\|_2^2$$

$$\leq C_5^{-1} n^{-1} y' y$$
$$\leq C_4^{-1} C_2, \ a.s..$$

In a similar way, we obtain

$$\|\tilde{w}\|_2^2 \leq C_4^{-1} C_2, \ a.s..$$

From Assumption 3 and (4), we complete the proof by taking $B_2 = \max\{2C_4^{-1}C_2, 2\sigma^4 C_4^{-2}\}$. ■

The following three lemmas characterize the degree of impact of removing an observation on $\hat{w}$, $\bar{w}$ and $\hat{w}^*$, and are used to prove Theorem 15.

Lemma 26 *Under Assumptions 1-5, we have*

$$E_S[\|\hat{w} - \hat{w}^{-n}\|_2^2] = O(n^{-2}K^4M^2).$$

Proof Let $\hat{\theta}^{-n} = (\pi_1\hat{\theta}_1^{-n}, \pi_2\hat{\theta}_2^{-n}, \cdots, \pi_M\hat{\theta}_M^{-n})$. Then we have

$$\begin{aligned}\hat{w} &= (\hat{\Omega}'\hat{\Omega} + \lambda_n I_M)^{-1}(\hat{\Omega}' y - \sigma^2\kappa) \\ &= (\hat{\theta}' X_M' X_M \hat{\theta} + \lambda_n I_M)^{-1}(\hat{\theta}' X_M' y - \sigma^2\kappa)\end{aligned}$$

and

$$\begin{aligned}\hat{w}^{-n} &= (\hat{\Omega}^{-n'}\hat{\Omega}^{-n} + \lambda_{n-1} I_M)^{-1}(\hat{\Omega}^{-n'} y^{-n} - \sigma^2\kappa) \\ &= (\hat{\theta}^{-n'} X_M^{-n'} X_M^{-n} \hat{\theta}^{-n} + \lambda_{n-1} I_M)^{-1}(\hat{\theta}^{-n'} X_M^{-n'} y^{-n} - \sigma^2\kappa).\end{aligned}$$

Thus, under Assumption 3, we obtain

$$\begin{aligned}&(nC_4 + \lambda_n)^2 E_S[\|\hat{w} - \hat{w}^{-n}\|_2^2] \\ &\leq E_S[\lambda_{min}^2(\hat{\Omega}'\hat{\Omega} + \lambda_n I_M)\|\hat{w} - \hat{w}^{-n}\|_2^2] \\ &\leq E_S[\|(\hat{\Omega}'\hat{\Omega} + \lambda_n I_M)(\hat{w} - \hat{w}^{-n})\|_2^2] \\ &\leq 2E_S[\|(\hat{\Omega}'\hat{\Omega} + \lambda_n I_M)\hat{w} - (\hat{\theta}^{-n'} X_M^{-n'} X_M^{-n} \hat{\theta}^{-n} + \lambda_{n-1} I_M)\hat{w}^{-n}\|_2^2] \\ &\quad + 2E_S[\|(\hat{\theta}^{-n'} X_M^{-n'} X_M^{-n} \hat{\theta}^{-n} + \lambda_{n-1} I_M)\hat{w}^{-n} - (\hat{\Omega}'\hat{\Omega} + \lambda_n I_M)\hat{w}^{-n}\|_2^2] \\ &= 2E_S[\|\hat{\theta}' X_M' y - \hat{\theta}^{-n'} X_M^{-n'} y^{-n}\|_2^2] \\ &\quad + 2E_S[\|(\hat{\Omega}^{-n'}\hat{\Omega}^{-n} + \lambda_{n-1} I_M - \hat{\Omega}'\hat{\Omega} - \lambda_n I_M)\hat{w}^{-n}\|_2^2] \\ &\leq 2E_S[\|\hat{\theta}' X_M' y - \hat{\theta}^{-n'} X_M^{-n'} y^{-n}\|_2^2] \\ &\quad + 4E_S[\|(\hat{\Omega}^{-n'}\hat{\Omega}^{-n} - \hat{\Omega}'\hat{\Omega})\hat{w}^{-n}\|_2^2] + 4(\lambda_{n-1} - \lambda_n)^2 E_S(\|\hat{w}^{-n}\|_2^2). \qquad (10)\end{aligned}$$

We now consider the first term on the right-hand side of (10). From (2), Assumption 2, Lemma 20 and (3), we have

$$E_S[\max_{1\leq m\leq M} |(\hat{\theta}_m - \hat{\theta}_m^{-n})' X_m' y|^2]$$

$$
\begin{aligned}
&= E_S[\max_{1\le m\le M} |x_n^{'}\pi_m^{'}(X_m^{'}X_m)^{-1}X_m^{'}y(y_n - x_n^{'}\pi_m\hat{\theta}_m^{-n})|^2] \\
&\le E_S\{\max_{1\le m\le M} \lambda_{max}(\pi_m x_n x_n^{'}\pi_m^{'})\lambda_{max}[X_m(X_m^{'}X_m)^{-1}(X_m^{'}X_m)^{-1}X_m^{'}]\|y\|_2^2|y_n - x_n^{'}\pi_m\hat{\theta}_m^{-n}|^2\} \\
&\le C_1^{-1}C_2C_3^2KE_S(\max_{1\le m\le M} |y_n - x_n^{'}\pi_m\hat{\theta}_m^{-n}|^2) \\
&\le 2C_1^{-1}C_2C_3^2K[E_S(|y_n|^2) + E_S(\max_{1\le m\le M} |x_n^{'}\pi_m\hat{\theta}_m^{-n}|^2)] \\
&\le 2C_1^{-1}C_2C_3^2K[C_2 + E_S(\|x_n\|_2^2 \max_{1\le m\le M} \|\hat{\theta}_m^{-n}\|_2^2)] \\
&\le 2C_1^{-1}C_2C_3^2K[C_2 + C_3^2B_1K] \\
&= O(K^2)
\end{aligned}
$$

and

$$
\begin{aligned}
&E_S[\max_{1\le m\le M} |\hat{\theta}_m^{'}\pi_m^{'}X_M^{'}y - \hat{\theta}_m^{-n'}\pi_m^{'}X_M^{-n'}y^{-n}|^2] \\
&= E_S[\max_{1\le m\le M} |\hat{\theta}_m^{'}\pi_m^{'}X_M^{'}y - \hat{\theta}_m^{-n'}\pi_m^{'}X_M^{'}y + \hat{\theta}_m^{-n'}\pi_m^{'}X_M^{'}y - \hat{\theta}_m^{-n'}\pi_m^{'}X_M^{-n'}y^{-n}|^2] \\
&= E_S[\max_{1\le m\le M} |(\hat{\theta}_m - \hat{\theta}_m^{-n})^{'}X_m^{'}y + \hat{\theta}_m^{-n'}(X_m^{'}y - X_m^{-n'}y^{-n})|^2] \\
&\le 2E_S[\max_{1\le m\le M} |(\hat{\theta}_m - \hat{\theta}_m^{-n})^{'}X_m^{'}y|^2] + 2E_S[\max_{1\le m\le M} |\hat{\theta}_m^{-n'}(X_m^{'}y - X_m^{-n'}y^{-n})|^2] \\
&\le 2E_S[\max_{1\le m\le M} |(\hat{\theta}_m - \hat{\theta}_m^{-n})^{'}X_m^{'}y|^2] + 2E_S(\max_{1\le m\le M} \|\hat{\theta}_m^{-n}\|_2^2 \max_{1\le m\le M} \|X_m^{'}y - X_m^{-n'}y^{-n}\|_2^2) \\
&\le 2E_S[\max_{1\le m\le M} |(\hat{\theta}_m - \hat{\theta}_m^{-n})^{'}X_m^{'}y|^2] + 2B_1\sum_{k=1}^{K} E_S(|x_{(k)n}y_n|^2) \\
&\le 2E_S[\max_{1\le m\le M} |(\hat{\theta}_m - \hat{\theta}_m^{-n})^{'}X_m^{'}y|^2] + 2C_2C_3^2B_1K \\
&= O(K^2).
\end{aligned}
$$

Hence, we obtain

$$
\begin{aligned}
&E_S[\|\hat{\theta}^{'}X_M^{'}y - \hat{\theta}^{-n'}X_M^{-n'}y^{-n}\|_2^2] \\
&= \sum_{m=1}^{M} E_S(|\hat{\theta}_m^{'}\pi_m^{'}X_M^{'}y - \hat{\theta}_m^{-n'}\pi_m^{'}X_M^{-n'}y^{-n}|^2) \\
&= O(K^2M). \qquad (11)
\end{aligned}
$$

We then consider the second term on the right-hand side of (10). It follows from Assumption 2 and Lemmas 20-21 that

$$
\begin{aligned}
&E_S(\max_{1\le m\le M}\max_{1\le t\le M} |x_{n-1}^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_{n-1} - x_{n-1}^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_{n-1}|^2) \\
&\le 2E_S[\max_{1\le m\le M}\max_{1\le t\le M} |x_{n-1}^{'}\pi_m(\hat{\theta}_m - \hat{\theta}_m^{-n})\hat{\theta}_t^{'}\pi_t^{'}x_{n-1}|^2] \\
&\quad + 2E_S[\max_{1\le m\le M}\max_{1\le t\le M} |x_{n-1}^{'}\pi_m\hat{\theta}_m^{-n}(\hat{\theta}_t - \hat{\theta}_t^{-n})^{'}\pi_t^{'}x_{n-1}|^2] \\
&\le 2E_S(\max_{1\le m\le M} \|\hat{\theta}_m - \hat{\theta}_m^{-n}\|_2^2 \max_{1\le m\le M}\max_{1\le t\le M} \|x_{n-1}^{'}\pi_t\hat{\theta}_t\pi_m^{'}x_{n-1}\|_2^2)
\end{aligned}
$$

$$
\begin{aligned}
&\quad + 2E_S(\max_{1\le m\le M}\max_{1\le t\le M}\|x_{n-1}^{'}\pi_m\hat{\theta}_m^{-n}\pi_t^{'}x_{n-1}\|_2^2\max_{1\le t\le M}\|\hat{\theta}_t-\hat{\theta}_t^{-n}\|_2^2)\\
&\le 2E_S(\max_{1\le m\le M}\|\hat{\theta}_m-\hat{\theta}_m^{-n}\|_2^2\max_{1\le t\le M}\|x_{n-1}\|_2^2\|\hat{\theta}_t\|_2^2\|x_{n-1}\|_2^2)\\
&\quad + 2E_S(\max_{1\le m\le M}\|x_{n-1}\|_2^2\|\hat{\theta}_m^{-n}\|_2^2\|x_{n-1}\|_2^2\max_{1\le t\le M}\|\hat{\theta}_t-\hat{\theta}_t^{-n}\|_2^2)\\
&\le 4C_3^4B_1K^2E_S(\max_{1\le m\le M}\|\hat{\theta}_m-\hat{\theta}_m^{-n}\|_2^2)\\
&= O(n^{-2}K^4)
\end{aligned}
$$

and

$$
\begin{aligned}
&E_S\big\{\max_{1\le m\le M}\max_{1\le t\le M}\big|\sum_{i=1}^{n-1}(x_i^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_i-x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_i)-x_n^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_n\big|^2\big\}\\
&\le 2E_S\big\{\max_{1\le m\le M}\max_{1\le t\le M}\big|\sum_{i=1}^{n-1}(x_i^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_i-x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_i)\big|^2\big\}\\
&\quad + 2E_S\big\{\max_{1\le m\le M}\max_{1\le t\le M}\big|x_n^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_n\big|^2\big\}\\
&\le 2(n-1)\sum_{i=1}^{n-1}E_S\big\{\max_{1\le m\le M}\max_{1\le t\le M}\big|x_i^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_i-x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_i\big|^2\big\}\\
&\quad + 2E_S\big\{\max_{1\le m\le M}\max_{1\le t\le M}\big|x_n^{'}\pi_m\hat{\theta}_m|^2|\hat{\theta}_t^{'}\pi_t^{'}x_n\big|^2\big\}\\
&\le 2(n-1)\sum_{i=1}^{n-1}E_S\big\{\max_{1\le m\le M}\max_{1\le t\le M}\big|x_i^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_i-x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_i\big|^2\big\}\\
&\quad + 2E_S\big\{\max_{1\le m\le M}\|x_n\|_2^4\|\hat{\theta}_m\|_2^4\big\}\\
&\le 2(n-1)\sum_{i=1}^{n-1}E_S\big\{\max_{1\le m\le M}\max_{1\le t\le M}\big|x_i^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_i-x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_i\big|^2\big\}+2C_3^4B_1^2K^2\\
&= 2(n-1)^2E_S\big\{\max_{1\le m\le M}\max_{1\le t\le M}\big|x_{n-1}^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_{n-1}-x_{n-1}^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_{n-1}\big|^2\big\}+2C_3^4B_1^2K^2\\
&= O(K^4).
\end{aligned}
$$

So from Lemma 25,

$$
\hat{\Omega}^{'}\hat{\Omega}=(y^{'}P_m^{'}P_ty)_{M\times M}=(\hat{\theta}_m^{'}X_m^{'}X_t\hat{\theta}_t)_{M\times M}=\Big(\sum_{i=1}^{n}x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_i\Big)_{M\times M}
$$

and

$$
\hat{\Omega}^{-n'}\hat{\Omega}^{-n}=(\hat{\theta}_m^{-n'}X_m^{-n'}X_t^{-n}\hat{\theta}_t^{-n})_{M\times M}=\Big(\sum_{i=1}^{n-1}x_i^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_i\Big)_{M\times M},
$$

we have

$$
E_S[\|(\hat{\Omega}^{-n'}\hat{\Omega}^{-n}-\hat{\Omega}^{'}\hat{\Omega})\hat{w}^{-n}\|_2^2]
$$

$$
\begin{aligned}
&= E_S\Big[\sum_{m=1}^{M}\sum_{t=1}^{M}\hat{w}_m^{-n}\hat{w}_t^{-n}\sum_{s=1}^{M}(\sum_{i=1}^{n-1}x_i^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_s^{-n'}\pi_s^{'}x_i-\sum_{i=1}^{n}x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_s^{'}\pi_s^{'}x_i)\\
&\quad(\sum_{i=1}^{n-1}x_i^{'}\pi_s\hat{\theta}_s^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_i-\sum_{i=1}^{n}x_i^{'}\pi_s\hat{\theta}_s\hat{\theta}_t^{'}\pi_t^{'}x_i)\Big]\\
&\le E_S\Big[\sum_{m=1}^{M}\sum_{t=1}^{M}|\hat{w}_m^{-n}\hat{w}_t^{-n}|\max_{1\le m\le M}\max_{1\le t\le M}\Big|\sum_{s=1}^{M}(\sum_{i=1}^{n-1}x_i^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_s^{-n'}\pi_s^{'}x_i\\
&\quad-\sum_{i=1}^{n}x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_s^{'}\pi_s^{'}x_i)(\sum_{i=1}^{n-1}x_i^{'}\pi_s\hat{\theta}_s^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_i-\sum_{i=1}^{n}x_i^{'}\pi_s\hat{\theta}_s\hat{\theta}_t^{'}\pi_t^{'}x_i)\Big|\Big]\\
&\le M^2E_S\Big[\sum_{m=1}^{M}|\hat{w}_m^{-n}|^2\max_{1\le m\le M}\max_{1\le t\le M}\Big|\sum_{i=1}^{n-1}x_i^{'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x_i-\sum_{i=1}^{n}x_i^{'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x_i\Big|^2\Big]\\
&= O(K^4M^2). \qquad (12)
\end{aligned}
$$

Finally, combining (10)-(12) and using Assumption 5, we see that Lemma 26 is true. ■

Lemma 27 *Under Assumptions 1-5, we have*

$$E_S[\|\bar{w}-\bar{w}^{-n}\|_2^2]=O(n^{-2}K^4M^2).$$

Proof Note that

$$\bar{w}=(\bar{\Omega}^{'}\bar{\Omega}+\lambda_nI_M)^{-1}\bar{\Omega}^{'}y$$

and

$$\bar{w}^{-n}=(\bar{\Omega}^{-n'}\bar{\Omega}^{-n}+\lambda_{n-1}I_M)^{-1}\bar{\Omega}^{-n'}y^{-n}.$$

So under Assumption 3, we obtain

$$
\begin{aligned}
&(nC_4+\lambda_n)^2E_S[\|\bar{w}-\bar{w}^{-n}\|_2^2]\\
&\le E_S[\lambda_{min}^2(\bar{\Omega}^{'}\bar{\Omega}+\lambda_nI_M)\|\bar{w}-\bar{w}^{-n}\|_2^2]\\
&\le E_S[\|(\bar{\Omega}^{'}\bar{\Omega}+\lambda_nI_M)(\bar{w}-\bar{w}^{-n})\|_2^2]\\
&\le 2E_S[\|(\bar{\Omega}^{'}\bar{\Omega}+\lambda_nI_M)\bar{w}-(\bar{\Omega}^{-n'}\bar{\Omega}^{-n}+\lambda_{n-1}I_M)\bar{w}^{-n}\|_2^2]\\
&\quad+2E_S[\|(\bar{\Omega}^{-n'}\bar{\Omega}^{-n}+\lambda_{n-1}I_M)\bar{w}^{-n}-(\bar{\Omega}^{'}\bar{\Omega}+\lambda_nI_M)\bar{w}^{-n}\|_2^2]\\
&= 2E_S[\|\bar{\Omega}^{'}y-\bar{\Omega}^{-n'}y^{-n}\|_2^2]\\
&\quad+2E_S[\|(\bar{\Omega}^{-n'}\bar{\Omega}^{-n}+\lambda_{n-1}I_M-\bar{\Omega}^{'}\bar{\Omega}-\lambda_nI_M)\bar{w}^{-n}\|_2^2]\\
&\le 2E_S[\|\bar{\Omega}^{'}y-\bar{\Omega}^{-n'}y^{-n}\|_2^2]\\
&\quad+4E_S[\|(\bar{\Omega}^{-n'}\bar{\Omega}^{-n}-\bar{\Omega}^{'}\bar{\Omega})\bar{w}^{-n}\|_2^2]+4(\lambda_{n-1}-\lambda_n)^2E_S(\|\bar{w}^{-n}\|_2^2). \qquad (13)
\end{aligned}
$$

We now consider the first term on the right-hand side of (13). From Assumption 2, (8) and the definition of D_m, we have

$$\begin{aligned}
&\|\bar{\Omega}^{'}y - \hat{\Omega}^{'}y\|_2^2 \\
&= \sum_{m=1}^{M}\{[y - D_m(I - P_m)y]^{'}y - y^{'}P_m y\}^2 \\
&= \sum_{m=1}^{M}\{y^{'}(I_n - D_m)y + y^{'}P_m(D_m - I_n)y\}^2 \\
&\le 2M \max_{1\le m\le M}\left[|y^{'}(D_m - I_n)y|^2 + |y^{'}P_m(D_m - I_n)y|^2\right] \\
&\le 2M \max_{1\le m\le M}\{|\max_{1\le i\le n}[(1-h_{ii}^m)^{-1} - 1]y^{'}y|^2 + y^{'}(D_m - I_n)P_m y y^{'}P_m(D_m - I_n)y\} \\
&\le 4M \max_{1\le m\le M}\max_{1\le i\le n}[(1-h_{ii}^m)^{-1} - 1]^2 (y^{'}y)^2 \\
&\le 4C_2^2 n^2 M[(1 - \frac{C_3^2 K}{nC_1})^{-1} - 1]^2 \\
&= 4C_2^2 C_3^4 n^2 M K^2 (nC_1 - C_3^2 K)^{-2}, a.s..
\end{aligned}$$

Similarly, it follows from (1) and (4) that

$$\|\hat{\Omega}^{-n^{'}}y^{-n} - \bar{\Omega}^{-n^{'}}y^{-n}\|_2^2 \le 4C_2^2 C_3^4 n^2 M K^2 (nC_1 - C_3^2 K)^{-2}, a.s..$$

Hence, from Assumption 5 and (11), we obtain

$$\begin{aligned}
&E_S[\|\bar{\Omega}^{'}y - \bar{\Omega}^{-n^{'}}y^{-n}\|_2^2] \\
&= E_S[\|\bar{\Omega}^{'}y - \hat{\Omega}^{'}y + \hat{\Omega}^{'}y - \hat{\Omega}^{-n^{'}}y^{-n} + \hat{\Omega}^{-n^{'}}y^{-n} - \bar{\Omega}^{-n^{'}}y^{-n}\|_2^2] \\
&\le 3E_S[\|\bar{\Omega}^{'}y - \hat{\Omega}^{'}y\|_2^2] + 3E_S[\|\hat{\Omega}^{'}y - \hat{\Omega}^{-n^{'}}y^{-n}\|_2^2] + E_S[\|\hat{\Omega}^{-n^{'}}y^{-n} - \bar{\Omega}^{-n^{'}}y^{-n}\|_2^2] \\
&= O(K^2 M). \qquad (14)
\end{aligned}$$

We then consider the second term on the right-hand side of (13). From Assumption 5, Lemma 25, (1), (4), (9) and (12), we have

$$\begin{aligned}
&E_S[\|(\bar{\Omega}^{-n^{'}}\bar{\Omega}^{-n} - \bar{\Omega}^{'}\bar{\Omega})\bar{w}^{-n}\|_2^2] \\
&= E_S[\|(\bar{\Omega}^{-n^{'}}\bar{\Omega}^{-n} - \hat{\Omega}^{-n^{'}}\hat{\Omega}^{-n} + \hat{\Omega}^{-n^{'}}\hat{\Omega}^{-n} - \hat{\Omega}^{'}\hat{\Omega} + \hat{\Omega}^{'}\hat{\Omega} - \bar{\Omega}^{'}\bar{\Omega})\bar{w}^{-n}\|_2^2] \\
&\le 3E_S[\|(\bar{\Omega}^{-n^{'}}\bar{\Omega}^{-n} - \hat{\Omega}^{-n^{'}}\hat{\Omega}^{-n})\bar{w}^{-n}\|_2^2] + 3E_S[\|(\hat{\Omega}^{-n^{'}}\hat{\Omega}^{-n} - \hat{\Omega}^{'}\hat{\Omega})\bar{w}^{-n}\|_2^2] \\
&\quad + 3E_S[\|(\hat{\Omega}^{'}\hat{\Omega} - \bar{\Omega}^{'}\bar{\Omega})\bar{w}^{-n}\|_2^2] \\
&= O(K^4 M^2). \qquad (15)
\end{aligned}$$

Thus, under Assumption 5, this proof can be completed by combining (13)-(15). ■

Lemma 28 *Under Assumptions 1-5, we have*

$$E_S[\|\hat{w}^* - \hat{w}^{-n*}\|_2^2] = O(n^{-2}K^4M^2).$$

Proof Note that

$$\begin{aligned}\hat{w}^* &= [E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})]^{-1}E_{z^*}(\hat{\gamma}^{'}y^*)\\ &= [\hat{\theta}^{'}E_{z^*}(x^*x^{*'})\hat{\theta}]^{-1}\hat{\theta}^{'}E_{z^*}(x^*y^*)\end{aligned}$$

and

$$\hat{w}^{-n*} = [\hat{\theta}^{-n'}E_{z^*}(x^*x^{*'})\hat{\theta}^{-n}]^{-1}\hat{\theta}^{-n'}E_{z^*}(x^*y^*).$$

So under Assumption 3, we obtain

$$\begin{aligned}&C_4^2E_S(\|\hat{w}^* - \hat{w}^{-n*}\|_2^2)\\ &\le E_S\{\lambda_{min}^2[\hat{\theta}^{'}E_{z^*}(x^*x^{*'})\hat{\theta}]\|\hat{w}^* - \hat{w}^{-n*}\|_2^2\}\\ &\le E_S[\|\hat{\theta}^{'}E_{z^*}(x^*x^{*'})\hat{\theta}(\hat{w}^* - \hat{w}^{-n*})\|_2^2]\\ &\le 2E_S[\|\hat{\theta}^{'}E_{z^*}(x^*x^{*'})\hat{\theta}\hat{w}^* - \hat{\theta}^{-n'}E_{z^*}(x^*x^{*'})\hat{\theta}^{-n}\hat{w}^{-n*}\|_2^2]\\ &\quad + 2E_S[\|\hat{\theta}^{-n'}E_{z^*}(x^*x^{*'})\hat{\theta}^{-n}\hat{w}^{-n*} - \hat{\theta}^{'}E_{z^*}(x^*x^{*'})\hat{\theta}\hat{w}^{-n*}\|_2^2]\\ &= 2E_S[\|\hat{\theta}^{'}E_{z^*}(x^*y^*) - \hat{\theta}^{-n'}E_{z^*}(x^*y^*)\|_2^2]\\ &\quad + 2E_S[\|\hat{\theta}^{-n'}E_{z^*}(x^*x^{*'})\hat{\theta}^{-n}\hat{w}^{-n*} - \hat{\theta}^{'}E_{z^*}(x^*x^{*'})\hat{\theta}\hat{w}^{-n*}\|_2^2]. \qquad (16)\end{aligned}$$

We now consider the first term on the right-hand side of (16). From (2), Assumption 2 and Lemma 20, we have

$$\begin{aligned}&E_S[\|\hat{\theta}^{'}E_{z^*}(x^*y^*) - \hat{\theta}^{-n'}E_{z^*}(x^*y^*)\|_2^2]\\ &= \sum_{m=1}^{M}E_S[|(\hat{\theta}_m - \hat{\theta}_m^{-n})^{'}\pi_m^{'}E_{z^*}(x^*y^*)|^2]\\ &= \sum_{m=1}^{M}E_S[|x_n^{'}\pi_m(X_m^{'}X_m)^{-1}\pi_m^{'}E_{z^*}(x^*y^*)(y_n - x_n^{'}\pi_m\hat{\theta}_m^{-n})|^2]\\ &\le C_1^{-2}C_3^2n^{-2}K\sum_{m=1}^{M}E_S[\|\pi_m^{'}E_{z^*}(x^*y^*)(y_n - x_n^{'}\pi_m\hat{\theta}_m^{-n})\|^2]\\ &\le C_1^{-2}C_3^2n^{-2}K\sum_{m=1}^{M}E_S[(y_n - x_n^{'}\pi_m\hat{\theta}_m^{-n})^2]\sum_{k=1}^{K}[E_{z^*}(x_{(k)}^*y^*)]^2\\ &\le C_1^{-2}C_2C_3^4n^{-2}K^2\sum_{m=1}^{M}E_S[(y_n - x_n^{'}\pi_m\hat{\theta}_m^{-n})^2]\\ &\le 2C_1^{-2}C_2C_3^4n^{-2}K^2\sum_{m=1}^{M}[E_S(|y_n|^2) + E_S(|x_n^{'}\pi_m\hat{\theta}_m^{-n}|^2)]\\ &= O(n^{-2}K^3M). \qquad (17)\end{aligned}$$

It follows from Assumption 2 and Lemmas 20-21 that

$$E_S[\max_{1\le m\le M}\max_{1\le t\le M}E_{z^*}(|x^{*'}\pi_m\hat{\theta}_m\hat{\theta}_t^{'}\pi_t^{'}x^* - x^{*'}\pi_m\hat{\theta}_m^{-n}\hat{\theta}_t^{-n'}\pi_t^{'}x^*|^2)]$$

$$
\begin{aligned}
&\le 2E_S\{\max_{1\le m\le M}\max_{1\le t\le M} E_{z^*}[|x^{*'}\pi_m(\hat\theta_m-\hat\theta_m^{-n})\hat\theta_t'\pi_t'x^*|^2]\}\\
&\quad+2E_S\{\max_{1\le m\le M}\max_{1\le t\le M} E_{z^*}[|x^{*'}\pi_m\hat\theta_m^{-n}(\hat\theta_t-\hat\theta_t^{-n})'\pi_t'x^*|^2]\}\\
&\le 2E_S[\max_{1\le m\le M}\|\hat\theta_m-\hat\theta_m^{-n}\|_2^2\max_{1\le m\le M}\max_{1\le t\le M} E_{z^*}(\|x^{*'}\pi_t\hat\theta_t\pi_m'x^*\|_2^2)]\\
&\quad+2E_S[\max_{1\le m\le M}\max_{1\le t\le M} E_{z^*}(\|x^{*'}\pi_m\hat\theta_m^{-n}\pi_t'x^*\|_2^2)\max_{1\le t\le M}\|\hat\theta_t-\hat\theta_t^{-n}\|_2^2]\\
&\le 2E_S[\max_{1\le m\le M}\|\hat\theta_m-\hat\theta_m^{-n}\|_2^2\max_{1\le t\le M} E_{z^*}(\|x^*\|_2^2\|\hat\theta_t\|_2^2\|x^*\|_2^2)]\\
&\quad+2E_S[\max_{1\le m\le M} E_{z^*}(\|x^*\|_2^2\|\hat\theta_m^{-n}\|_2^2\|x^*\|_2^2)\max_{1\le t\le M}\|\hat\theta_t-\hat\theta_t^{-n}\|_2^2]\\
&\le 4C_3^4B_1K^2E_S(\max_{1\le m\le M}\|\hat\theta_m-\hat\theta_m^{-n}\|_2^2)\\
&= O(n^{-2}K^4).
\end{aligned}
$$

We then consider the second term on the right-hand side of (16). From Assumption 4,

$$\hat\gamma'\hat\gamma=(x^{*'}\pi_m\hat\theta_m\hat\theta_t'\pi_t'x^*)_{M\times M}$$

and

$$\hat\gamma^{-n'}\hat\gamma^{-n}=(x^{*'}\pi_m\hat\theta_m^{-n}\hat\theta_t^{-n'}\pi_t'x^*)_{M\times M},$$

we have

$$
\begin{aligned}
&E_S[\|\hat\theta^{-n'}E_{z^*}(x^*x^{*'})\hat\theta^{-n}\hat w^{-n*}-\hat\theta'E_{z^*}(x^*x^{*'})\hat\theta\hat w^{-n*}\|_2^2]\\
&=E_S\Big[\sum_{m=1}^M\sum_{t=1}^M\hat w_m^{-n*}\hat w_t^{-n*}\sum_{s=1}^M E_{z^*}(x^{*'}\pi_m\hat\theta_m\hat\theta_s'\pi_s'x^*-x^{*'}\pi_m\hat\theta_m^{-n}\hat\theta_s^{-n'}\pi_s'x^*)\\
&\quad E_{z^*}(x^{*'}\pi_s\hat\theta_s\hat\theta_t'\pi_t'x^*-x^{*'}\pi_s\hat\theta_s^{-n}\hat\theta_t^{-n'}\pi_t'x^*)\Big]\\
&\le M^2E_S\Big[\sum_{m=1}^M|\hat w_m^{-n*}|^2\max_{1\le m\le M}\max_{1\le t\le M}E_{z^*}(|x^{*'}\pi_m\hat\theta_m\hat\theta_t'\pi_t'x^*-x^{*'}\pi_m\hat\theta_m^{-n}\hat\theta_t^{-n'}\pi_t'x^*|^2)\Big]\\
&=O(n^{-2}K^4M^2).
\end{aligned}
\tag{18}
$$

Thus, under Assumption 5, this proof can be completed by combining (16)-(18). ■

The following lemma gives a case where Assumption 4 holds.

Lemma 29 *Assume* $y^*=\sum_{k=1}^K x_{(k)}^*\theta_k^*+e^*$ *with* $x^*=(x_{(1)}^*,x_{(2)}^*,\cdots,x_{(K)}^*)'\sim N(0,I_K)$ *and* $E(e^*|x^*)=0$. *If there is a constant* $C_{10}>0$ *such that* $\sum_{k=1}^K\theta_k^{*2}\le C_{10}$, *then under Assumption 3, Assumption 4 holds.*

Proof Similar to the proof of (3), it follows from Assumption 3, $\hat\gamma=x^{*'}\hat\theta$ and $E_{z^*}(x^*x^{*'})=I_K$ that

$$\lambda_{max}[\hat\theta(\hat\theta'\hat\theta)^{-2}\hat\theta']\le C_4^{-1},\ a.s..$$

And then, from the definition of $\hat{w}^*$, we have

$$\begin{aligned}
\|\hat{w}^*\|_2^2 &= \|[E_{z^*}(\hat{\gamma}'\hat{\gamma})]^{-1}E_{z^*}(\hat{\gamma}'y^*)\|_2^2 \\
&= \|[\hat{\theta}'E_{z^*}(x^*x^{*'})\hat{\theta}]^{-1}\hat{\theta}'E_{z^*}(x^*y^*)\|_2^2 \\
&= E_{z^*}(x^{*'}y^*)\hat{\theta}[\hat{\theta}'E_{z^*}(x^*x^{*'})\hat{\theta}]^{-2}\hat{\theta}'E_{z^*}(x^*y^*) \\
&\leq C_4^{-1}\sum_{k=1}^{K}\left[E_{z^*}(x^*_{(k)}y^*)\right]^2 \\
&= C_4^{-1}\sum_{k=1}^{K}\left\{E_{z^*}\left[x^*_{(k)}(\sum_{j=1}^{K}x^*_{(j)}\theta^*_j + e^*)\right]\right\}^2 \\
&= C_4^{-1}\sum_{k=1}^{K}\left[\sum_{j=1}^{K}\theta^*_j E_{z^*}(x^*_{(k)}x^*_{(j)}) + E_{z^*}(x^*_{(k)}e^*)\right]^2 \\
&= C_4^{-1}\sum_{k=1}^{K}\theta^{*2}_k \\
&\leq C_4^{-1}C_{10},\ a.s..
\end{aligned}$$

Similarly, we have $\|\hat{w}^{-n*}\|_2^2 \leq C_4^{-1}C_{10}$ a.s.. ■

Appendix B. Proofs of Theorems

Proof of Theorem 13: We first prove 1). Let $(\hat{c}_1, \hat{c}_2, ..., \hat{c}_M)' = \hat{P}'\hat{w}^0$ and $(\hat{d}_1, \hat{d}_2, ..., \hat{d}_M)' = \hat{P}'\hat{w}^*$. Then, we have

$$\begin{aligned}
\hat{M}_1(\lambda_n) &= \|\hat{Z}\hat{w}^0 - \hat{Z}\hat{w}^*\|_2^2 + \|\hat{Z}\hat{w}^* - \hat{w}^*\|_2^2 \\
&= \sum_{m=1}^{M}\frac{(\hat{c}_m - \hat{d}_m)^2\hat{\zeta}_m^2}{(\lambda_n + \hat{\zeta}_m)^2} + \sum_{m=1}^{M}\hat{d}_m^2(\frac{\hat{\zeta}_m}{\lambda_n + \hat{\zeta}_m} - 1)^2 \\
&= \sum_{m=1}^{M}\frac{(\hat{c}_m - \hat{d}_m)^2\hat{\zeta}_m^2}{(\lambda_n + \hat{\zeta}_m)^2} + \sum_{m=1}^{M}\frac{\hat{d}_m^2\lambda_n^2}{(\lambda_n + \hat{\zeta}_m)^2}
\end{aligned}$$

and

$$\frac{d}{d\lambda_n}\hat{M}_1(\lambda_n) = \sum_{m=1}^{M}\frac{-2(\hat{c}_m - \hat{d}_m)^2\hat{\zeta}_m^2}{(\lambda_n + \hat{\zeta}_m)^3} + \sum_{m=1}^{M}\frac{2\hat{d}_m^2\lambda_n\hat{\zeta}_m}{(\lambda_n + \hat{\zeta}_m)^3}.$$

From

$$\begin{aligned}
&\sum_{m=1}^{M}\frac{(\hat{c}_m - \hat{d}_m)^2\hat{\zeta}_m^2}{(\lambda_n + \hat{\zeta}_m)^3} \\
&\geq \frac{1}{\lambda_n + \hat{\zeta}_M}\sum_{m=1}^{M}\frac{(\hat{c}_m - \hat{d}_m)^2\hat{\zeta}_m^2}{(\lambda_n + \hat{\zeta}_m)^2}
\end{aligned}$$

$$= \frac{1}{\lambda_n + \hat{\zeta}_M} \|\hat{Z}(\hat{w}^0 - \hat{w}^*)\|_2^2,$$

we see that when $\hat{w}^0 \neq \hat{w}^*$, $\sum_{m=1}^{M} \frac{(\hat{c}_m - \hat{d}_m)^2 \hat{\zeta}_m^2}{\hat{\zeta}_m^3} > 0$. So we have $\hat{\lambda}_n > 0$ and $\hat{M}_1(\hat{\lambda}_n) < \hat{M}_1(0)$ when $\hat{w}^0 \neq \hat{w}^*$.

Below we prove 2). Let $(\bar{c}_1, \bar{c}_2, ..., \bar{c}_M)' = \bar{P}' \bar{w}^0$ and $(\bar{d}_1, \bar{d}_2, ..., \bar{d}_M)' = \bar{P}' \hat{w}^*$. Then, we have

$$\begin{aligned}
\bar{M}_1(\lambda_n) &= \|\bar{Z}\bar{w}^0 - \bar{Z}\hat{w}^*\|_2^2 + \|\bar{Z}\hat{w}^* - \hat{w}^*\|_2^2 \\
&= \sum_{m=1}^{M} \frac{(\bar{c}_m - \bar{d}_m)^2 \bar{\zeta}_m^2}{(\lambda_n + \bar{\zeta}_m)^2} + \sum_{m=1}^{M} \bar{d}_m^2 (\frac{\bar{\zeta}_m}{\lambda_n + \bar{\zeta}_m} - 1)^2 \\
&= \sum_{m=1}^{M} \frac{(\bar{c}_m - \bar{d}_m)^2 \bar{\zeta}_m^2}{(\lambda_n + \bar{\zeta}_m)^2} + \sum_{m=1}^{M} \frac{\bar{d}_m^2 \lambda_n^2}{(\lambda_n + \bar{\zeta}_m)^2}
\end{aligned}$$

and

$$\frac{d}{d\lambda_n} \bar{M}_1(\lambda_n) = \sum_{m=1}^{M} \frac{-2(\bar{c}_m - \bar{d}_m)^2 \bar{\zeta}_m^2}{(\lambda_n + \bar{\zeta}_m)^3} + \sum_{m=1}^{M} \frac{2\bar{d}_m^2 \lambda_n \bar{\zeta}_m}{(\lambda_n + \bar{\zeta}_m)^3}.$$

Similarly, from

$$\begin{aligned}
&\sum_{m=1}^{M} \frac{(\bar{c}_m - \bar{d}_m)^2 \bar{\zeta}_m^2}{(\lambda_n + \bar{\zeta}_m)^3} \\
&\geq \frac{1}{\lambda_n + \bar{\zeta}_M} \sum_{m=1}^{M} \frac{(\bar{c}_m - \bar{d}_m)^2 \bar{\zeta}_m^2}{(\lambda_n + \bar{\zeta}_m)^2} \\
&= \frac{1}{\lambda_n + \bar{\zeta}_M} \|\bar{Z}(\bar{w}^0 - \hat{w}^*)\|_2^2,
\end{aligned}$$

we see that when $\bar{w}^0 \neq \hat{w}^*$, $\sum_{m=1}^{M} \frac{(\bar{c}_m - \bar{d}_m)^2 \bar{\zeta}_m^2}{\bar{\zeta}_m^3} > 0$. So we have $\bar{\lambda}_n > 0$ and $\bar{M}_1(\bar{\lambda}_n) < \bar{M}_1(0)$ when $\bar{w}^0 \neq \hat{w}^*$. ■

Proof of Theorem 14: We first prove that $C(w, S)$ is an AMER. It follows from Assumption 5 and Lemma 25 that

$$\begin{aligned}
&E_S[\hat{F}(\hat{w}, S) - \hat{F}(\tilde{w}, S)] \\
&= E_S[\hat{F}(\hat{w}, S) - \frac{1}{n} C(\hat{w}, S) + \frac{1}{n} C(\hat{w}, S) - \hat{F}(\tilde{w}, S)] \\
&\leq E_S[\hat{F}(\hat{w}, S) - \frac{1}{n} C(\hat{w}, S) + \frac{1}{n} C(\tilde{w}, S) - \hat{F}(\tilde{w}, S)] \\
&= E_S\Big(\frac{2\sigma^2 \tilde{w}' \kappa}{n} + \frac{\lambda_n \tilde{w}' \tilde{w}}{n} - \frac{2\sigma^2 \hat{w}' \kappa}{n} - \frac{\lambda_n \hat{w}' \hat{w}}{n}\Big) \\
&\leq \frac{4\sigma^2 B_2^{\frac{1}{2}} K M^{\frac{1}{2}} (1 + n^{-2} K^2 M)^{\frac{1}{2}}}{n} + \frac{2B_2 \lambda_n (1 + n^{-2} K^2 M)}{n}
\end{aligned}$$

$$= O[n^{-1}\max(\lambda_n, KM^{\frac{1}{2}})].$$

So $C(w,S)$ is an AMER with rate $n^{-1}\max(\lambda_n, KM^{\frac{1}{2}})$. Next, we prove that $J(w,S)$ is an AMER. It follows from Lemmas 20, 21 and 25 that

$$\begin{aligned}
\|\hat{\theta}(\bar{w})\|_2^2 &= \sum_{m=1}^{M}\sum_{t=1}^{M}\bar{w}_m\bar{w}_t\hat{\theta}_m^{'}\pi_m^{'}\pi_t\hat{\theta}_t \\
&\leq \sum_{m=1}^{M}\sum_{t=1}^{M}|\bar{w}_m\bar{w}_t||\hat{\theta}_m^{'}\pi_m^{'}\pi_t\hat{\theta}_t| \\
&\leq \max_{1\leq m\leq M}\max_{1\leq t\leq M}|\hat{\theta}_m^{'}\pi_m^{'}\pi_t\hat{\theta}_t|\sum_{m=1}^{M}\sum_{t=1}^{M}|\bar{w}_m\bar{w}_t| \\
&\leq \max_{1\leq m\leq M}\max_{1\leq t\leq M}\|\hat{\theta}_m\|_2\|\hat{\theta}_t\|_2\Big(\sum_{m=1}^{M}|\bar{w}_m|\Big)^2 \\
&= M\max_{1\leq m\leq M}\|\hat{\theta}_m\|_2^2\sum_{m=1}^{M}\bar{w}_m^2 \\
&\leq B_1B_2M, a.s. \qquad (19)
\end{aligned}$$

and

$$\begin{aligned}
&E_S[\|\hat{\theta}(\bar{w}) - \hat{\theta}^{-n}(\bar{w})\|_2^2] \\
&= E_S\Big[\sum_{m=1}^{M}\sum_{t=1}^{M}\bar{w}_m\bar{w}_t(\hat{\theta}_m - \hat{\theta}_m^{-n})^{'}\pi_m^{'}\pi_t(\hat{\theta}_t - \hat{\theta}_t^{-n})\Big] \\
&\leq E_S\Big(\sum_{m=1}^{M}\sum_{t=1}^{M}|\bar{w}_m\bar{w}_t||(\hat{\theta}_m - \hat{\theta}_m^{-n})^{'}\pi_m^{'}\pi_t(\hat{\theta}_t - \hat{\theta}_t^{-n})|\Big) \\
&\leq E_S\Big(\sum_{m=1}^{M}\sum_{t=1}^{M}|\bar{w}_m\bar{w}_t|\|\hat{\theta}_m - \hat{\theta}_m^{-n}\|_2\|\hat{\theta}_t - \hat{\theta}_t^{-n}\|_2\Big) \\
&\leq ME_S\Big(\max_{1\leq m\leq M}\max_{1\leq t\leq M}\|\hat{\theta}_m - \hat{\theta}_m^{-n}\|_2\|\hat{\theta}_t - \hat{\theta}_t^{-n}\|_2\sum_{m=1}^{M}\bar{w}_m^2\Big) \\
&\leq B_2ME_S\Big(\max_{1\leq m\leq M}\|\hat{\theta}_m - \hat{\theta}_m^{-n}\|_2^2\Big) \\
&= O(n^{-2}K^2M). \qquad (20)
\end{aligned}$$

In a similar way to (19), we obtain

$$\|\hat{\theta}^{-n}(\bar{w})\|_2^2 \leq B_1B_2M, a.s..$$

Further, from Assumption 2, we have

$$E_S\Big\{\|x_n[2y_n - x_n^{'}\hat{\theta}(\bar{w}) - x_n^{'}\hat{\theta}^{-n}(\bar{w})]\|_2^2\Big\}$$

$$
\begin{aligned}
&= \sum_{k=1}^{K} E_S\Big\{x_{(k)n}^2[2y_n - x_n^{'}\hat{\theta}(\bar{w}) - x_n^{'}\hat{\theta}^{-n}(\bar{w})]^2\Big\} \\
&\le C_3^2 K E_S\Big\{[2y_n - x_n^{'}\hat{\theta}(\bar{w}) - x_n^{'}\hat{\theta}^{-n}(\bar{w})]^2\Big\} \\
&\le C_3^2 K E_S\Big\{12y_n^2 + 3[x_n^{'}\hat{\theta}(\bar{w})]^2 + 3[x_n^{'}\hat{\theta}^{-n}(\bar{w})]^2\Big\} \\
&\le C_3^2 K E_S[12y_n^2 + 3\|x_n\|_2^2\|\hat{\theta}(\bar{w})\|_2^2 + 3\|x_n\|_2^2\|\hat{\theta}^{-n}(\bar{w})\|_2^2] \\
&\le 12C_2C_3^2K + 6C_3^4B_1B_2K^2M.
\end{aligned}
$$

Therefore, we obtain

$$
\begin{aligned}
&|E_S[\hat{F}(\bar{w},S) - \frac{1}{n}J(\bar{w},S)]| \\
&= \Big|E_S\Big\{\frac{1}{n}\sum_{i=1}^{n}\big[[y_i - x_i^{'}\hat{\theta}(\bar{w})]^2 - [y_i - x_i^{'}\hat{\theta}^{-i}(\bar{w})]^2\big] - \frac{\lambda_n\bar{w}^{'}\bar{w}}{n}\Big\}\Big| \\
&= \Big|E_S\Big\{[y_n - x_n^{'}\hat{\theta}(\bar{w})]^2 - [y_n - x_n^{'}\hat{\theta}^{-n}(\bar{w})]^2\Big\} - E_S(\frac{\lambda_n\bar{w}^{'}\bar{w}}{n})\Big| \\
&= \Big|E_S\Big\{[2y_n - x_n^{'}\hat{\theta}(\bar{w}) - x_n^{'}\hat{\theta}^{(-n)}(\bar{w})][x_n^{'}\hat{\theta}^{-n}(\bar{w}) - x_n^{'}\hat{\theta}(\bar{w})]\Big\} - E_S(\frac{\lambda_n\bar{w}^{'}\bar{w}}{n})\Big| \\
&\le \sqrt{E_S[\|x_n(2y_n - x_n^{'}\hat{\theta}(\bar{w}) - x_n^{'}\hat{\theta}^{-n}(\bar{w})\|_2^2]}\sqrt{E_S[\|\hat{\theta}(\bar{w}) - \hat{\theta}^{-n}(\bar{w})\|_2^2]} + B_2n^{-1}\lambda_n \\
&= O[n^{-1}\max(\lambda_n, K^2M)].
\end{aligned} \tag{21}
$$

In a similar way, it is seen that

$$
|E_S[\frac{1}{n}J_S(\tilde{w}) - \hat{F}(\tilde{w},S)]| = O[n^{-1}\max(\lambda_n, K^2M)].
$$

Thus, we have

$$
\begin{aligned}
&E_S[\hat{F}(\bar{w},S) - \hat{F}(\tilde{w},S)] \\
&= E_S[\hat{F}(\bar{w},S) - \frac{1}{n}J(\bar{w},S) + \frac{1}{n}J(\bar{w},S) - \hat{F}(\tilde{w},S)] \\
&\le E_S[\hat{F}(\bar{w},S) - \frac{1}{n}J(\bar{w},S) + \frac{1}{n}J(\tilde{w},S) - \hat{F}(\tilde{w},S)] \\
&= O[n^{-1}\max(\lambda_n, K^2M)].
\end{aligned}
$$

So $J(w,S)$ is an AMER with rate $n^{-1}\max(\lambda_n, K^2M)$. ■

Proof of Theorem 15: We first prove that $C(w,S)$ has PLoo and FLoo stability. From Lemma 26, we have

$$
\begin{aligned}
&E_S[\|\hat{\theta}(\hat{w}) - \hat{\theta}(\hat{w}^{-n})\|_2^2] \\
&= E_S\Big[\sum_{m=1}^{M}\sum_{t=1}^{M}(\hat{w}_m - \hat{w}_m^{-n})(\hat{w}_t - \hat{w}_t^{-n})\hat{\theta}_m^{'}\pi_m^{'}\pi_t\hat{\theta}_t\Big] \\
&\le E_S\Big[\sum_{m=1}^{M}\sum_{t=1}^{M}|(\hat{w}_m - \hat{w}_m^{-n})(\hat{w}_t - \hat{w}_t^{-n})||\hat{\theta}_m^{'}\pi_m^{'}\pi_t\hat{\theta}_t|\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\leq E_S\Big[\sum_{m=1}^{M}\sum_{t=1}^{M}|(\hat{w}_m-\hat{w}_m^{-n})(\hat{w}_t-\hat{w}_t^{-n})|\|\hat{\theta}_m\|_2\|\hat{\theta}_t\|_2\Big]\\
&\leq ME_S\Big[\max_{1\leq m\leq M}\max_{1\leq t\leq M}\|\hat{\theta}_m\|_2\|\hat{\theta}_t\|_2\sum_{m=1}^{M}(\hat{w}_m-\hat{w}_m^{-n})^2\Big]\\
&\leq B_1ME_S\Big[\sum_{m=1}^{M}(\hat{w}_m-\hat{w}_m^{-n})^2\Big]\\
&= O(n^{-2}K^4M^3). \qquad (22)
\end{aligned}
$$

Similar to (21), we see that

$$
\Big|E_S\Big\{[2y_n-x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})-x_n^{'}\hat{\theta}(\hat{w})][x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})-x_n^{'}\hat{\theta}(\hat{w}^{-n})]\Big\}\Big|=O(n^{-1}K^2M)
$$

and

$$
\Big|E_S\Big\{[2y_n-x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})-x_n^{'}\hat{\theta}(\hat{w})][x_n^{'}\hat{\theta}(\hat{w}^{-n})-x_n^{'}\hat{\theta}(\hat{w})]\Big\}\Big|=O(n^{-1}K^3M^2).
$$

Noting that

$$
\begin{aligned}
&|E_S\Big\{[y_n-x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})]^2-[y_n-x_n^{'}\hat{\theta}(\hat{w})]^2\Big\}|\\
&=|E_S\Big\{[2y_n-x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})-x_n^{'}\hat{\theta}(\hat{w})][x_n^{'}\hat{\theta}(\hat{w})-x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})]\Big\}|\\
&\leq|E_S\Big\{[2y_n-x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})-x_n^{'}\hat{\theta}(\hat{w})][x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})-x_n^{'}\hat{\theta}(\hat{w}^{-n})]\Big\}|\\
&\quad+|E_S\Big\{[2y_n-x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})-x_n^{'}\hat{\theta}(\hat{w})][x_n^{'}\hat{\theta}(\hat{w}^{-n})-x_n^{'}\hat{\theta}(\hat{w})]\Big\}|,
\end{aligned}
$$

we have

$$
E_S\Big\{[y_n-x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n})]^2-[y_n-x_n^{'}\hat{\theta}(\hat{w})]^2\Big\}=O(n^{-1}K^3M^2).
$$

In a similar way, we obtain

$$
E_{S,z^*}\Big\{[y^*-x^{*'}\hat{\theta}^{-n}(\hat{w}^{-n})]^2-[y^*-x^{*'}\hat{\theta}(\hat{w})]^2\Big\}=O(n^{-1}K^3M^2).
$$

Therefore, from Definitions 3-4, $C(w,S)$ has PLoo and FLoo stability with rate $n^{-1}K^3M^2$.

Next, we prove that $J(w,S)$ has PLoo and FLoo stability. From Lemma 27, we have

$$
|E_S\Big\{[2y_n-x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n})-x_n^{'}\hat{\theta}(\bar{w})][x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n})-x_n^{'}\hat{\theta}(\bar{w}^{-n})]\Big\}|=O(n^{-1}K^2M)
$$

and

$$
|E_S\Big\{[2y_n-x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n})-x_n^{'}\hat{\theta}(\bar{w})][x_n^{'}\hat{\theta}(\bar{w}^{-n})-x_n^{'}\hat{\theta}(\bar{w})]\Big\}|=O(n^{-1}K^3M^2).
$$

Further, since

$$
|E_S\Big\{[y_n-x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n})]^2-[y_n-x_n^{'}\hat{\theta}(\bar{w})]^2\Big\}|
$$

$$
\begin{aligned}
&= |E_S\Big\{[2y_n - x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n}) - x_n^{'}\hat{\theta}(\bar{w})][x_n^{'}\hat{\theta}(\bar{w}) - x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n})]\Big\}| \\
&\leq |E_S\Big\{[2y_n - x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n}) - x_n^{'}\hat{\theta}(\bar{w})][x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n}) - x_n^{'}\hat{\theta}(\bar{w}^{-n})]\Big\}| \\
&+ |E_S\Big\{[2y_n - x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n}) - x_n^{'}\hat{\theta}(\bar{w})][x_n^{'}\hat{\theta}(\bar{w}^{-n}) - x_n^{'}\hat{\theta}(\bar{w})]\Big\}|,
\end{aligned}
$$

it is seen that

$$
E_S\Big\{[y_n - x_n^{'}\hat{\theta}^{-n}(\bar{w}^{-n})]^2 - [y_n - x_n^{'}\hat{\theta}(\bar{w})]^2\Big\} = O(n^{-1}K^3M^2).
$$

In a similar way, we obtain

$$
E_{S,z^*}\Big\{[y^* - x^{*'}\hat{\theta}^{-n}(\bar{w}^{-n})]^2 - [y^* - x^{*'}\hat{\theta}(\bar{w})]^2\Big\} = O(n^{-1}K^3M^2).
$$

Thus, from Definitions 3-4, $J(w, S)$ has PLoo and FLoo stability with rate $n^{-1}K^3M^2$.

Finally, we prove that $F(w, S)$ has PLoo and FLoo stability. Similarly, it follows from Lemma 28 that

$$
|E_S\Big\{[2y_n - x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*}) - x_n^{'}\hat{\theta}(\hat{w}^*)][x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*}) - x_n^{'}\hat{\theta}(\hat{w}^{-n*})]\Big\}| = O(n^{-1}K^2M)
$$

and

$$
|E_S\Big\{[2y_n - x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*}) - x_n^{'}\hat{\theta}(\hat{w}^*)][x_n^{'}\hat{\theta}(\hat{w}^{-n*}) - x_n^{'}\hat{\theta}(\hat{w}^*)]\Big\}| = O(n^{-1}K^3M^2).
$$

Since

$$
\begin{aligned}
&|E_S\Big\{[y_n - x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*})]^2 - [y_n - x_n^{'}\hat{\theta}(\hat{w}^*)]^2\Big\}| \\
&= |E_S\Big\{[2y_n - x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*}) - x_n^{'}\hat{\theta}(\hat{w}^*)][x_n^{'}\hat{\theta}(\hat{w}^*) - x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*})]\Big\}| \\
&\leq |E_S\Big\{(2y_n - x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*}) - x_n^{'}\hat{\theta}(\hat{w}^*)][x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*}) - x_n^{'}\hat{\theta}(\hat{w}^{-n*})]\Big\}| \\
&+ |E_S\Big\{[2y_n - x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*}) - x_n^{'}\hat{\theta}(\hat{w}^*)][x_n^{'}\hat{\theta}(\hat{w}^{-n*}) - x_n^{'}\hat{\theta}(\hat{w}^*)]\Big\}|,
\end{aligned}
$$

we see that

$$
E_S\Big\{[y_n - x_n^{'}\hat{\theta}^{-n}(\hat{w}^{-n*})]^2 - [y_n - x_n^{'}\hat{\theta}(\hat{w}^*)]^2\Big\} = O(n^{-1}K^3M^2).
$$

In a similar way, we obtain

$$
E_{S,z^*}\Big\{[y^* - x^{*'}\hat{\theta}^{-n}(\hat{w}^{-n*})]^2 - [y^* - x^{*'}\hat{\theta}(\hat{w}^*)]^2\Big\} = O(n^{-1}K^3M^2).
$$

So from Definitions 3-4, $F(w, S)$ has PLoo and FLoo stability with rate $n^{-1}K^3M^2$.

■

Proof of Theorem 17: From Assumption 5, Assumption 6 and Lemma 25, we know that

$$
E_S\{\Lambda_{max}^2[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma}) - \hat{\Omega}^{'}\hat{\Omega}/n]\|\hat{w}\|_2^2\} = o(n^{-1}K),
$$

$$
(\lambda_n/n)^2 E_S(\|\hat{w}\|_2^2) = o(n^{-1}K)
$$

and

$$E_S(\|\sigma^2\kappa/n\|_2^2) = o(n^{-1}K).$$

Then, from

$$E_S[F(\hat{w}, S) - F(\hat{w}^*, S)] = E_S[(\hat{w} - \hat{w}^*)^{'} E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})(\hat{w} - \hat{w}^*)]$$

and

$$\begin{aligned}
&C_4 E_S[(\hat{w} - \hat{w}^*)^{'} E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})(\hat{w} - \hat{w}^*)] \\
&\leq E_S\{\lambda_{min}[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})](\hat{w} - \hat{w}^*)^{'} E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})(\hat{w} - \hat{w}^*)\} \\
&\leq E_S[\|E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})(\hat{w} - \hat{w}^*)\|_2^2] \\
&= E_S[\|E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})\hat{w} - (\hat{\Omega}^{'}\hat{\Omega} + \lambda_n I_M)\hat{w}/n + (\hat{\Omega}^{'}\hat{\Omega} + \lambda_n I_M)\hat{w}/n - E_{z^*}(\hat{\gamma}^{'}\hat{\gamma})\hat{w}^*\|_2^2] \\
&= E_S\{\|[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma}) - (\hat{\Omega}^{'}\hat{\Omega}/n + \lambda_n/nI_M)]\hat{w} + (\hat{\theta}^{'} X_M^{'} y - \sigma^2\kappa)/n - \hat{\theta}^{'} E_{z^*}(x^* y^*)\|_2^2\} \\
&= E_S\{\|[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma}) - (\hat{\Omega}^{'}\hat{\Omega}/n + \lambda_n/nI_M)]\hat{w} + \hat{\theta}^{'}[X_M^{'} y/n - E_{z^*}(x^* y^*)] - \sigma^2\kappa/n\|_2^2\} \\
&\leq 4E_S\{\|[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma}) - \hat{\Omega}^{'}\hat{\Omega}/n]\hat{w}\|_2^2\} + 4(\lambda_n/n)^2 E_S(\|\hat{w}\|_2^2) \\
&\quad + 4E_S\{\|\hat{\theta}^{'}[X_M^{'} y/n - E_{z^*}(x^* y^*)]\|_2^2\} + 4E_S(\|\sigma^2\kappa/n\|_2^2) \\
&\leq 4E_S\{\Lambda_{max}^2[E_{z^*}(\hat{\gamma}^{'}\hat{\gamma}) - \hat{\Omega}^{'}\hat{\Omega}/n]\|\hat{w}\|_2^2\} + 4(\lambda_n/n)^2 E_S(\|\hat{w}\|_2^2) \\
&\quad + 4E_S\{\|\hat{\theta}^{'}[X_M^{'} y/n - E_{z^*}(x^* y^*)]\|_2^2\} + 4E_S(\|\sigma^2\kappa/n\|_2^2),
\end{aligned}$$

we see that proving $E_S\{\|\hat{\theta}^{'}[X_M^{'} y/n - E_{z^*}(x^* y^*)]\|_2^2\} = O(n^{-1}KM)$ is sufficient to demonstrate that $C(w, S)$ is consistent with rate $O(n^{-1}KM)$. It follows from Lemma 20 and Marcinkiewicz-Zygmund-Burkholder inequality in Lin and Bai (2010) that

$$\begin{aligned}
&E_S\{\|\hat{\theta}^{'}[X_M^{'} y/n - E_{z^*}(x^* y^*)]\|_2^2\} \\
&\leq E_S[\lambda_{max}(\hat{\theta}\hat{\theta}^{'})\|X_M^{'} y/n - E_{z^*}(x^* y^*)\|_2^2] \\
&\leq B_1 M E_S[\|X_M^{'} y/n - E_{z^*}(x^* y^*)\|_2^2] \\
&= B_1 M \sum_{k=1}^{K} E_S\Big\{\Big[\frac{1}{n}\sum_{i=1}^{n}[x_{(k)i} y_i - E_{z^*}(x_{(k)}^* y^*)]\Big]^2\Big\} \\
&\leq 4B_1 n^{-2} KM \max_{1\leq k\leq K} E_S\Big\{\sum_{i=1}^{n}[x_{(k)i} y_i - E_{z^*}(x_{(k)}^* y^*)]^2\Big\} \\
&= 4B_1 n^{-1} KM \max_{1\leq k\leq K} var(x_{(k)i} y_i) \\
&\leq 4B_1 C_6 n^{-1} KM.
\end{aligned}$$

Thus, we have completed the proof of Theorem 17. ■

Appendix C. Figures and Tables

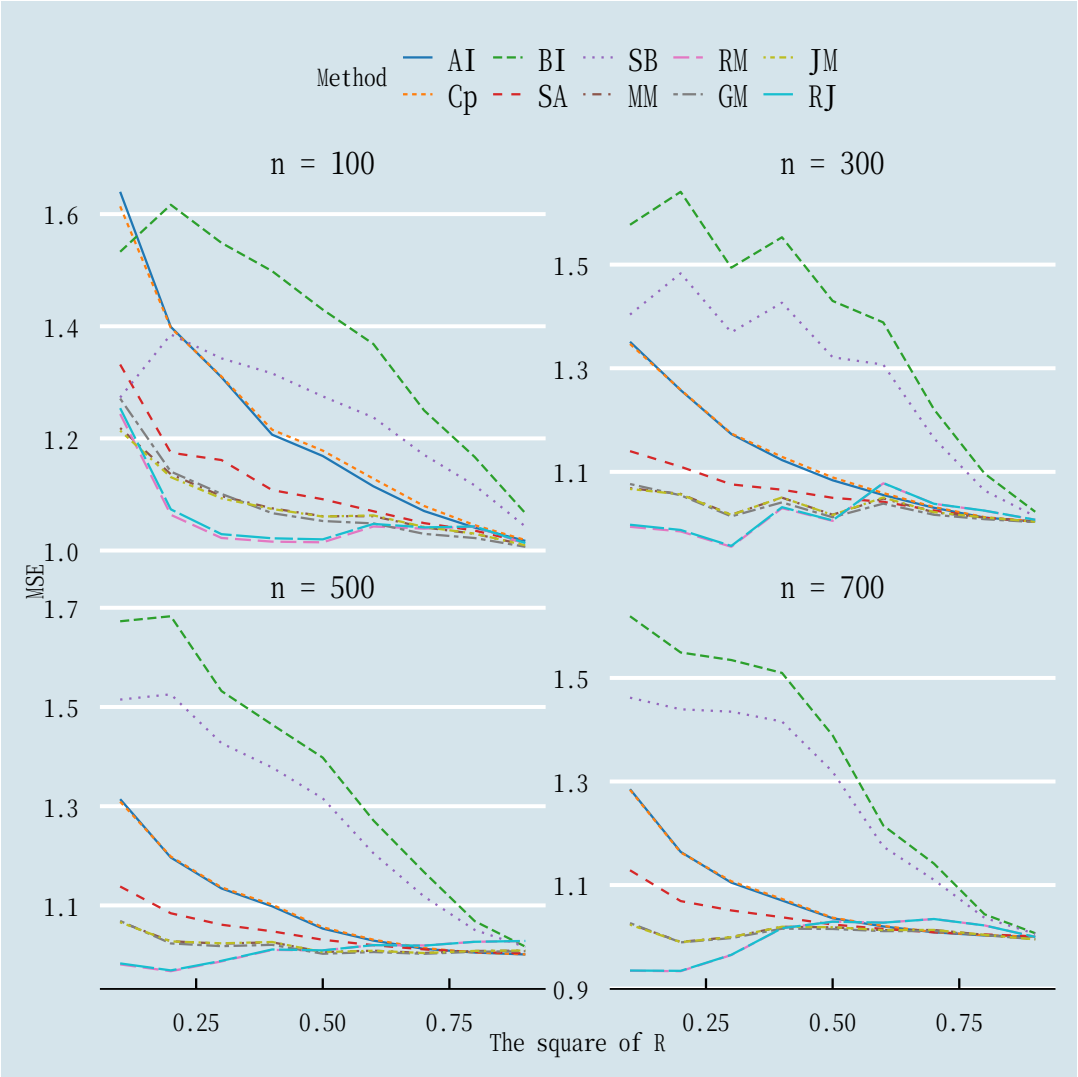

Figure 1: The mean of MSEs under homoskedastic errors with $\alpha = 0.5$ for nested setting of simulation study

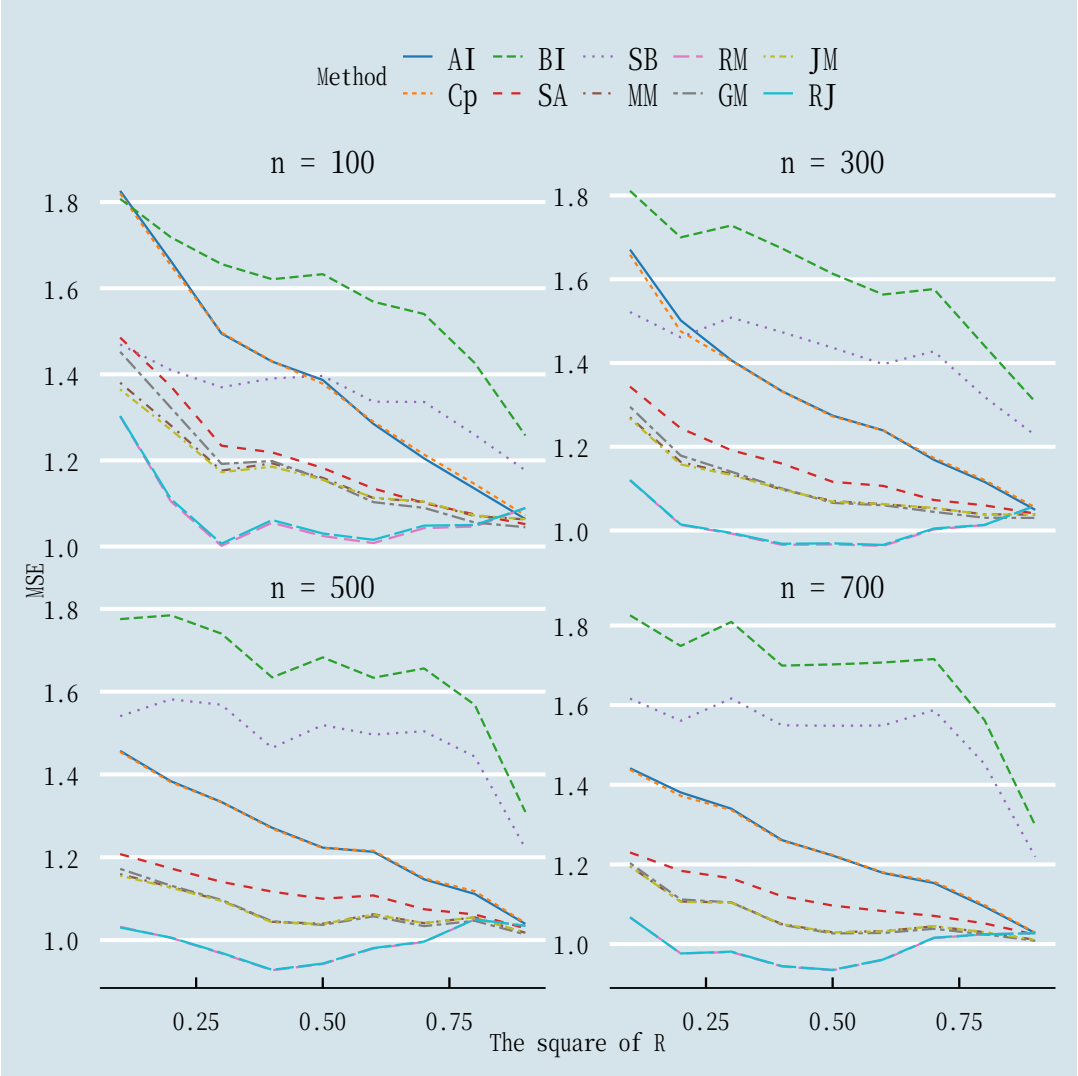

Figure 2: The mean of MSEs under homoskedastic errors with $\alpha = 1.0$ for nested setting of simulation study

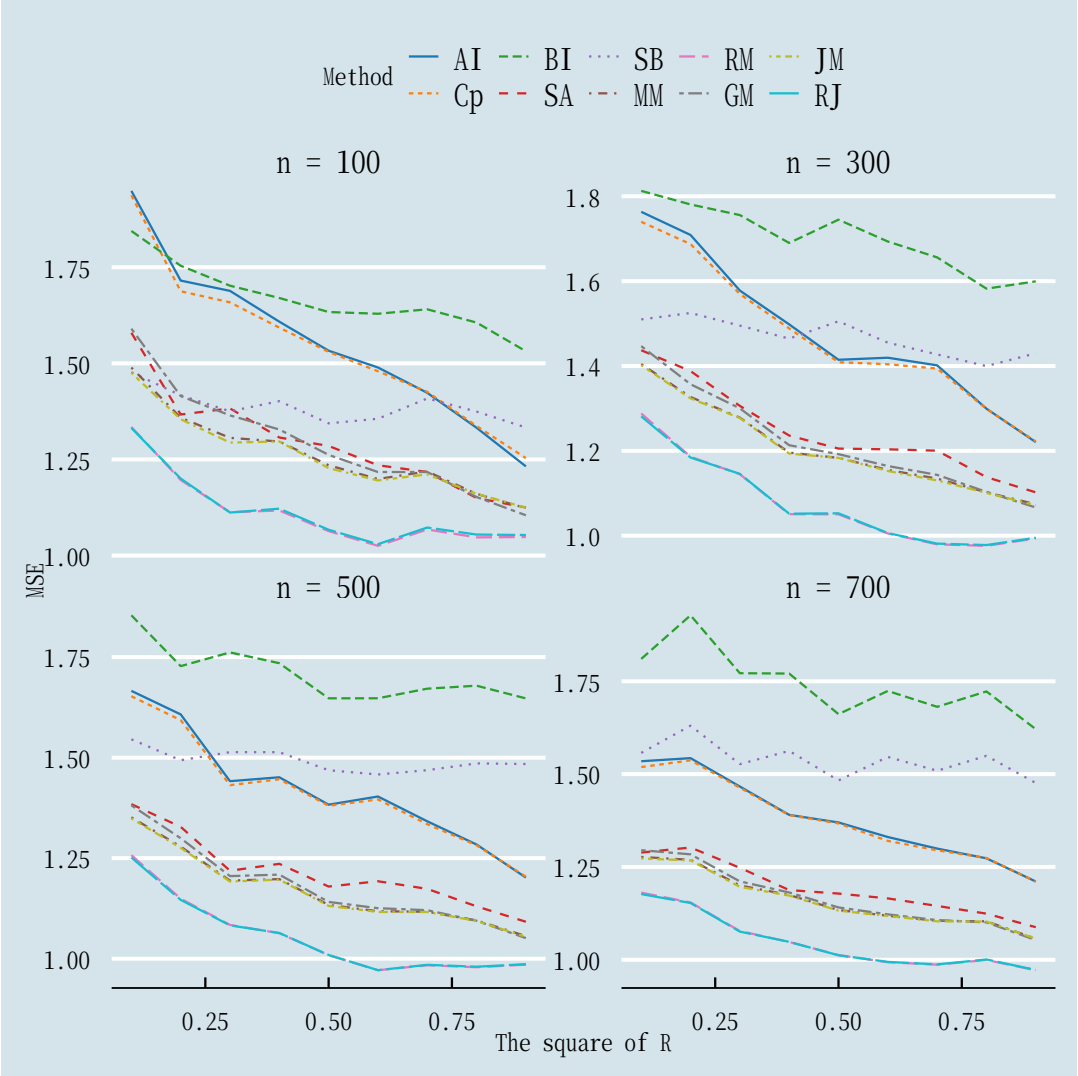

Figure 3: The mean of MSEs under homoskedastic errors with $\alpha = 1.5$ for nested setting of simulation study

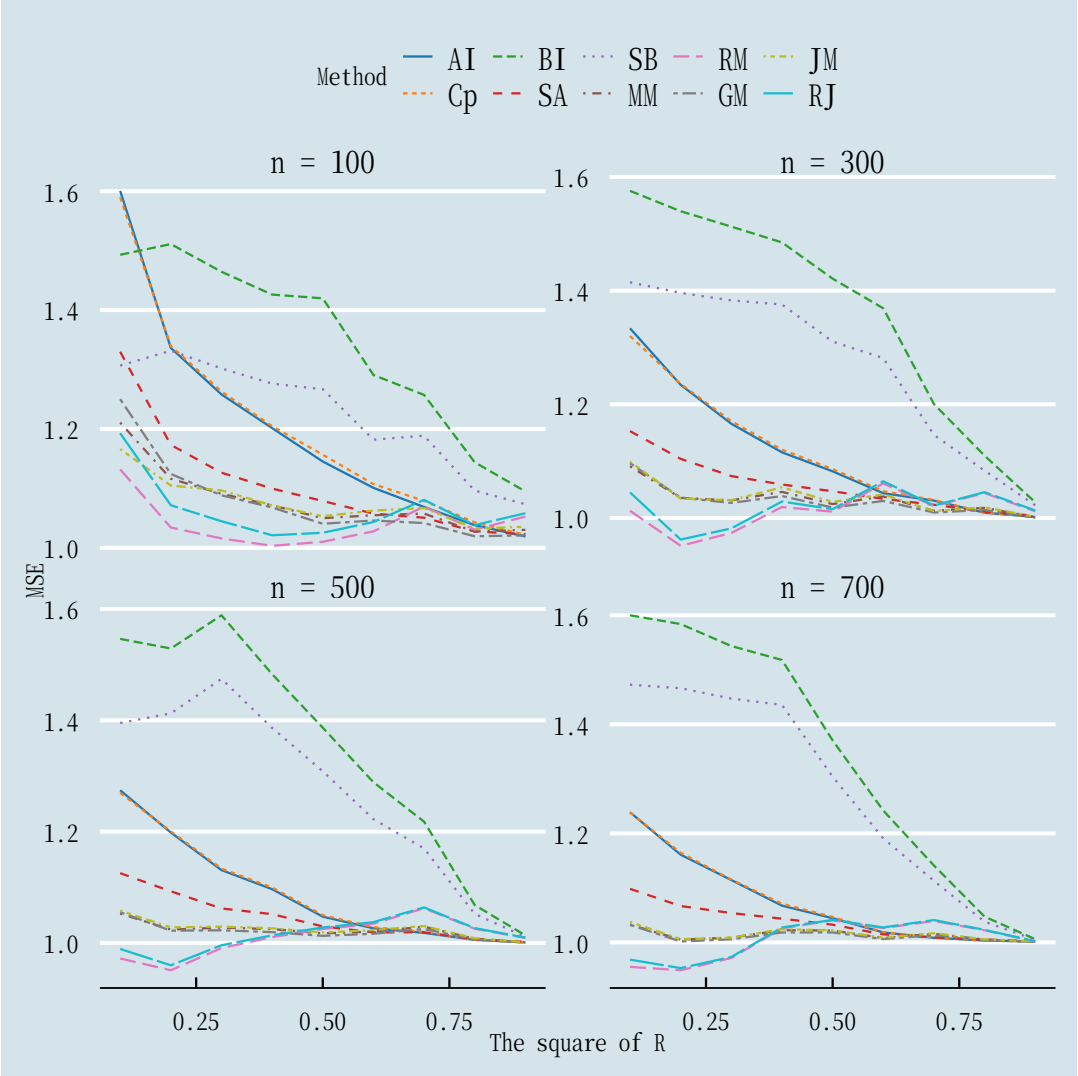

Figure 4: The mean of MSEs under heteroskedastic errors with $\alpha = 0.5$ for nested setting of simulation study

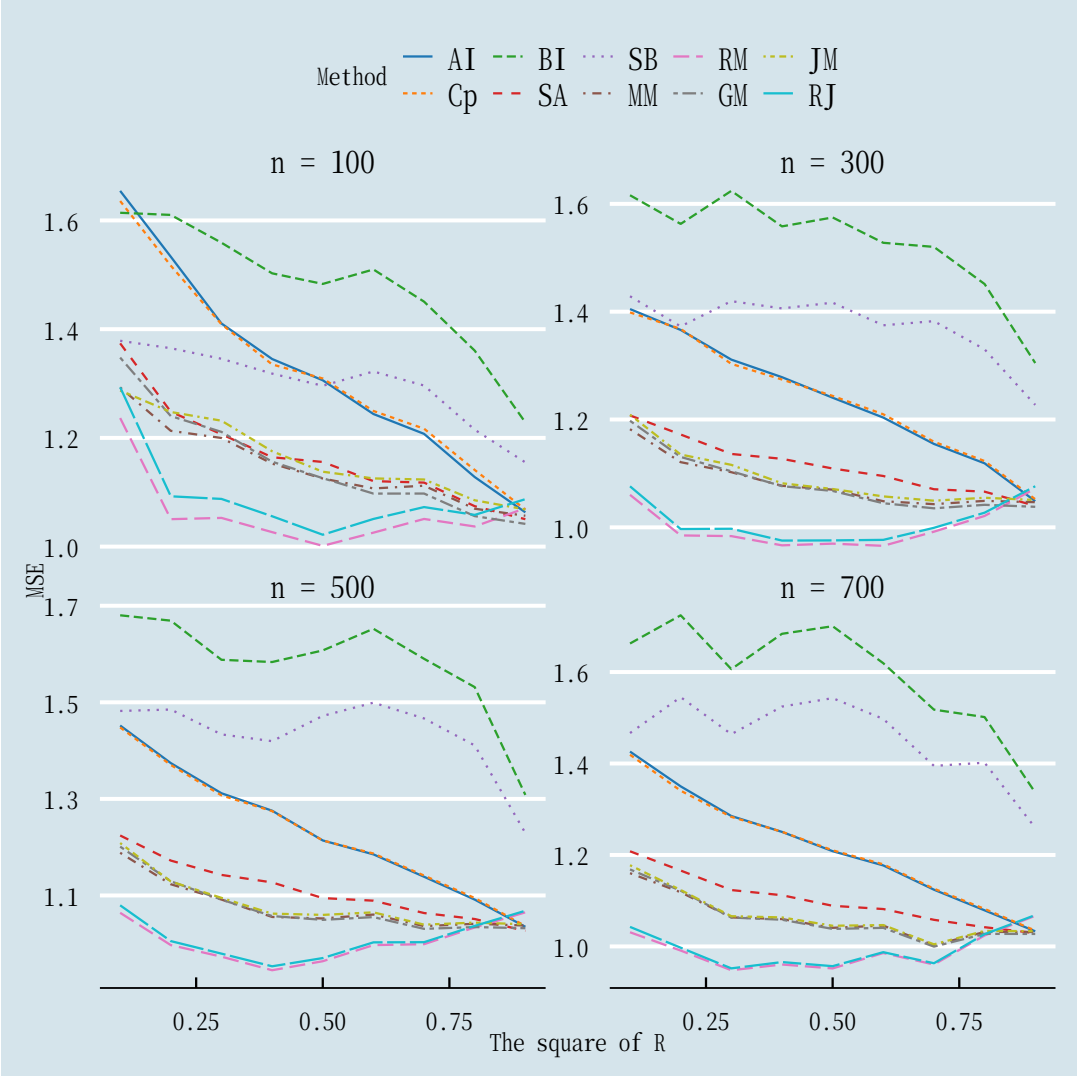

Figure 5: The mean of MSEs under heteroskedastic errors with $\alpha = 1.0$ for nested setting of simulation study

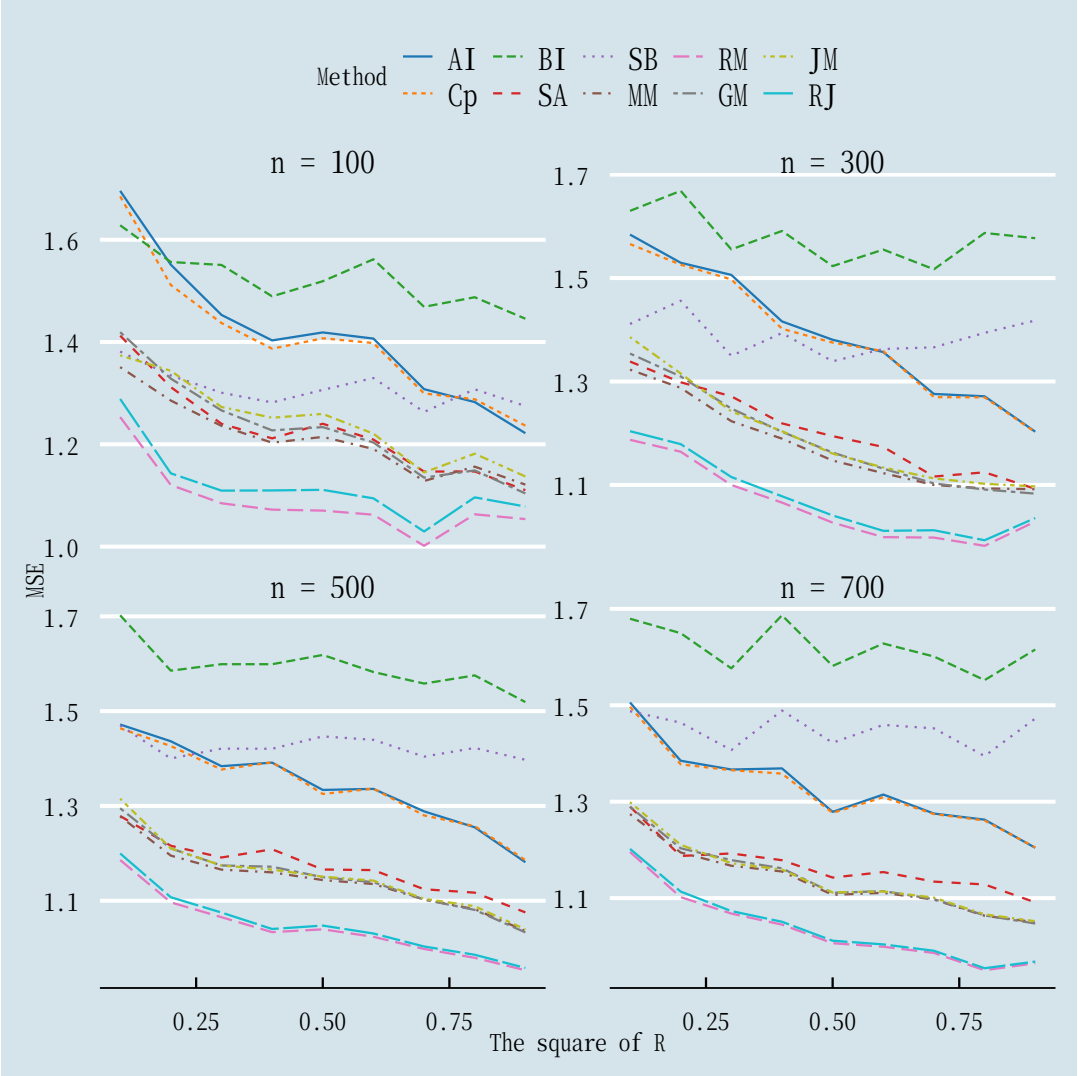

Figure 6: The mean of MSEs under heteroskedastic errors with $\alpha = 1.5$ for nested setting of simulation study

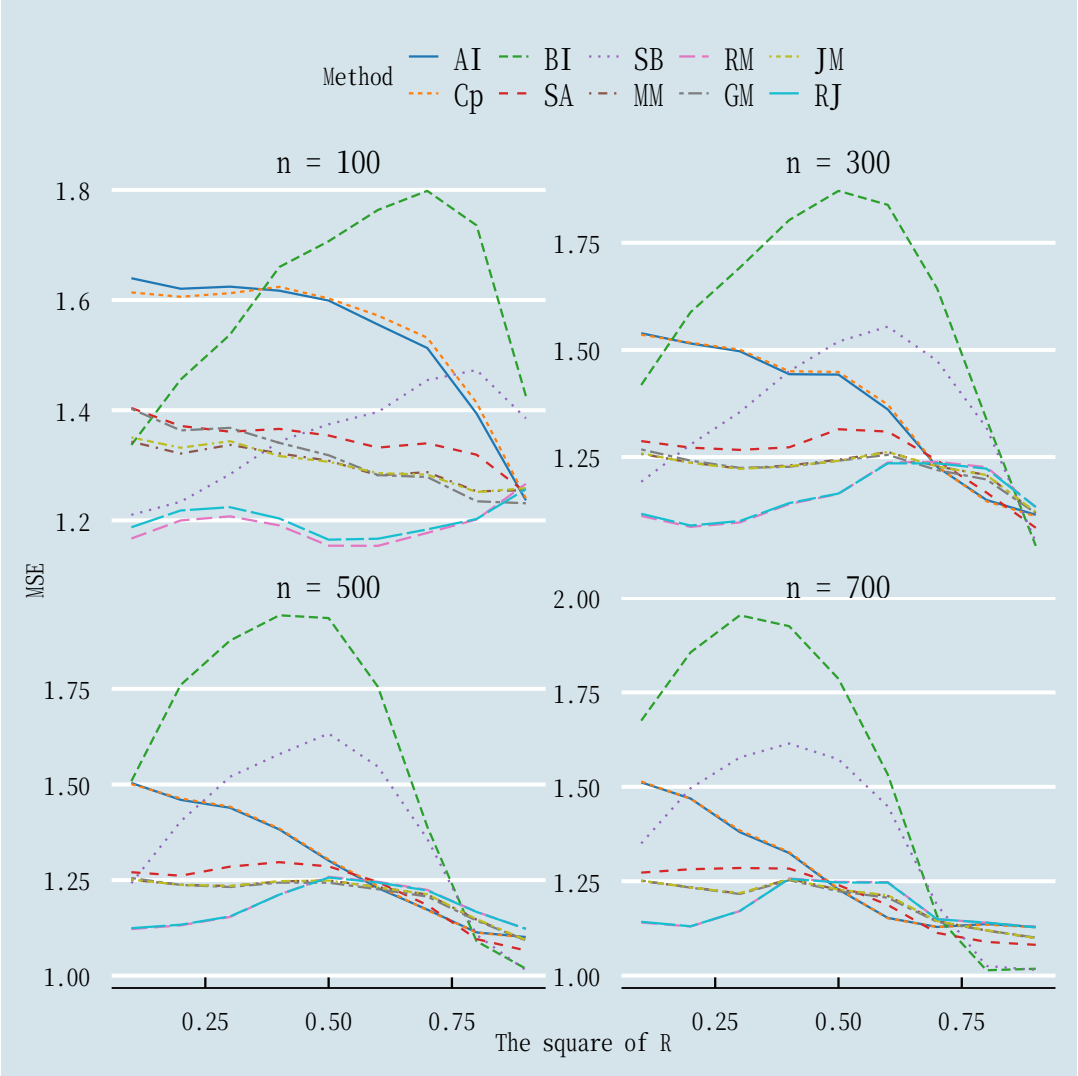

Figure 7: The mean of MSEs under homoskedastic errors with $\alpha = 0.5$ for non-nested setting of simulation study

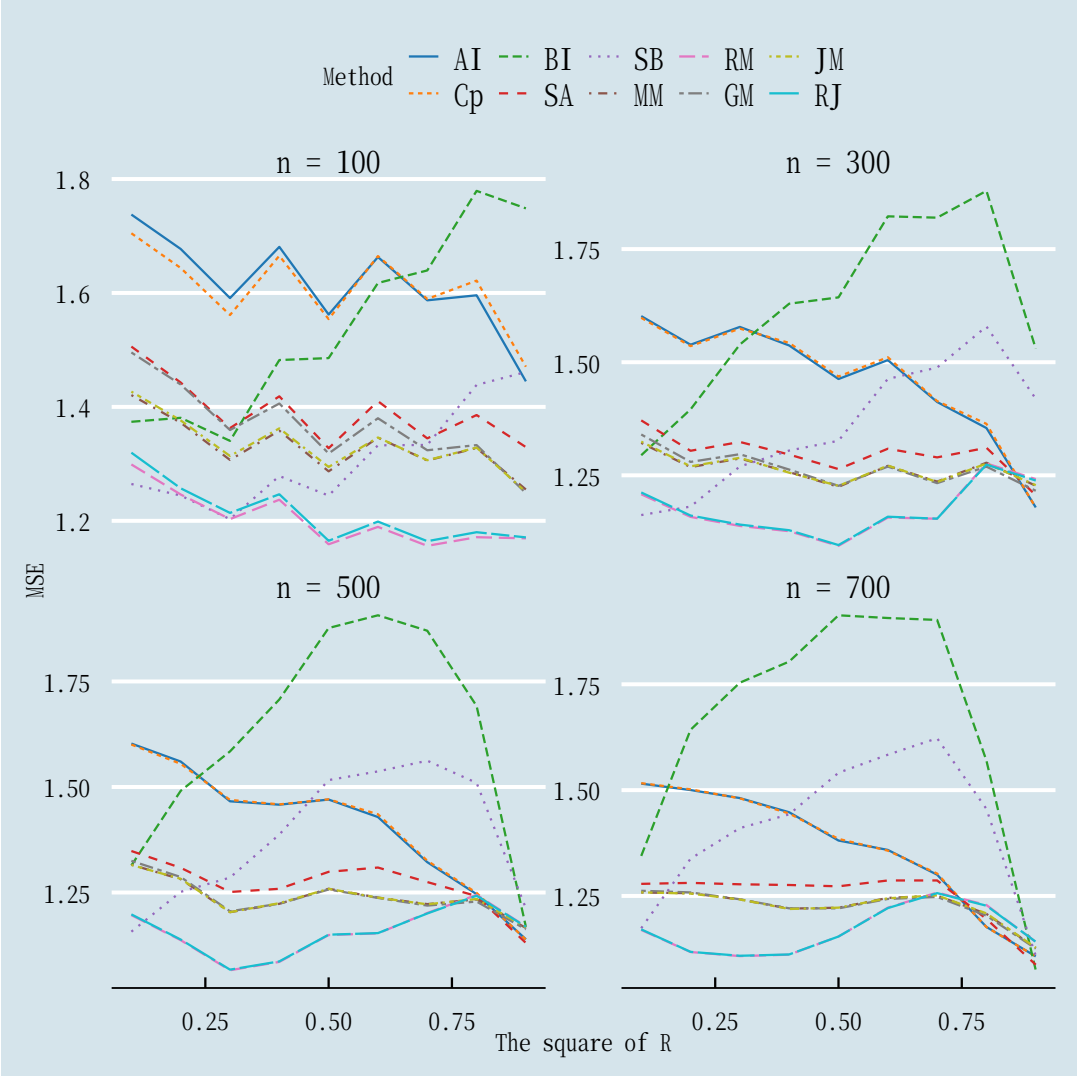

Figure 8: The mean of MSEs under homoskedastic errors with $\alpha = 1.0$ for non-nested setting of simulation study

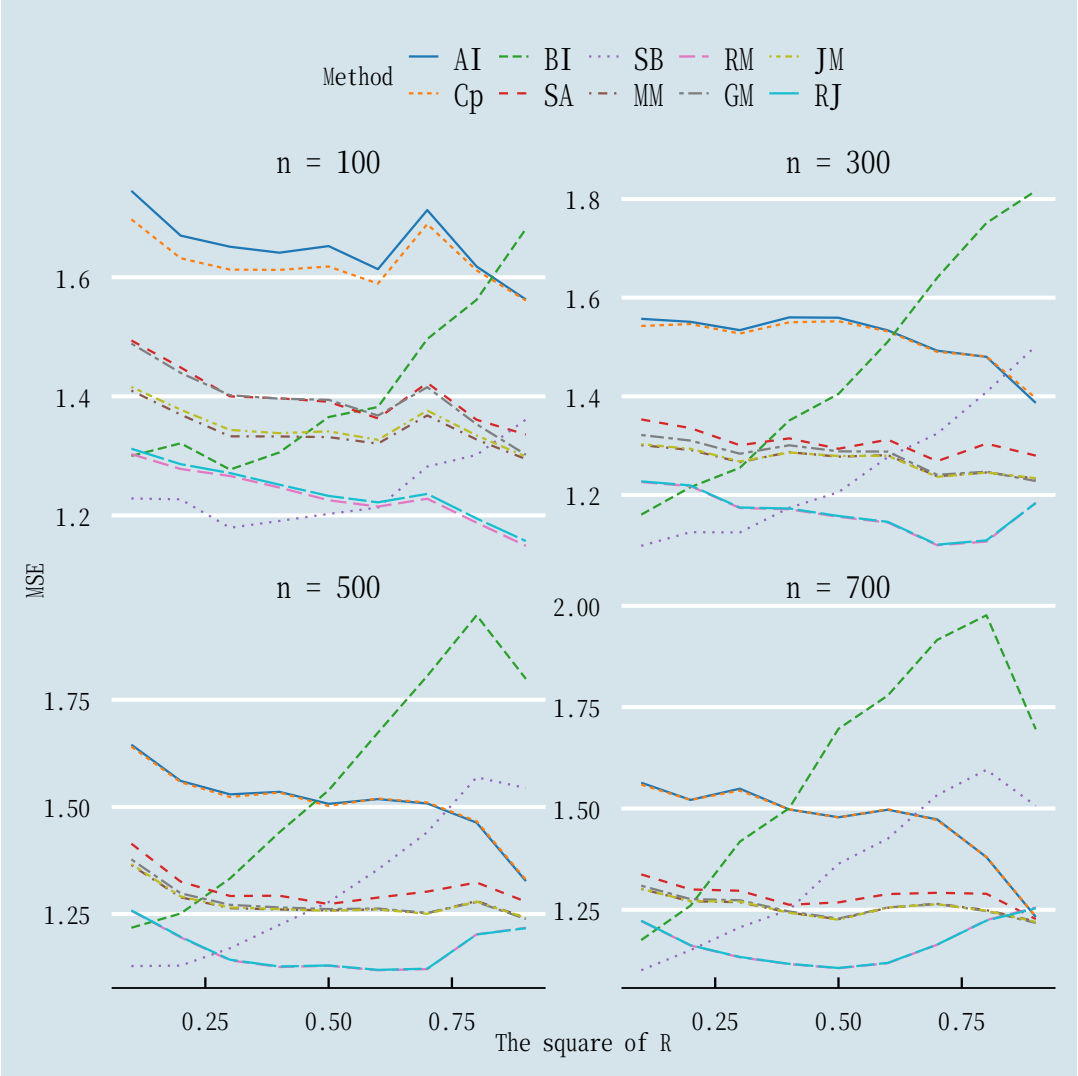

Figure 9: The mean of MSEs under homoskedastic errors with $\alpha = 1.5$ for non-nested setting of simulation study

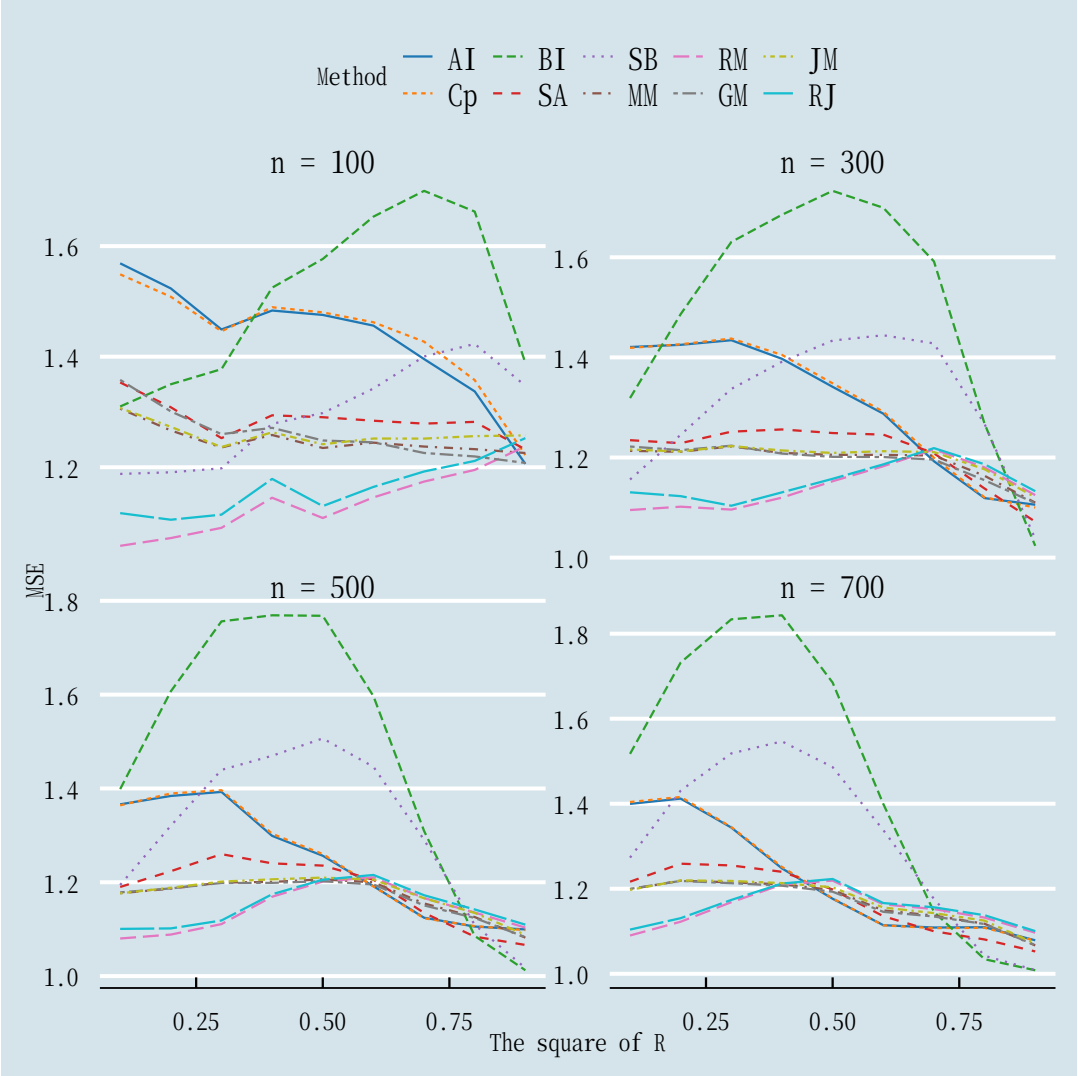

Figure 10: The mean of MSEs under heteroskedastic errors with $\alpha = 0.5$ for non-nested setting of simulation study

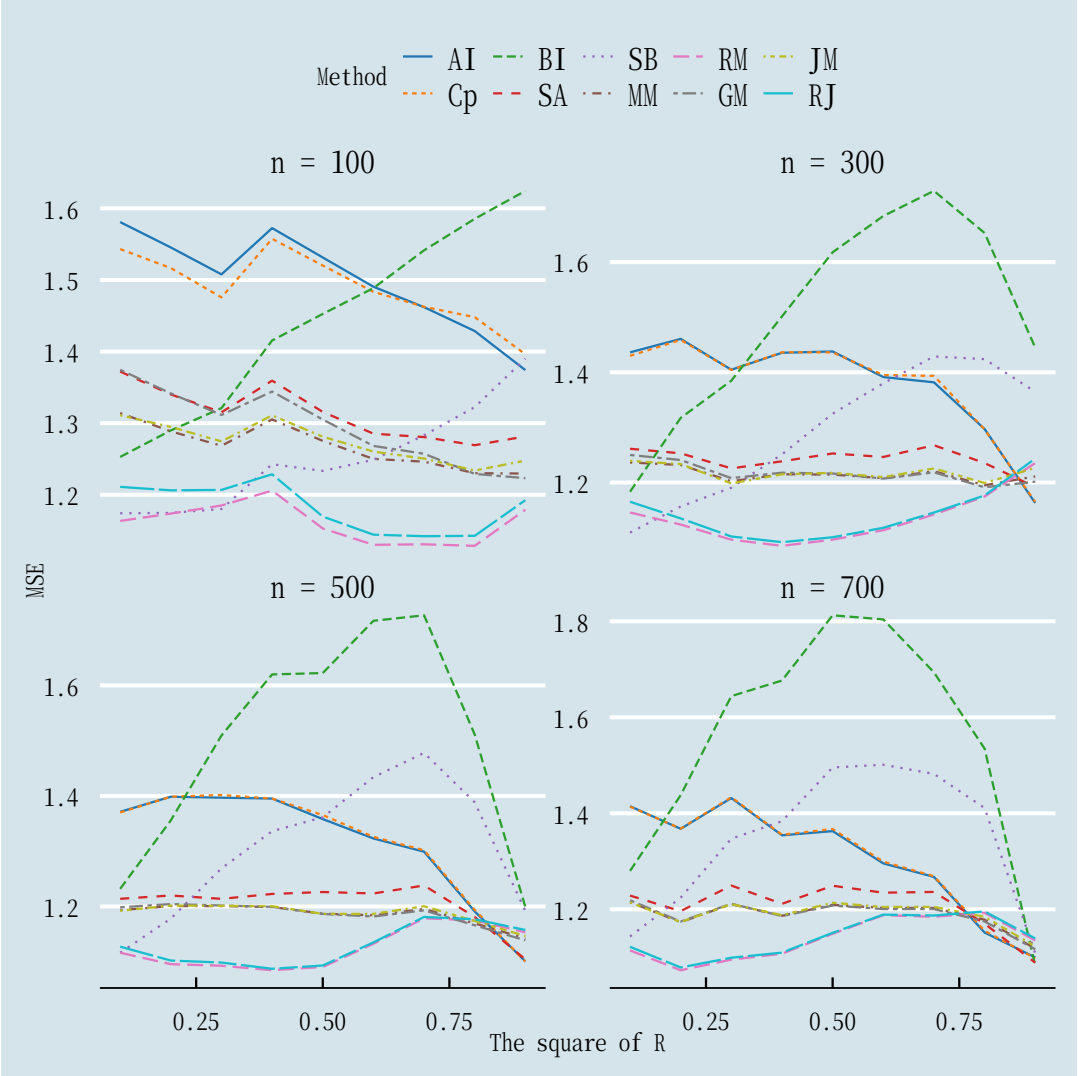

Figure 11: The mean of MSEs under heteroskedastic errors with $\alpha = 1.0$ for non-nested setting of simulation study

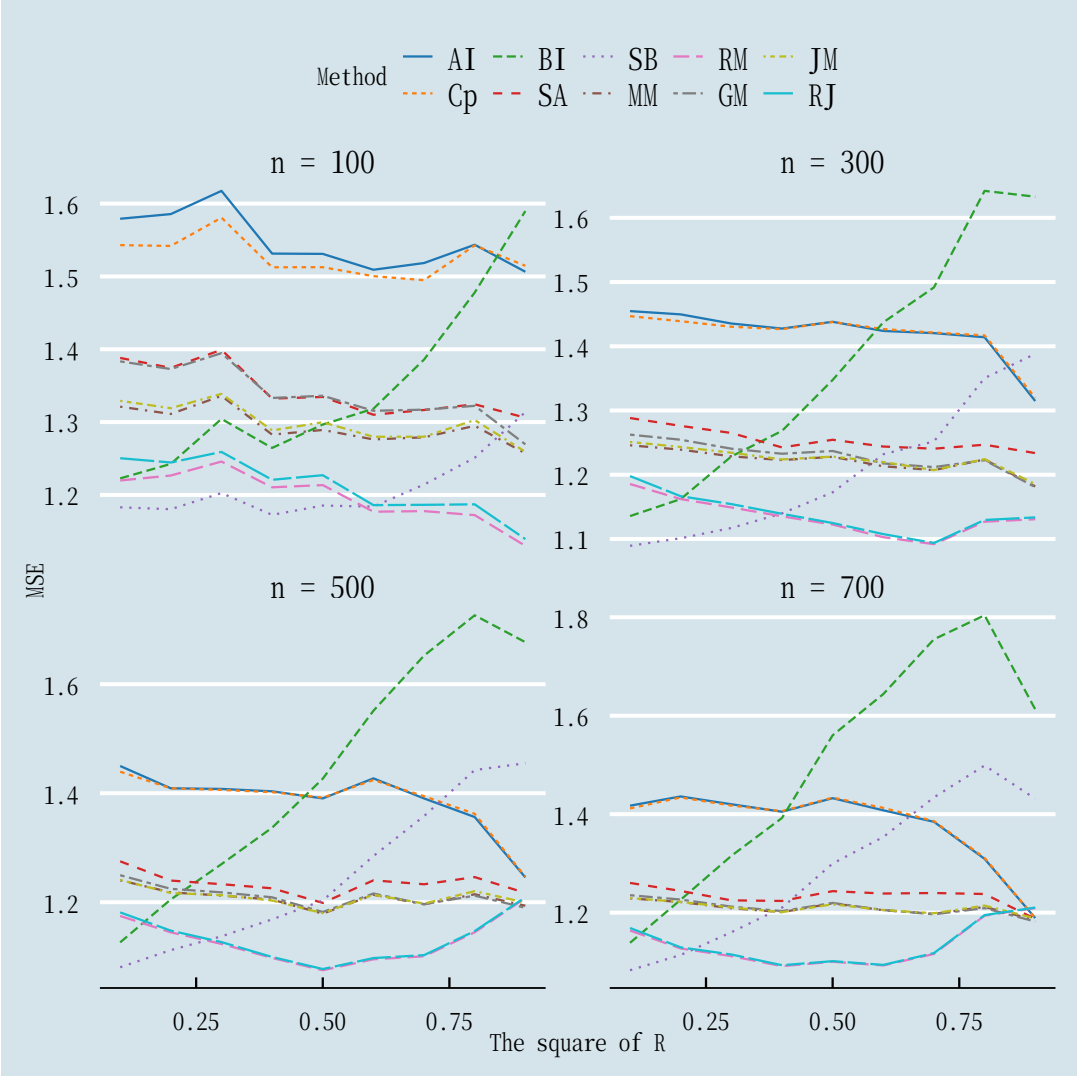

Figure 12: The mean of MSEs under heteroskedastic errors with $\alpha = 1.5$ for non-nested setting of simulation study

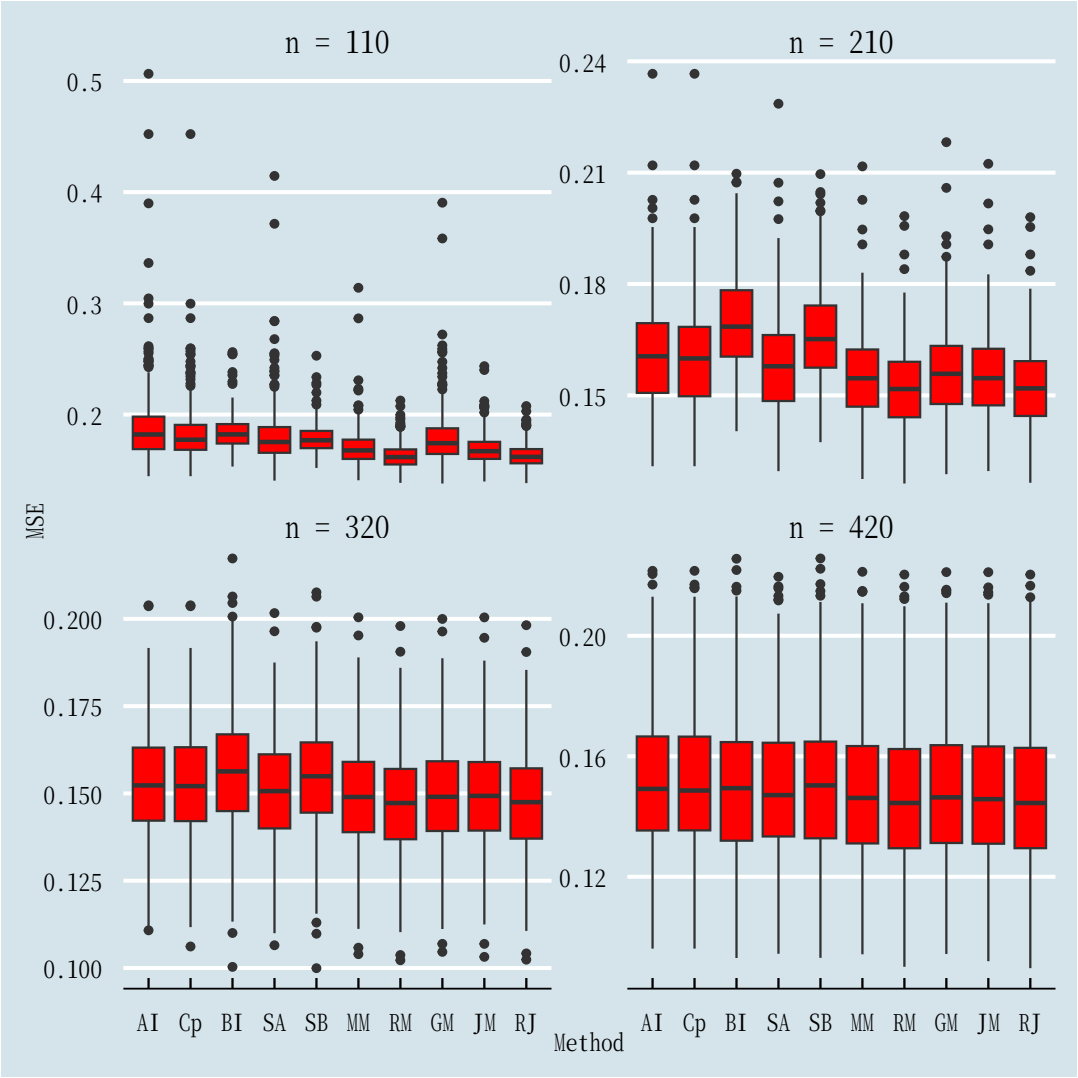

Figure 13: The box plot of MSEs in Case 1 of real data analysis

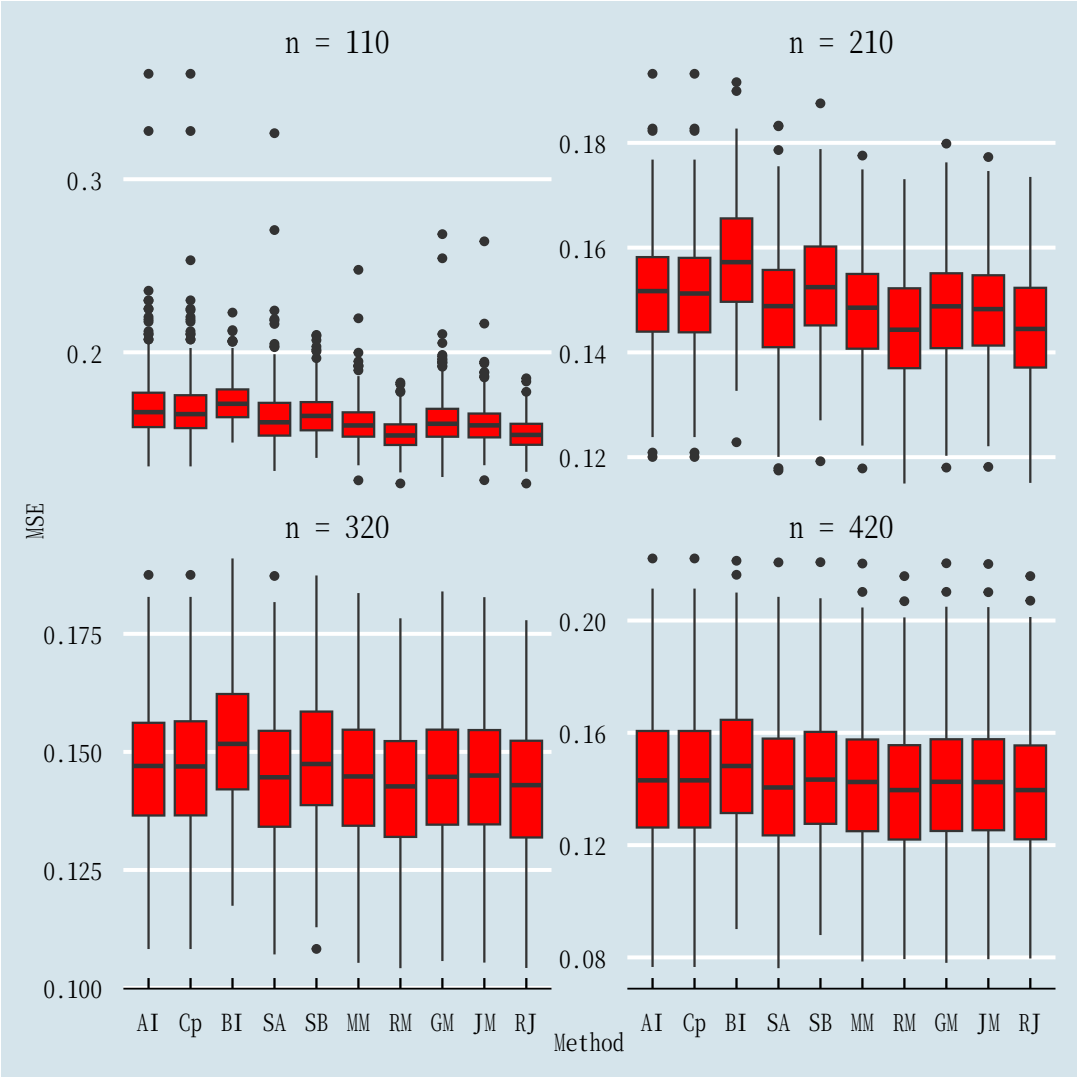

Figure 14: The box plot of MSEs in Case 2 of real data analysis

Table 1: The mean, median and BPR of MSEs in Case 1 of real data analysis

n		AI	Cp	BI	SA	SB	MM	RM	GM	JM	RJ
110	Mean	0.190	0.183	0.184	0.181	0.178	0.170	**0.163**	0.179	0.169	**0.163**
	Median	0.182	0.177	0.182	0.175	0.177	0.168	**0.162**	0.174	0.167	**0.162**
	BPR	0.009	0.012	0.016	0.037	0.043	0.040	0.345	0.034	0.053	**0.410**
210	Mean	0.161	0.160	0.170	0.158	0.166	0.155	**0.152**	0.156	0.155	**0.152**
	Median	0.161	0.160	0.169	0.158	0.165	0.155	**0.152**	0.156	0.155	**0.152**
	BPR	0.008	0.008	0.000	0.118	0.020	0.033	**0.383**	0.078	0.035	0.320
320	Mean	0.152	0.152	0.157	0.150	0.155	0.149	0.147	0.149	0.149	**0.147**
	Median	0.152	0.152	0.156	0.151	0.155	0.149	**0.147**	0.149	0.149	0.148
	BPR	0.033	0.005	0.045	0.143	0.033	0.030	**0.328**	0.083	0.015	0.288
420	Mean	0.151	0.151	0.150	0.149	0.151	0.148	**0.147**	0.148	0.148	**0.147**
	Median	0.149	0.149	0.149	0.147	0.150	0.146	**0.145**	0.146	0.146	**0.145**
	BPR	0.018	0.008	0.130	0.188	0.050	0.013	0.223	0.068	0.045	**0.260**

Table 2: The mean, median and BPR of MSE in Case 2 of real data analysis

n		AI	Cp	BI	SA	SB	MM	RM	GM	JM	RJ
110	Mean	0.169	0.168	0.172	0.163	0.164	0.159	**0.153**	0.161	0.159	**0.153**
	Median	0.165	0.164	0.170	0.160	0.163	0.158	**0.152**	0.159	0.158	**0.152**
	BPR	0.019	0.000	0.000	0.082	0.036	0.011	**0.477**	0.014	0.027	0.334
210	Mean	0.151	0.151	0.157	0.148	0.152	0.148	**0.145**	0.148	0.148	**0.145**
	Median	0.152	0.151	0.157	0.149	0.152	0.149	**0.144**	0.149	0.148	**0.144**
	BPR	0.010	0.000	0.003	0.145	0.018	0.003	**0.568**	0.003	0.010	0.243
320	Mean	0.147	0.147	0.152	0.144	0.148	0.145	**0.142**	0.145	0.145	**0.142**
	Median	0.147	0.147	0.152	0.145	0.147	0.145	**0.143**	0.145	0.145	**0.143**
	BPR	0.013	0.000	0.000	0.283	0.025	0.000	**0.428**	0.003	0.013	0.238
420	Mean	0.144	0.144	0.149	0.141	0.145	0.142	**0.140**	0.142	0.142	**0.140**
	Median	0.143	0.143	0.148	0.141	0.143	0.142	**0.140**	0.143	0.142	**0.140**
	BPR	0.015	0.000	0.003	**0.338**	0.040	0.000	0.310	0.000	0.015	0.280