# Inference after Model Averaging in Linear Regression Models*

Xinyu Zhang† and Chu-An Liu‡

June 28, 2018

**Abstract**

This paper considers the problem of inference for nested least squares averaging estimators. We study the asymptotic behavior of the Mallows model averaging estimator (MMA; Hansen, 2007) and the jackknife model averaging estimator (JMA; Hansen and Racine, 2012) under the standard asymptotics with fixed parameters setup. We find that both MMA and JMA estimators asymptotically assign zero weight to the under-fitted models, and MMA and JMA weights of just-fitted and over-fitted models are asymptotically random. Building on the asymptotic behavior of model weights, we derive the asymptotic distributions of MMA and JMA estimators and propose a simulation-based confidence interval for the least squares averaging estimator. Monte Carlo simulations show that the coverage probabilities of proposed confidence intervals achieve the nominal level.

Keywords: Confidence intervals, Inference post-model-averaging, Jackknife model averaging, Mallows model averaging.

JEL Classification: C51, C52

*We thank three anonymous referees, the co-editor Liangjun Su, and the editor Peter C.B. Phillips for many constructive comments and suggestions. We also thank conference participants of SETA 2016, AMES 2016, and CFE 2017 for their discussions and suggestions. Xinyu Zhang gratefully acknowledges the research support from National Natural Science Foundation of China (Grant numbers 71522004, 11471324 and 71631008). Chu-An Liu gratefully acknowledges the research support from the Ministry of Science and Technology of Taiwan (MOST 104-2410-H-001-092-MY2). All errors remain the authors'.

†Academy of Mathematics and Systems Science, Chinese Academy of Sciences. Email: xinyu@amss.ac.cn.

‡Institute of Economics, Academia Sinica. Email: caliu@econ.sinica.edu.tw.

# 1 Introduction

In the past two decades, model averaging from the frequentist perspective has received much attention in both econometrics and statistics. Model averaging considers the uncertainty across different models as well as the model bias from each candidate model via effectively averaging over all potential models. Different methods of weight selection have been proposed based on distinct criteria; see Claeskens and Hjort (2008) and Moral-Benito (2015) for a literature review. Despite the growing literature on frequentist model averaging, little work has been done on examining the asymptotic behavior of the model averaging estimator.

Recently, Hansen (2014) and Liu (2015) study the limiting distributions of the least squares averaging estimators in a local asymptotic framework where the regression coefficients are in a local $n^{-1/2}$ neighborhood of zero. The merit of the local asymptotic framework is that both squared model biases and estimator variances have the same order $O(n^{-1})$. Thus, the asymptotic mean squared error remains finite and provides a good approximation to finite sample mean squared error in this context. However, there has been a discussion about the realism of the local asymptotic framework; see Hjort and Claeskens (2003b) and Raftery and Zheng (2003). Furthermore, the local asymptotic framework induces the local parameters in the asymptotics, which generally cannot be estimated consistently.

In this paper, instead of assuming drifting sequences of parameters, we consider the standard asymptotics with fixed parameters setup and investigate the asymptotic distribution of the nested least squares averaging estimator. Under the fixed parameter framework, we study the asymptotic behavior of model weights selected by the Mallows model averaging (MMA) estimator and the jackknife model averaging (JMA) estimator. We find that both MMA and JMA estimators asymptotically assign zero weight to the under-fitted model, that is, a model with omitted variables. This result implies that both MMA and JMA estimators only average over just-fitted and over-fitted models but not under-fitted models as the sample size goes to infinity. Unlike the weight of the under-fitted model, MMA and JMA weights of just-fitted and over-fitted models have nonstandard limiting distributions, but they could be characterized by a normal random vector. Building on the asymptotic behavior of model weights, we show that the asymptotic distributions of MMA and JMA estimators are both nonstandard and not pivotal.

To address the problem of inference for least squares averaging estimators, we follow Claeskens and Hjort (2008), Lu (2015), and DiTraglia (2016) and consider a simulation-based method to construct the confidence intervals. The idea of the simulation-based confidence

interval is to simulate the limiting distributions of averaging estimators and use this simulated distribution to conduct inference. Unlike the naive method, which ignores the model selection step and takes the selected model as the true model to construct the confidence intervals, the proposed method takes the model averaging step into account and has asymptotically the correct coverage probability. Monte Carlo simulations show that the coverage probabilities of the simulation-based confidence intervals achieve the nominal level, while the naive confidence intervals that ignore the model selection step lead to distorted inference.

As an alternative approach to the simulation-based confidence interval, we consider imposing a larger penalty term in the weight selection criterion such that the resulting weights of over-fitted models could converge to zeros. We show that this modified averaging estimator is asymptotically normal with the same covariance matrix as the least squares estimator for the just-fitted model. Therefore, we can use the critical value of the standard normal distribution to construct the traditional confidence interval.

There are two main limitations of our results. First, we do not demonstrate that the proposed simulation-based confidence intervals are better than those based on the just-fitted or over-fitted models in the asymptotic theory. The simulations show that the average length of the proposed confidence intervals is shorter than those of other estimators. However, this could be a finite sample improvement, and it would be greatly desirable to provide the theoretical justification in a future study. Second, we do not demonstrate any advantage of model averaging in the fixed parameter framework. We show that both MMA and JMA estimators asymptotically average over the just-fitted model along with the over-fitted models. In general, however, there is no advantage of using over-fitting models in the asymptotic theory. Although our simulations show that both MMA and JMA estimators could achieve the mean square error reduction, we do not provide any theoretical justification of this finite sample improvement.

We now discuss the related literature. There are two main model averaging approaches, Bayesian model averaging and frequentist model averaging. Bayesian model averaging has a long history, and has been widely used in statistical and economic analysis; see Hoeting et al. (1999) for a literature review. In contrast to Bayesian model averaging, there is a growing body of literature on frequentist model averaging, including information criterion weighting (Buckland et al., 1997; Hjort and Claeskens, 2003a; Zhang and Liang, 2011; Zhang et al., 2012), adaptive regression by mixing models (Yang, 2000, 2001; Yuan and Yang, 2005), Mallows' $C_p$-type averaging (Hansen, 2007; Wan et al., 2010; Liu and Okui, 2013; Zhang

et al., 2014), optimal mean squared error averaging (Liang et al., 2011), jackknife model averaging (Hansen and Racine, 2012; Zhang et al., 2013; Lu and Su, 2015), and plug-in averaging (Liu, 2015). There are also many alternative approaches to model averaging, for example, bagging (Breiman, 1996; Inoue and Kilian, 2008), LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006), and the model confidence set (Hansen et al., 2011), among others.

There is a large literature on inference after model selection, including Pötscher (1991), Kabaila (1995, 1998), Pötscher and Leeb (2009), and Leeb and Pötscher (2003, 2005, 2006, 2008, 2017). These papers point out that the coverage probabilities of naive confidence intervals are lower than the nominal values. They also claim that no uniformly consistent estimator exists for the conditional and unconditional distributions of post-model-selection estimators.

The existing literature on inference after model averaging is comparatively small. Hjort and Claeskens (2003a) and Claeskens and Hjort (2008) show that the traditional confidence interval based on normal approximations leads to distorted inference. Pötscher (2006) argues that the finite sample distribution of the averaging estimator cannot be uniformly consistently estimated. Our paper is closely related to Hansen (2014) and Liu (2015), who investigate the asymptotic distributions of the least squares averaging estimators in a local asymptotic framework. The main difference is that our limiting distribution is a nonlinear function of the normal random vector with mean zero, while their limiting distributions depend on a nonlinear function of the normal random vector plus the local parameters.

The outline of the paper is as follows. Section 2 presents the model and the averaging estimator. Section 3 presents the MMA and JMA estimators. Section 4 presents the asymptotic framework and derives the limiting distributions of the MMA and JMA estimators. Section 5 proposes a simulation-based confidence interval and a modified least squares averaging estimator with asymptotic normality. Section 6 examines the finite sample properties of proposed methods, and Section 7 concludes the paper. Proofs are included in the Appendix.

# 2 Model and Estimation

We consider a linear regression model:

$$y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + e_i, \tag{1}$$

$$E(e_i|\mathbf{x}_i) = 0, \tag{2}$$

$$E(e_i^2|\mathbf{x}_i) = \sigma^2(\mathbf{x}_i), \tag{3}$$

where $y_i$ is a scalar dependent variable, $\mathbf{x}_i = (\mathbf{x}_{1i}', \mathbf{x}_{2i}')'$, $\mathbf{x}_{1i}(k_1 \times 1)$ and $\mathbf{x}_{2i}(k_2 \times 1)$ are vectors of regressors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown parameter vectors, and $e_i$ is an unobservable regression error. The error term is allowed to be homoskedastic or heteroskedastic, and there is no further assumption on the distribution of the error term. Here, $\mathbf{x}_{1i}$ contain the core regressors that must be included in the model based on theoretical grounds, while $\mathbf{x}_{2i}$ contain the auxiliary regressors that may or may not be included in the model. The auxiliary regressors could be any nonlinear transformations of the original variables or the interaction terms between the regressors. Note that $\mathbf{x}_{1i}$ may only include a constant term or even an empty matrix. The model (1) is widely used in the model averaging literature, for example, Magnus et al. (2010), Liang et al. (2011), and Liu (2015).

Let $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{X}_1 = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n})'$, $\mathbf{X}_2 = (\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n})'$, and $\mathbf{e} = (e_1, \ldots, e_n)'$. In matrix notation, we write the model (1) as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{4}$$

where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. Let $K = k_1 + k_2$ be the number of regressors in the model (4). We assume that $\mathbf{X}$ has full column rank $K$.

Suppose that we have a set of $M$ candidate models. We follow Hansen (2007, 2008, 2014) and consider a sequence of nested candidate models. In most applications, we have $M = k_2 + 1$ candidate models. The $m$th submodel includes all regressors in $\mathbf{X}_1$ and the first $m-1$ regressors in $\mathbf{X}_2$, but excludes the remaining regressors. We use $\mathbf{X}_{2m}$ to denote the auxiliary regressors included in the $m$th submodel. Note that the $m$th model has $k_1 + k_{2m} = K_m$ regressors.

In empirical applications, practitioners can order regressors by some manner or prior and then combine nested models. Similar to Hansen (2014), for all the following theoretical results, we do not impose any assumption on the ordering of regressors, i.e., the ordering is not required to be "correct" in any sense. A candidate model is called *under-fitted* if the model has omitted variables with nonzero slope coefficients. A candidate model is called *just-fitted* if the model has no omitted variable and no irrelevant variable, while a candidate model is called *over-fitted* if the model has no omitted variable but has irrelevant variables.[1] Without loss of generality, we assume that the first $M_0$ candidate models are under-fitted. Obviously, we have $M > M_0 \geqslant 0$.

Let $\mathbf{I}$ denote an identity matrix and $\mathbf{0}$ a zero matrix. Let $\mathbf{\Pi}_m$ be a selection matrix so that $\mathbf{\Pi}_m = (\mathbf{I}_{K_m}, \mathbf{0}_{K_m \times (K-K_m)})$ or a column permutation thereof and thus $\mathbf{X}_m = (\mathbf{X}_1, \mathbf{X}_{2m}) = \mathbf{X}\mathbf{\Pi}_m'$. The least squares estimator of $\boldsymbol{\beta}$ in the $m$th candidate model is $\widehat{\boldsymbol{\beta}}_m = \mathbf{\Pi}_m'(\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{y}$. We now define the least squares averaging estimator of $\boldsymbol{\beta}$. Let $w_m$ be the weight corresponding to the $m$th candidate model and $\mathbf{w} = (w_1, \ldots, w_M)'$ be a weight vector belonging to the weight set $\mathcal{W} = \{\mathbf{w} \in [0,1]^M : \sum_{m=1}^M w_m = 1\}$. That is, the weight vector lies in the unit simplex in $\mathbb{R}^M$. The least squares averaging estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{m=1}^{M} w_m \widehat{\boldsymbol{\beta}}_m. \tag{5}$$

# 3 Least Squares Averaging Estimator

In this section, we consider two commonly used methods of least squares averaging estimators, the Mallows model averaging (MMA) estimator and the jackknife model averaging (JMA) estimator.

Hansen (2007) introduces the Mallows model averaging estimator for the homoskedastic linear regression model. Let $\mathbf{P}_m = \mathbf{X}_m(\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'$ and $\mathbf{P}(\mathbf{w}) = \sum_{m=1}^M w_m \mathbf{P}_m$ be the projection matrices. Let $\|\cdot\|^2$ stand for the Euclidean norm. The MMA estimator selects the model weights by minimizing a Mallows criterion

$$\mathcal{C}(\mathbf{w}) = \|(\mathbf{I}_n - \mathbf{P}(\mathbf{w}))\mathbf{y}\|^2 + 2\sigma^2\mathbf{w}'\mathbf{K}, \tag{6}$$

where $\sigma^2 = E(e_i^2)$ and $\mathbf{K} = (K_1, \ldots, K_M)'$. In practice, $\sigma^2$ can be estimated by $\widehat{\sigma}^2 = (n-K)^{-1}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_M\|^2$. Denote $\widehat{\mathbf{w}}_{\text{MMA}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \mathcal{C}(\mathbf{w})$ as the MMA weights. Note that the criterion function $\mathcal{C}(\mathbf{w})$ is a quadratic function of the weight vector. Therefore, the MMA weights can be found numerically via quadratic programming. The MMA estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}_{\text{MMA}}) = \sum_{m=1}^{M} \widehat{w}_{\text{MMA},m} \widehat{\boldsymbol{\beta}}_m. \tag{7}$$

Hansen (2007) demonstrates the asymptotic optimality of the MMA estimator for nested and homoskedastic linear regression models, i.e., the MMA estimator asymptotically achieves the lowest possible mean squared error among all candidates. However, the optimality of MMA fails under heteroskedasticity (Hansen, 2007).

Hansen and Racine (2012) introduce the jackknife model averaging estimator and demonstrate its optimality in the linear regression model with heteroskedastic errors. Let $h_{ii}^m$ be

the $i$th diagonal element of $\mathbf{P}_m$. Define $\mathbf{D}_m$ as a diagonal matrix with $(1-h_{ii}^m)^{-1}$ being its $i$th diagonal element. Let $\widetilde{\mathbf{P}}_m = \mathbf{D}_m(\mathbf{P}_m - \mathbf{I}_n) + \mathbf{I}_n$ and $\widetilde{\mathbf{P}}(\mathbf{w}) = \sum_{m=1}^M w_m \widetilde{\mathbf{P}}_m$. The JMA estimator selects the weights by minimizing a cross-validation (or jackknife) criterion

$$\mathcal{J}(\mathbf{w}) = \|(\mathbf{I}_n - \widetilde{\mathbf{P}}(\mathbf{w}))\mathbf{y}\|^2. \tag{8}$$

Similar to the MMA estimator, the JMA weights can also be found numerically via quadratic programming. Denote $\widehat{\mathbf{w}}_{\text{JMA}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \mathcal{J}(\mathbf{w})$ as the JMA weights. Thus, the JMA estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}_{\text{JMA}}) = \sum_{m=1}^M \widehat{w}_{\text{JMA},m}\widehat{\boldsymbol{\beta}}_m. \tag{9}$$

Hansen (2007) and Hansen and Racine (2012) demonstrate the asymptotic optimality of the MMA and JMA estimators in homoskedastic and heteroskedastic settings, respectively.[2] To yield a good approximation to the finite sample behavior, Hansen (2014) and Liu (2015) investigate the asymptotic distributions of the MMA and JMA estimators in a local asymptotic framework where the regression coefficients are in a local $n^{-1/2}$ neighborhood of zero. Unlike Hansen (2014) and Liu (2015), which assume a drifting sequence of the parameter, we study the asymptotic distributions of the MMA and JMA estimators under the standard asymptotics with fixed parameters setup in the next section.

# 4 Asymptotic Theory

We first state the regularity conditions required for asymptotic results, where all limiting processes here and throughout the text are with respect to $n \to \infty$.

**Condition (C.1).** $\mathbf{Q}_n \equiv n^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbf{Q}$, *where* $\mathbf{Q} = E(\mathbf{x}_i\mathbf{x}_i')$ *is a positive definite matrix.*

**Condition (C.2).** $\mathbf{Z}_n \equiv n^{-1/2}\mathbf{X}'\mathbf{e} \xrightarrow{d} \mathbf{Z} \sim \mathbf{N}(0, \boldsymbol{\Omega})$, *where* $\boldsymbol{\Omega} = E(\mathbf{x}_i\mathbf{x}_i'e_i^2)$ *is a positive definite matrix.*

**Condition (C.3).** $\bar{h}_n \equiv \max_{1\leqslant m\leqslant M}\max_{1\leqslant i\leqslant n} h_{ii}^m = o_p(n^{-1/2})$.

**Condition (C.4).** $\boldsymbol{\Omega}_n \equiv n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'e_i^2 \xrightarrow{p} \boldsymbol{\Omega}$.

Conditions (C.1), (C.2) and (C.4) are high-level conditions that permits the application of cross-section, panel, and time-series data. Conditions (C.1) and (C.2) hold under appropriate

primitive assumptions. For example, if $y_i$ is a stationary and ergodic martingale difference sequence with finite fourth moments, then these conditions follow from the weak law of large numbers and the central limit theorem for martingale difference sequences. The sufficient condition for Condition (C.4) is that $e_i$ is i.i.d. or a martingale difference sequence with finite fourth moments. Condition (C.3) is quite mild. Note that Li (1987) and Andrews (1991) assumed that $h_{ii}^m \leqslant cK_m n^{-1}$ for some constant $c < \infty$, which is more restrictive than Condition (C.3) under our model (1). Conditions (C.1) and (C.2) are similar to Assumption 2 of Liu (2015). Condition (C.3) is similar to Condition A.9 in Hansen and Racine (2012) and Assumption 2.4 in Liu and Okui (2013). Condition (C.4) is similar to the condition in Theorem 3 of Liu (2015).

## 4.1 Asymptotic Distribution of the MMA Estimator

The weights selected by the MMA estimator are random, and this must be taken into account in the asymptotic distribution of the MMA estimator. The following theorem describes the asymptotic behavior of the MMA weights of under-fitted models.

**Theorem 1.** *Suppose that Conditions (C.1)-(C.2) hold. Then for any $m \in \{1, \ldots, M_0\}$,*

$$\widehat{w}_{\mathrm{MMA},m} = O_p(n^{-1}). \tag{10}$$

Theorem 1 shows that the MMA weights of under-fitted models are $O_p(n^{-1})$. This result implies that the MMA estimator asymptotically assigns zero weight to the model that has omitted variables with nonzero parameters $\boldsymbol{\beta}_2$.

We next study the MMA weights of just-fitted and over-fitted models, i.e., $m \in \{M_0 + 1, ..., M\}$, and the asymptotic distribution of the MMA estimator. Let $S = M - M_0$ be the number of just-fitted and over-fitted models, which is not smaller than 1. Excluding the under-fitted models, we define a new weight vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)'$ that belongs to a weight set

$$\mathcal{L} = \left\{ \boldsymbol{\lambda} \in [0,1]^S : \sum_{s=1}^{S} \lambda_s = 1 \right\}. \tag{11}$$

Note that the weight vector $\boldsymbol{\lambda}$ lies in the unit simplex in $\mathbb{R}^S$. For $s = 1, \ldots, S$, let $\boldsymbol{\Omega}_s = \boldsymbol{\Pi}_{M_0+s} \boldsymbol{\Omega} \boldsymbol{\Pi}'_{M_0+s}$, $\mathbf{Q}_s = \boldsymbol{\Pi}_{M_0+s} \mathbf{Q} \boldsymbol{\Pi}'_{M_0+s}$, and $\mathbf{V}_s = \boldsymbol{\Pi}'_{M_0+s} \mathbf{Q}_s^{-1} \boldsymbol{\Pi}_{M_0+s}$ be the covariance matrices associated with the new weight vector, where $\boldsymbol{\Omega}$ and $\mathbf{Q}$ are defined in Conditions (C.1)-(C.2).

**Theorem 2.** *Suppose that Conditions (C.1)-(C.2) hold. Then we have*

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}_{\text{MMA}}) - \boldsymbol{\beta}) &= \sum_{m=1}^{M_0} \widehat{\mathbf{w}}_{\text{MMA},m}\sqrt{n}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) + \sum_{m=M_0+1}^{M} \widehat{\mathbf{w}}_{\text{MMA},m}\sqrt{n}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \\ &= O_p(n^{-1/2}) + \sum_{m=M_0+1}^{M} \widehat{\mathbf{w}}_{\text{MMA},m}\boldsymbol{\Pi}_m'(\boldsymbol{\Pi}_m\mathbf{Q}_n\boldsymbol{\Pi}_m')^{-1}\boldsymbol{\Pi}_m\mathbf{Z}_n \\ &\to \sum_{s=1}^{S} \widetilde{\lambda}_{\text{MMA},s}\mathbf{V}_s\mathbf{Z} \qquad (12)\end{aligned}$$

*in distribution, where* $\widetilde{\boldsymbol{\lambda}}_{\text{MMA}} = (\widetilde{\lambda}_{\text{MMA},1}, \ldots, \widetilde{\lambda}_{\text{MMA},S})' = \arg\min_{\boldsymbol{\lambda}\in\mathcal{L}} \boldsymbol{\lambda}'\boldsymbol{\Gamma}\boldsymbol{\lambda}$ *and* $\boldsymbol{\Gamma}$ *is an* $S\times S$ *matrix with the* $(s,j)$*th element*

$$\Gamma_{sj} = 2\sigma^2 K_{M_0+s} - \mathbf{Z}'\mathbf{V}_{\max\{s,j\}}\mathbf{Z}. \qquad (13)$$

Theorem 2 shows that the MMA weights of just-fitted and over-fitted models have nonstandard asymptotic distributions since $\boldsymbol{\Gamma}$ is a nonlinear function of the normal random vector $\mathbf{Z}$. Furthermore, the MMA estimator has a nonstandard limiting distribution, which can be expressed in terms of the normal random vector $\mathbf{Z}$. The representation (12) also implies that in the large sample sense, the just-fitted and over-fitted models can receive positive weight, while the under-fitted models receive zero weight. Note that the least squares estimator with more variables tends to have a larger variance in the nested framework. Thus, there is no advantage of using irrelevant regressors or over-fitting models in the asymptotic theory in general.[3]

Hansen (2014) and Liu (2015) also derive the asymptotic distribution of the MMA estimator. Both papers consider the local-to-zero asymptotic framework, that is, $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2n} = \boldsymbol{\delta}/\sqrt{n}$ and $\boldsymbol{\delta}$ is an unknown local parameter. Note that the local parameters generally cannot be estimated consistently. The main difference between Theorem 2 and results in Hansen (2014) and Liu (2015) is that our limiting distribution does not depend on the local parameters.

## 4.2 Asymptotic Distribution of the JMA Estimator

We now study the asymptotic behavior of the JMA weights and the asymptotic distribution of the JMA estimator.

**Theorem 3.** *Suppose that Conditions (C.1)-(C.3) hold. Then for any* $m \in \{1, \ldots, M_0\}$,

$$\widehat{w}_{\text{JMA},m} = o_p(n^{-1/2}). \qquad (14)$$

Similar to Theorem 1, Theorem 3 shows that the JMA estimator asymptotically assigns zero weight to under-fitted models. The next theorem provides the asymptotic distribution of the JMA estimator.

**Theorem 4.** *Suppose that Conditions (C.1)-(C.4) hold. Then we have*

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}_{\text{JMA}}) - \boldsymbol{\beta}) &= \sum_{m=1}^{M_0} \widehat{\mathbf{w}}_{\text{JMA},m}\sqrt{n}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) + \sum_{m=M_0+1}^{M} \widehat{\mathbf{w}}_{\text{JMA},m}\sqrt{n}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \\ &= o_p(1) + \sum_{m=M_0+1}^{M} \widehat{\mathbf{w}}_{\text{JMA},m}\boldsymbol{\Pi}_m'(\boldsymbol{\Pi}_m\mathbf{Q}_n\boldsymbol{\Pi}_m')^{-1}\boldsymbol{\Pi}_m\mathbf{Z}_n \\ &\to \sum_{s=1}^{S} \widetilde{\lambda}_{\text{JMA},s}\mathbf{V}_s\mathbf{Z} \qquad (15)\end{aligned}$$

*in distribution, where* $\widetilde{\boldsymbol{\lambda}}_{\text{JMA}} = (\widetilde{\lambda}_{\text{JMA},1}, \ldots, \widetilde{\lambda}_{\text{JMA},S})' = \arg\min_{\boldsymbol{\lambda}\in\mathcal{L}} \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}$ *and* $\boldsymbol{\Sigma}$ *is an* $S \times S$ *matrix with the* $(s, j)$*th element*

$$\Sigma_{sj} = tr(\mathbf{Q}_s^{-1}\boldsymbol{\Omega}_s) + tr(\mathbf{Q}_j^{-1}\boldsymbol{\Omega}_j) - \mathbf{Z}'\mathbf{V}_{\max\{s,j\}}\mathbf{Z}. \qquad (16)$$

Similar to Theorem 2, Theorem 4 shows that the JMA estimator has a nonstandard asymptotic distribution. The main difference between Theorem 2 and 4 is the limiting behavior of the weight vector, i.e., $\widetilde{\lambda}_{\text{MMA},s}$ and $\widetilde{\lambda}_{\text{JMA},s}$. The first term of $\Gamma_{sj}$ in (13) is the limit of the penalty term of the Mallows criterion, and the second term of $\Gamma_{sj}$ is the limit of the in-sample squared error $\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{M_0+\max\{s,j\}}\|^2$ subtracting the term $\|\mathbf{e}\|^2$, where $\|\mathbf{e}\|^2$ is unrelated to $\boldsymbol{\lambda}$. Since the second term of $\Gamma_{sj}$ is the same as the last term of $\Sigma_{sj}$ in (16), the asymptotic distributions of MMA and JMA estimators differ only in the limit of the penalty terms. Note that for the homoskedastic situation, $\boldsymbol{\Omega} = \sigma^2\mathbf{Q}$, by which we have $\text{tr}(\mathbf{Q}_s^{-1}\boldsymbol{\Omega}_s) = \sigma^2 K_{M_0+s}$, and thus $\widetilde{\lambda}_{\text{MMA},s} = \widetilde{\lambda}_{\text{JMA},s}$. This result means that the limiting distributions of the MMA and JMA estimators are the same for the homoskedastic situation, which is reasonable and expected.

# 5 Inference for Least Squares Averaging Estimators

In this section, we investigate the problem of inference for least squares averaging estimators. In the first subsection, we propose a simulation-based method to construct the confidence intervals. In the second subsection, we propose a modified JMA estimator and demonstrate its asymptotic normality.

## 5.1 Simulation-Based Confidence Intervals

As shown in the previous section, the least squares averaging estimator with data-dependent weights has a nonstandard asymptotic distribution. Since the asymptotic distributions derived in Theorems 2 and 4 are not pivotal, they cannot be directly used for inference. To address this issue, we follow Claeskens and Hjort (2008), Lu (2015), and DiTraglia (2016), and consider a simulation-based method to construct the confidence intervals.

In Theorems 2 and 4, we show that the asymptotic distribution of the least squares averaging estimator is a nonlinear function of unknown parameters $\sigma^2$, $\mathbf{\Omega}$, and $\mathbf{Q}$, and the normal random vector $\mathbf{Z}$. Suppose that $\sigma^2$, $\mathbf{\Omega}$, and $\mathbf{Q}$ were all known. Then, by simulating from $\mathbf{Z}$ defined in Condition (C.2), we could approximate the limiting distributions defined in Theorems 2 and 4 to arbitrary precision. This is the main idea of the simulation-based confidence intervals. In practice, we replace the unknown parameters with the consistent estimators. We then simulate the limiting distributions of least squares averaging estimators and use this simulated distribution to conduct inference.

We now describe the simulation-based confidence intervals in details. Let $\widehat{e}_i$ be the least squares residual from the full model, i.e., $\widehat{e}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_M$, where $\widehat{\boldsymbol{\beta}}_M = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Then, $\widehat{\sigma}^2 = (n-K)^{-1}\sum_{i=1}^n \widehat{e}_i^2$ is the consistent estimator of $\sigma^2$. Also, $\widehat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' = \mathbf{Q}_n$ and $\widehat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\widehat{e}_i^2$ are consistent estimators of $\mathbf{Q}$ and $\mathbf{\Omega}$, respectively. We propose the following algorithm to obtain the simulation-based confidence interval for $\beta_j$.

- Step 1: Estimate the full model and obtain the consistent estimators $\widehat{\sigma}^2$, $\widehat{\mathbf{Q}}$, and $\widehat{\mathbf{\Omega}}$.

- Step 2: Generate a sufficiently large number of $K \times 1$ normal random vector $\mathbf{Z}^{(r)} \sim \mathbf{N}(0, \widehat{\mathbf{\Omega}})$ for $r = 1, ..., R$. For each $r$, we compute the quantities of the asymptotic distributions derived in Theorem 2 or 4 based on the sample analogue $\widehat{\sigma}^2$, $\widehat{\mathbf{Q}}$, and $\widehat{\mathbf{\Omega}}$. That is, we first calculate $\widehat{\mathbf{V}}_s = \mathbf{\Pi}'_{M_0+s}\widehat{\mathbf{Q}}_s^{-1}\mathbf{\Pi}_{M_0+s}$ and $\widehat{\mathbf{Q}}_s = \mathbf{\Pi}_{M_0+s}\widehat{\mathbf{Q}}\mathbf{\Pi}'_{M_0+s}$ for a given $M_0$. We then compute $\sum_{s=1}^S \widetilde{\lambda}_{\text{MMA},s}^{(r)}(M_0)\widehat{\mathbf{V}}_s\mathbf{Z}^{(r)}$ or $\sum_{s=1}^S \widetilde{\lambda}_{\text{JMA},s}^{(r)}(M_0)\widehat{\mathbf{V}}_s\mathbf{Z}^{(r)}$, where $\widetilde{\boldsymbol{\lambda}}_{\text{MMA}}^{(r)}(M_0) = \arg\min_{\boldsymbol{\lambda}\in\mathcal{L}} \boldsymbol{\lambda}'\widehat{\mathbf{\Gamma}}^{(r)}(M_0)\boldsymbol{\lambda}$, $\widetilde{\boldsymbol{\lambda}}_{\text{JMA}}^{(r)} = \arg\min_{\boldsymbol{\lambda}\in\mathcal{L}} \boldsymbol{\lambda}'\widehat{\mathbf{\Sigma}}^{(r)}(M_0)\boldsymbol{\lambda}$, and the $(s,j)$th element of $\widehat{\mathbf{\Gamma}}^{(r)}(M_0)$ and $\widehat{\mathbf{\Sigma}}^{(r)}(M_0)$ are

$$\widehat{\Gamma}_{sj}^{(r)}(M_0) = 2\widehat{\sigma}^2 K_{M_0+s} - \mathbf{Z}^{(r)\prime}\widehat{\mathbf{V}}_{\max\{s,j\}}\mathbf{Z}^{(r)},$$
$$\widehat{\Sigma}_{sj}^{(r)}(M_0) = \text{tr}(\widehat{\mathbf{Q}}_s^{-1}\widehat{\mathbf{\Omega}}_s) + \text{tr}(\widehat{\mathbf{Q}}_j^{-1}\widehat{\mathbf{\Omega}}_j) - \mathbf{Z}^{(r)\prime}\widehat{\mathbf{V}}_{\max\{s,j\}}\mathbf{Z}^{(r)},$$

for $M_0 = 0, ..., M-1$, respectively.

Let $\mathbf{\Lambda}^{(r)}_{\text{MMA},j}(M_0)$ and $\mathbf{\Lambda}^{(r)}_{\text{JMA},j}(M_0)$ be the $j$th component of $\sum_{s=1}^{S} \widetilde{\lambda}^{(r)}_{\text{MMA},s}(M_0)\widehat{\mathbf{V}}_s\mathbf{Z}^{(r)}$ and $\sum_{s=1}^{S} \widetilde{\lambda}^{(r)}_{\text{JMA},s}(M_0)\widehat{\mathbf{V}}_s\mathbf{Z}^{(r)}$, respectively. We then compute

$$\mathbf{\Lambda}^{(r)}_{\text{MMA},j}(\widetilde{\mathbf{w}}_{\text{JMA}}) = \sum_{M_0=0}^{M-1} \widetilde{w}_{\text{JMA},M_0+1}\mathbf{\Lambda}^{(r)}_{\text{MMA},j}(M_0),$$
$$\mathbf{\Lambda}^{(r)}_{\text{JMA},j}(\widetilde{\mathbf{w}}_{\text{JMA}}) = \sum_{M_0=0}^{M-1} \widetilde{w}_{\text{JMA},M_0+1}\mathbf{\Lambda}^{(r)}_{\text{JMA},j}(M_0),$$

where $\widetilde{\mathbf{w}}_{\text{JMA}}$ are the modified JMA weights defined in the next subsection.[4]

- Step 3: Let $\widehat{q}_j(\alpha/2)$ and $\widehat{q}_j(1-\alpha/2)$ be the $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of $\mathbf{\Lambda}^{(r)}_{\text{MMA},j}(\widetilde{\mathbf{w}}_{\text{JMA}})$ or $\mathbf{\Lambda}^{(r)}_{\text{JMA},j}(\widetilde{\mathbf{w}}_{\text{JMA}})$ for $r = 1, ..., R$, respectively.

- Step 4: Let $\widehat{\beta}_j(\widehat{\mathbf{w}})$ be the $j$th component of $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$, where $\widehat{\mathbf{w}}$ is either $\widehat{\mathbf{w}}_{\text{MMA}}$ or $\widehat{\mathbf{w}}_{\text{JMA}}$. The confidence interval of $\beta_j$ is constructed as

$$\text{CI}_n = \left[\widehat{\beta}_j(\widehat{\mathbf{w}}) - n^{-1/2}\widehat{q}_j(1-\alpha/2), \widehat{\beta}_j(\widehat{\mathbf{w}}) - n^{-1/2}\widehat{q}_j(\alpha/2)\right]. \tag{17}$$

Given the consistent estimators $\widehat{\sigma}^2$, $\widehat{\mathbf{Q}}$, and $\widehat{\mathbf{\Omega}}$, the proposed confidence interval $\text{CI}_n$ yields valid inference for $\beta_j$.[5] Thus, the confidence interval $\text{CI}_n$ has asymptotically the correct coverage probability as $R$, $n \to \infty$. [6]

Note that both Lu (2015), and DiTraglia (2016) propose a two-step algorithm to construct the simulation-based confidence intervals since they need to construct a confidence region for the local parameters first. Our proposed algorithm is a one-step procedure since the limiting distributions derived in Theorems 2 and 4 do not depend on the local parameters.

## 5.2 Asymptotic Normality of Averaging Estimators

From the analysis in Section 4, we know that the MMA and JMA weights of under-fitted models converge to zeros, but the weights of the over-fitted models converge to random vectors $\widetilde{\lambda}_{\text{MMA},1}, \ldots, \widetilde{\lambda}_{\text{MMA},S}$ or $\widetilde{\lambda}_{\text{JMA},1}, \ldots, \widetilde{\lambda}_{\text{JMA},S}$. This is the main reason that the asymptotic distributions of the MMA and JMA estimators are nonstandard. As an alternative approach to the simulation-based confidence interval, we can consider a larger penalty term in the weight selection criterion such that the resulting weights of over-fitted models could converge to zeros. Utilizing this idea, we could have asymptotic normality of the least squares averaging estimator.

We now present the details. We add a penalty term to the cross-validation criterion defined in (8) and obtain the following criterion

$$\widetilde{\mathcal{J}}(\mathbf{w}) = \|(\mathbf{I}_n - \widetilde{\mathbf{P}}(\mathbf{w}))\mathbf{y}\|^2 + \phi_n \mathbf{w}'\mathbf{K}, \tag{18}$$

where $\phi_n$ is a tuning parameter to control the penalization level. Note that the modified JMA weights can also be found numerically via quadratic programming. Denote $\widetilde{\mathbf{w}}_{\text{JMA}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \widetilde{\mathcal{J}}(\mathbf{w})$ as the modified JMA weights. Thus, the modified JMA estimator of $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}}(\widetilde{\mathbf{w}}_{\text{JMA}}) = \sum_{m=1}^{M} \widetilde{w}_{\text{JMA},m} \widehat{\boldsymbol{\beta}}_m. \tag{19}$$

The following theorem presents the asymptotic normality of the modified JMA estimator.

**Theorem 5.** *Suppose that Conditions (C.1)-(C.3) hold and $\phi_n \to \infty$. Then for any $m \in \{M_0 + 2, \ldots, M\}$, we have*

$$\widetilde{w}_{\text{JMA},m} = O_p(\phi_n^{-1}). \tag{20}$$

*Further if $\phi_n n^{-1/2} \to 0$, then we have*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}(\widetilde{\mathbf{w}}_{\text{JMA}}) - \boldsymbol{\beta}) \to \mathbf{V}_1\mathbf{Z} \sim \mathbf{N}(0, \boldsymbol{\Pi}'_{M_0+1}\mathbf{Q}_1^{-1}\boldsymbol{\Omega}_1\mathbf{Q}_1^{-1}\boldsymbol{\Pi}_{M_0+1}) \tag{21}$$

*in distribution, where $\mathbf{V}_1 = \boldsymbol{\Pi}'_{M_0+1}\mathbf{Q}_1^{-1}\boldsymbol{\Pi}_{M_0+1}$, $\mathbf{Q}_1 = \boldsymbol{\Pi}_{M_0+1}\mathbf{Q}\boldsymbol{\Pi}'_{M_0+1}$, and $\boldsymbol{\Omega}_1 = \boldsymbol{\Pi}_{M_0+1}\boldsymbol{\Omega}\boldsymbol{\Pi}'_{M_0+1}$.*

The first part of Theorem 5 shows that the modified JMA weights of over-fitted models are $O_p(\phi_n^{-1})$, which implies that the modified JMA estimator asymptotically assigns zero weight to the over-fitted models as $\phi_n \to \infty$. For the tuning parameter $\phi_n$, we suggest to use $\log(n)$, which corresponds to the penalty term in the Bayesian information criterion.

Recall that the $(M_0 + 1)$th model is the just-fitted model. The second result of Theorem 5 shows that the modified JMA estimator is asymptotically normal with the same covariance matrix as the least squares estimator for the just-fitted model. Since the asymptotic distribution is normal, we can use the standard normal critical value to construct the traditional confidence interval for $\beta_j$.

# 6 Simulation Study

In this section, we study the finite sample mean squared error and the coverage probability of the least square averaging estimator in comparison with other alternative approaches to model averaging.

## 6.1 Simulation Setup

We consider a linear regression model with a finite number of regressions

$$y_i = \sum_{j=1}^{k} \beta_j x_{ji} + e_i, \tag{22}$$

where $x_{1i} = 1$ and $(x_{2i}, ..., x_{ki})' \sim N(0, \mathbf{\Sigma_x})$. The diagonal elements of $\mathbf{\Sigma_x}$ are $\rho$, and off-diagonal elements are $\rho^2$. We set $\rho = 0.7$. The error term is generated by $e_i = \sigma_i \eta_i$. For the homoskedastic simulation, $\eta_i$ is generated from a standard normal distribution, and $\sigma_i = 2.5$ for $i = 1, .., n$. For the heteroskedastic simulation, $\eta_i$ is generated from a t-distribution with 4 degrees of freedom, and $\sigma_i = (1 + 2|x_{4i}| + 4|x_{ki}|)/3$ for $i = 1, .., n$. We let $(x_{1i}, x_{2i})$ be the core regressors and consider all other regressors auxiliary. We set $k = 10$ and consider a sequence of nested submodels. Thus, the number of models is $M = 9$.

Three cases of the regression coefficients are studied:

Case 1: $\boldsymbol{\beta} = (1, 1, c, c^2, c^3, c^4, 0, 0, 0, 0)'$,

Case 2: $\boldsymbol{\beta} = (1, 1, c^4, c^3, c^2, c, 0, 0, 0, 0)'$,

Case 3: $\boldsymbol{\beta} = (1, 1, c, c^2, 0, 0, c^3, c^4, 0, 0)'$.

We set $c = 0.5$. The numbers of under-fitted models are $M_0 = 4$, 4, and 6 for Cases 1, 2, and 3, respectively. The regression coefficient is a decreasing sequence for Cases 1 and 3, but not for Case 2. Only in Case 1 is the ordering of regressors correct. The ordering of regressors is not correct from the 2nd to 5th models in Case 2, and the ordering of regressors is not correct from the 4th to 6th models in Case 3.

## 6.2 Comparison with Other Approaches

In the Monte Carlo experiments, we consider the following estimators:

1. Least squares estimator for the just-fitted model (labeled JUST).
2. Least squares estimator for the largest model (labeled FULL).
3. Akaike information criterion model selection estimator (labeled AIC).
4. Bayesian information criterion model selection estimator (labeled BIC).
5. Adaptive LASSO estimator with bootstrap confidence intervals (labeled ALASSO)

6. MMA estimator with simulation-based confidence intervals (labeled MMA-S).
7. JMA estimator with simulation-based confidence intervals (labeled JMA-S).
8. MMA estimator with bootstrap confidence intervals (labeled MMA-B).
9. JMA estimator with bootstrap confidence intervals (labeled JMA-B).
10. Modified JMA estimator (labeled JMA-M).

We briefly discuss each estimator and how to construct the confidence intervals for each estimator. The JUST estimator is the least squares estimator for the $(M_0+1)$th model, while the FULL estimator is the least squares estimator for the $M$th model, i.e., the largest model. Let $\widehat{\beta}_j(m)$ denote the $j$th component of $\widehat{\boldsymbol{\beta}}_m$ for $m = M_0+1$ or $M$. For JUST and FULL, the confidence interval of $\beta_j$ is constructed as

$$\mathrm{CI}_n = \left[\widehat{\beta}_j(m) - z_{1-\alpha/2}s(\widehat{\beta}_j(m)), \widehat{\beta}_j(m) + z_{1-\alpha/2}s(\widehat{\beta}_j(m))\right], \tag{23}$$

where $z_{1-\alpha/2}$ is $1-\alpha/2$ quantile of the standard normal distribution and $s(\widehat{\beta}_j(m))$ is the standard error computed based on the $m$th model.

The AIC criterion for the $m$th model is $\mathrm{AIC}_m = n\log(\widehat{\sigma}_m^2) + 2K_m$, where $\widehat{\sigma}_m^2 = \frac{1}{n}\sum_{i=1}^n \widehat{e}_{mi}^2$ and $\widehat{e}_{mi}$ is the least squares residual from the model $m$. The BIC criterion for the $m$th model is $\mathrm{BIC}_m = n\log(\widehat{\sigma}_m^2) + \log(n)K_m$. For both AIC and BIC, we construct the confidence intervals by a naive method. The naive approach ignores the model selection step and takes the selected model as the true model to construct the confidence intervals. For AIC and BIC, the naive confidence interval of $\beta_j$ is constructed as

$$\mathrm{CI}_n = \left[\widehat{\beta}_j(\widehat{m}) - z_{1-\alpha/2}s(\widehat{\beta}_j(\widehat{m})), \widehat{\beta}_j(\widehat{m}) + z_{1-\alpha/2}s(\widehat{\beta}_j(\widehat{m}))\right], \tag{24}$$

where $\widehat{m}$ is the model selected by the AIC or BIC criterion, $\widehat{\beta}_j(\widehat{m})$ is the coefficient estimator under the selected model, and $s(\widehat{\beta}_j(\widehat{m}))$ is the standard error computed by taking $\widehat{m}$ as the true model.

Zou (2006) proposed the adaptive LASSO estimator that simultaneously performs variable selection and estimation of the nonzero parameters in a linear regression model. The adaptive LASSO estimator minimizes the residual sum of squares subject to an $\ell_1$ penalty:

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^K \frac{|\beta_j|}{|\widehat{\beta}_j|^\gamma}\right), \tag{25}$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, ..., \widehat{\beta}_K)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \widehat{\boldsymbol{\beta}}_M$, $\lambda_n$ is a turning parameter, and $\gamma > 0$.[7]

We follow Chatterjee and Lahiri (2011) and use a residual bootstrap method to construct the confidence intervals for the adaptive LASSO estimator; see Chatterjee and Lahiri (2013) for the higher-order refinement of the bootstrap method, and Camponovo (2015) for a pairs bootstrap method for LASSO estimators. Let $\widetilde{e}_i = y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}$ be the ALASSO residual, and $\widetilde{e}_i^+ = \widetilde{e}_i - n^{-1}\sum_{i=1}^n \widetilde{e}_i$ be the centered value. The random samples $\{e_1^*, ..., e_n^*\}$ are drawn from the centered residuals $\{\widetilde{e}_1^+, ..., \widetilde{e}_n^+\}$ with replacement. The bootstrap samples are constructed by $y_i^* = \mathbf{x}_i'\widetilde{\boldsymbol{\beta}} + e_i^*$, and the bootstrap estimator is

$$\widetilde{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}} \left( \sum_{i=1}^n (y_i^* - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^K \frac{|\beta_j|}{|\widehat{\beta}_j^*|^\gamma} \right), \tag{26}$$

where $\widehat{\boldsymbol{\beta}}^* = (\widehat{\beta}_1^*, ..., \widehat{\beta}_K^*)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*$. Let $\widetilde{q}_j^*(\alpha)$ be the $\alpha$th quantile of the bootstrap distribution of $|\sqrt{n}(\widetilde{\beta}_j^* - \widetilde{\beta}_j)|$, where $\widetilde{\beta}_j^*$ and $\widetilde{\beta}_j$ are $j$th component of $\widetilde{\boldsymbol{\beta}}^*$ and $\widetilde{\boldsymbol{\beta}}$, respectively. The bootstrap confidence interval of $\beta_j$ is constructed as

$$\text{CI}_n = \left[ \widetilde{\beta}_j - n^{-1/2}\widetilde{q}_j^*(\alpha), \widetilde{\beta}_j + n^{-1/2}\widetilde{q}_j^*(\alpha) \right]. \tag{27}$$

The MMA and JMA estimators are defined in (7) and (9), respectively. The simulation-based confidence intervals for MMA-S and JMA-S are based on (17). We also consider a pairs bootstrap method to construct the confidence intervals for both MMA and JMA estimators.[8] More precisely, let $\{\mathbf{z}_1, ..., \mathbf{z}_n\}$ be the observation sample, where $\mathbf{z}_i = (y_i, \mathbf{x}_i')'$. The random samples $\{\mathbf{z}_1^*, ..., \mathbf{z}_n^*\}$ are drawn from $\{\mathbf{z}_1, ..., \mathbf{z}_n\}$ with replacement. Let $\widehat{\mathbf{w}}_{\text{MMA}}^*$ and $\widehat{\mathbf{w}}_{\text{JMA}}^*$ be the bootstrap MMA and JMA weights by minimizing (6) and (8) based on the bootstrap sample $\{\mathbf{z}_1^*, ..., \mathbf{z}_n^*\}$, respectively. The bootstrap MMA and JMA estimators are defined as $\widehat{\boldsymbol{\beta}}^*(\widehat{\mathbf{w}}_{\text{MMA}}^*) = \sum_{m=1}^M \widehat{w}_{\text{MMA},m}^* \widehat{\boldsymbol{\beta}}_m^*$ and $\widehat{\boldsymbol{\beta}}^*(\widehat{\mathbf{w}}_{\text{JMA}}^*) = \sum_{m=1}^M \widehat{w}_{\text{JMA},m}^* \widehat{\boldsymbol{\beta}}_m^*$ where $\widehat{\boldsymbol{\beta}}_m^* = \boldsymbol{\Pi}_m'(\mathbf{X}_m^{*\prime}\mathbf{X}_m^*)^{-1}\mathbf{X}_m^{*\prime}\mathbf{y}^*$. Let $\widehat{\beta}_j^*(\widehat{\mathbf{w}}^*)$ be the $j$th component of $\widehat{\boldsymbol{\beta}}^*(\widehat{\mathbf{w}}^*)$, where $\widehat{\mathbf{w}}^*$ is either $\widehat{\mathbf{w}}_{\text{MMA}}^*$ or $\widehat{\mathbf{w}}_{\text{JMA}}^*$. Let $\widehat{q}_j^*(\alpha)$ be the $\alpha$th quantile of the bootstrap distribution of $|\sqrt{n}(\widehat{\beta}_j^*(\widehat{\mathbf{w}}^*) - \widehat{\beta}_j(\widehat{\mathbf{w}})|$. For MMA-B and JMA-B, the bootstrap confidence interval of $\beta_j$ is constructed as

$$\text{CI}_n = \left[ \widehat{\beta}_j(\widehat{\mathbf{w}}) - n^{-1/2}\widehat{q}_j^*(\alpha), \widehat{\beta}_j(\widehat{\mathbf{w}}) + n^{-1/2}\widehat{q}_j^*(\alpha) \right]. \tag{28}$$

The modified JMA estimator is calculated based on (19). Let $\widehat{\beta}_j(\widetilde{\mathbf{w}}_{\text{JMA}})$ be the $j$th component of $\widehat{\boldsymbol{\beta}}(\widetilde{\mathbf{w}}_{\text{JMA}})$. For JMA-M, the confidence interval of $\beta_j$ is constructed as

$$\text{CI}_n = \left[ \widehat{\beta}_j(\widetilde{\mathbf{w}}_{\text{JMA}}) - z_{1-\alpha/2}s(\widehat{\beta}_j(m)), \widehat{\beta}_j(\widetilde{\mathbf{w}}_{\text{JMA}}) + z_{1-\alpha/2}s(\widehat{\beta}_j(m)) \right], \tag{29}$$

where $m = M_0 + 1$.

For each setting, we generate $R = 499$ and $B = 499$ random samples to construct the simulation-based confidence intervals and bootstrap confidence intervals, respectively. To evaluate the finite sample behavior of each estimator, we focus on the estimate of $\beta_4$. We report the variance (Var), the mean square error (MSE), the median absolute deviation (MAD), the coverage probability of a nominal 95% confidence interval (CP(95)), and the average length of the confidence intervals (Len). The number of Monte Carlo experiments is 500.

## 6.3 Simulation Results

Tables 1 and 2 present the finite sample performance of the least square averaging estimators and other approaches for $n = 100$ and $n = 400$, respectively. We first compare the variance, MSE, and MAD of the least square averaging estimators with alternative estimators. The simulation results show that JMA-M performs well and dominates other estimators in most cases. When the sample size is small, i.e., $n = 100$, MMA and JMA have similar variances and MSEs for the homoskedastic setup, but JMA has a smaller variance and MSE than MMA for the heteroskedastic setup. Both MMA and JMA achieve lower variances and MSEs than other estimators, including JUST, FULL, AIC, and ALASSO, in both homoskedastic and heteroskedastic setups. When the sample size is large, i.e., $n = 400$, the variances and MSEs of MMA and JMA are similar, and both MMA and JMA have smaller variances and MSEs than other estimators except JMA-M. The MAD of most estimators are quite similar, except that JUST and FULL have larger MADs for $n = 100$.

Recall that Theorems 1 and 3 show that both MMA and JMA estimators average over the just-fitted model along with the over-fitted models as $n \to \infty$. In general, there is no advantage of using over-fitting models in the asymptotic theory. The simulations show that it is possible that MMA, JMA, and JMA-M perform better than JUST and FULL in terms of MSE in finite samples.[9] However, we do not have any theoretical justification of this finite sample improvement. A rigorous demonstration is beyond the scope of the present paper and is left for future research.

We now compare the coverage probability and the average length of the confidence intervals of the least square averaging estimators with other estimators. As we expected, the coverage probabilities of JUST and FULL are close to the nominal values in most cases, but the coverage probabilities of the naive confidence intervals for AIC and BIC are much lower than the nominal values. Unlike the naive method, the coverage probabilities of MMA-S

Table 1: Simulation results in three cases for $n = 100$

| | | Homoskedastic setup | | | | | Heteroskedastic setup | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | Var | MSE | MAD | CP(95) | Len | Var | MSE | MAD | CP(95) | Len |
| Case 1 | JUST | 0.244 | 0.244 | 0.329 | 0.950 | 1.988 | 0.341 | 0.343 | 0.415 | 0.934 | 2.033 |
| | FULL | 0.289 | 0.289 | 0.335 | 0.952 | 2.129 | 0.376 | 0.379 | 0.442 | 0.924 | 2.079 |
| | AIC | 0.186 | 0.198 | 0.250 | 0.256 | 1.870 | 0.245 | 0.250 | 0.250 | 0.328 | 1.963 |
| | BIC | 0.060 | 0.106 | 0.250 | 0.020 | 1.749 | 0.085 | 0.125 | 0.250 | 0.030 | 2.047 |
| | ALASSO | 0.224 | 0.229 | 0.250 | 0.960 | 2.033 | 0.257 | 0.260 | 0.250 | 0.956 | 2.080 |
| | MMA-S | 0.102 | 0.119 | 0.250 | 0.964 | 1.396 | 0.134 | 0.149 | 0.250 | 0.948 | 1.409 |
| | JMA-S | 0.100 | 0.118 | 0.250 | 0.966 | 1.396 | 0.113 | 0.129 | 0.250 | 0.962 | 1.408 |
| | MMA-B | 0.102 | 0.119 | 0.250 | 0.984 | 1.824 | 0.134 | 0.149 | 0.250 | 0.988 | 1.957 |
| | JMA-B | 0.100 | 0.118 | 0.250 | 0.982 | 1.799 | 0.113 | 0.129 | 0.250 | 0.990 | 1.926 |
| | JMA-M | 0.072 | 0.098 | 0.250 | 0.984 | 1.988 | 0.082 | 0.105 | 0.250 | 0.992 | 2.033 |
| Case 2 | JUST | 0.291 | 0.292 | 0.340 | 0.932 | 2.007 | 0.322 | 0.322 | 0.390 | 0.934 | 2.048 |
| | FULL | 0.319 | 0.319 | 0.363 | 0.930 | 2.148 | 0.359 | 0.359 | 0.400 | 0.924 | 2.109 |
| | AIC | 0.222 | 0.222 | 0.125 | 0.376 | 1.942 | 0.246 | 0.246 | 0.125 | 0.390 | 1.921 |
| | BIC | 0.074 | 0.082 | 0.125 | 0.018 | 1.867 | 0.093 | 0.097 | 0.125 | 0.040 | 1.919 |
| | ALASSO | 0.217 | 0.217 | 0.125 | 0.942 | 2.045 | 0.243 | 0.243 | 0.125 | 0.934 | 2.060 |
| | MMA-S | 0.126 | 0.127 | 0.135 | 0.958 | 1.434 | 0.138 | 0.138 | 0.160 | 0.936 | 1.442 |
| | JMA-S | 0.123 | 0.125 | 0.137 | 0.958 | 1.434 | 0.121 | 0.122 | 0.138 | 0.952 | 1.441 |
| | MMA-B | 0.126 | 0.127 | 0.135 | 0.978 | 1.824 | 0.138 | 0.138 | 0.160 | 0.984 | 1.945 |
| | JMA-B | 0.123 | 0.125 | 0.137 | 0.976 | 1.800 | 0.121 | 0.122 | 0.138 | 0.986 | 1.907 |
| | JMA-M | 0.091 | 0.093 | 0.125 | 0.982 | 2.007 | 0.091 | 0.092 | 0.125 | 0.992 | 2.048 |
| Case 3 | JUST | 0.291 | 0.291 | 0.377 | 0.950 | 2.083 | 0.417 | 0.417 | 0.422 | 0.906 | 2.107 |
| | FULL | 0.305 | 0.305 | 0.389 | 0.948 | 2.141 | 0.453 | 0.453 | 0.434 | 0.894 | 2.133 |
| | AIC | 0.201 | 0.206 | 0.250 | 0.274 | 1.898 | 0.355 | 0.356 | 0.250 | 0.308 | 2.042 |
| | BIC | 0.065 | 0.103 | 0.250 | 0.014 | 1.763 | 0.166 | 0.184 | 0.250 | 0.030 | 2.006 |
| | ALASSO | 0.222 | 0.224 | 0.250 | 0.968 | 2.042 | 0.389 | 0.389 | 0.250 | 0.912 | 2.136 |
| | MMA-S | 0.110 | 0.120 | 0.250 | 0.974 | 1.411 | 0.203 | 0.208 | 0.250 | 0.912 | 1.451 |
| | JMA-S | 0.106 | 0.117 | 0.250 | 0.976 | 1.411 | 0.175 | 0.182 | 0.250 | 0.934 | 1.462 |
| | MMA-B | 0.110 | 0.120 | 0.250 | 0.992 | 1.845 | 0.203 | 0.208 | 0.250 | 0.962 | 1.973 |
| | JMA-B | 0.106 | 0.117 | 0.250 | 0.990 | 1.815 | 0.175 | 0.182 | 0.250 | 0.958 | 1.949 |
| | JMA-M | 0.077 | 0.094 | 0.250 | 0.994 | 2.083 | 0.139 | 0.151 | 0.250 | 0.972 | 2.107 |

Table 2: Simulation results in three cases for $n = 400$

| | | Homoskedastic setup | | | | | Heteroskedastic setup | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | Var | MSE | MAD | CP(95) | Len | Var | MSE | MAD | CP(95) | Len |
| Case 1 | JUST | 0.063 | 0.063 | 0.173 | 0.942 | 0.974 | 0.081 | 0.082 | 0.186 | 0.952 | 1.099 |
| | FULL | 0.070 | 0.070 | 0.176 | 0.942 | 1.025 | 0.087 | 0.087 | 0.189 | 0.954 | 1.128 |
| | AIC | 0.074 | 0.075 | 0.250 | 0.520 | 0.927 | 0.081 | 0.083 | 0.250 | 0.480 | 1.072 |
| | BIC | 0.058 | 0.083 | 0.250 | 0.084 | 0.896 | 0.055 | 0.084 | 0.250 | 0.092 | 1.075 |
| | ALASSO | 0.076 | 0.080 | 0.250 | 0.962 | 1.016 | 0.087 | 0.093 | 0.250 | 0.960 | 1.082 |
| | MMA-S | 0.049 | 0.054 | 0.213 | 0.942 | 0.770 | 0.053 | 0.060 | 0.222 | 0.956 | 0.874 |
| | JMA-S | 0.049 | 0.054 | 0.212 | 0.940 | 0.770 | 0.048 | 0.057 | 0.227 | 0.954 | 0.842 |
| | MMA-B | 0.049 | 0.054 | 0.213 | 0.962 | 0.911 | 0.053 | 0.060 | 0.222 | 0.984 | 1.028 |
| | JMA-B | 0.049 | 0.054 | 0.212 | 0.962 | 0.909 | 0.048 | 0.057 | 0.227 | 0.986 | 1.014 |
| | JMA-M | 0.043 | 0.053 | 0.235 | 0.972 | 0.974 | 0.041 | 0.056 | 0.250 | 0.990 | 1.099 |
| Case 2 | JUST | 0.067 | 0.067 | 0.182 | 0.938 | 0.974 | 0.079 | 0.079 | 0.178 | 0.958 | 1.108 |
| | FULL | 0.070 | 0.070 | 0.193 | 0.960 | 1.025 | 0.081 | 0.081 | 0.192 | 0.962 | 1.125 |
| | AIC | 0.074 | 0.075 | 0.181 | 0.780 | 0.969 | 0.083 | 0.084 | 0.159 | 0.730 | 1.092 |
| | BIC | 0.054 | 0.056 | 0.125 | 0.120 | 0.934 | 0.062 | 0.063 | 0.125 | 0.134 | 1.071 |
| | ALASSO | 0.057 | 0.057 | 0.125 | 0.976 | 1.013 | 0.065 | 0.066 | 0.125 | 0.972 | 1.078 |
| | MMA-S | 0.049 | 0.049 | 0.150 | 0.954 | 0.808 | 0.055 | 0.055 | 0.140 | 0.952 | 0.913 |
| | JMA-S | 0.049 | 0.049 | 0.149 | 0.954 | 0.808 | 0.049 | 0.050 | 0.131 | 0.966 | 0.886 |
| | MMA-B | 0.049 | 0.049 | 0.150 | 0.976 | 0.892 | 0.055 | 0.055 | 0.140 | 0.972 | 1.005 |
| | JMA-B | 0.049 | 0.049 | 0.149 | 0.974 | 0.890 | 0.049 | 0.050 | 0.131 | 0.976 | 0.982 |
| | JMA-M | 0.042 | 0.042 | 0.130 | 0.980 | 0.974 | 0.042 | 0.043 | 0.125 | 0.986 | 1.108 |
| Case 3 | JUST | 0.068 | 0.068 | 0.174 | 0.948 | 1.004 | 0.089 | 0.090 | 0.206 | 0.952 | 1.114 |
| | FULL | 0.070 | 0.070 | 0.179 | 0.950 | 1.024 | 0.090 | 0.091 | 0.198 | 0.954 | 1.124 |
| | AIC | 0.070 | 0.071 | 0.250 | 0.502 | 0.931 | 0.085 | 0.086 | 0.250 | 0.482 | 1.055 |
| | BIC | 0.047 | 0.079 | 0.250 | 0.062 | 0.910 | 0.062 | 0.086 | 0.250 | 0.088 | 1.065 |
| | ALASSO | 0.077 | 0.080 | 0.250 | 0.974 | 1.024 | 0.085 | 0.089 | 0.250 | 0.970 | 1.078 |
| | MMA-S | 0.045 | 0.049 | 0.201 | 0.962 | 0.766 | 0.056 | 0.061 | 0.240 | 0.962 | 0.872 |
| | JMA-S | 0.045 | 0.049 | 0.200 | 0.960 | 0.766 | 0.050 | 0.057 | 0.239 | 0.966 | 0.841 |
| | MMA-B | 0.045 | 0.049 | 0.201 | 0.988 | 0.911 | 0.056 | 0.061 | 0.240 | 0.986 | 1.023 |
| | JMA-B | 0.045 | 0.049 | 0.200 | 0.988 | 0.909 | 0.050 | 0.057 | 0.239 | 0.988 | 1.008 |
| | JMA-M | 0.039 | 0.047 | 0.222 | 0.988 | 1.004 | 0.044 | 0.056 | 0.250 | 0.994 | 1.114 |

and JMA-S are close to the nominal values in most cases. Furthermore, the average length of the confidence intervals of MMA-S and JMA-S is shorter than those of other estimators. However, we do not demonstrate that the proposed confidence intervals are better than those of JUST and FULL in the asymptotic theory. It would be greatly desirable to provide the theoretical justification in a future study.

The simulation results also show that the coverage probabilities of ALASSO and JMA-M generally achieve the nominal values. The average length of the confidence intervals of JMA-M is the same as that based on JUST, which is shorter than that based on FULL. Both MMA-B and JMA-B perform well, and the average length of the confidence intervals of MMA-B and JMA-B is shorter than those based on JUST and ALASSO. Comparing the results of Cases 1–3, we find that the performance of least square averaging estimators is relatively unaffected by the ordering of regressors.

We now consider an extended setup to investigate the effect of the number of models on the mean squared error and the coverage probability. The data generating process is based on (22) and the regression coefficients are determined by the following rule: $\boldsymbol{\beta} = \left(1, 1, c, c^2, ..., c^{k_0}, \mathbf{0}_{1\times k_0}\right)'$. The number of irrelevant variables $k_0$ is varied between 3, 5, and 7, and hence the numbers of models are 7, 11, and 15 for $k_0 = 3$, 5, and 7, respectively.

Tables 3 and 4 report simulation results for $M = 7$, 11, and 15. The variances and MSEs of most estimators slightly increase as the number of models increases when $n = 400$, but they are quite similar for $M = 7$, 11, and 15 when $n = 100$. Similar to Tables 1 and 2, the coverage probabilities of MMA-S and JMA-S generally achieve the nominal values in most cases, and the average length of the confidence intervals is shorter than those of other estimators in all cases. Overall, the finite sample performance of most estimators is quite robust to different values of $k_0$.

# 7 Conclusion

In this paper, we study the asymptotic behavior of two commonly used model averaging estimators, the MMA and JMA estimators, under the standard asymptotics with fixed parameters setup. We investigate the behavior of the MMA and JMA weights as the sample size goes to infinity, and show that both MMA and JMA estimators have nonstandard asymptotic distributions. To address the problem of inference after model averaging, we provide a simulation-based confidence interval for the least squares averaging estimator and propose

Table 3: Simulation results for different numbers of models for $n = 100$

| | | Homoskedastic setup | | | | | Heteroskedastic setup | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | Var | MSE | MAD | CP(95) | Len | Var | MSE | MAD | CP(95) | Len |
| $M = 7$ | JUST | 0.261 | 0.261 | 0.346 | 0.936 | 1.922 | 0.344 | 0.344 | 0.401 | 0.952 | 2.066 |
| | FULL | 0.285 | 0.285 | 0.352 | 0.934 | 2.069 | 0.385 | 0.385 | 0.408 | 0.932 | 2.129 |
| | AIC | 0.194 | 0.202 | 0.250 | 0.232 | 1.855 | 0.244 | 0.245 | 0.250 | 0.298 | 2.077 |
| | BIC | 0.070 | 0.113 | 0.250 | 0.016 | 1.742 | 0.099 | 0.129 | 0.250 | 0.028 | 2.170 |
| | ALASSO | 0.216 | 0.221 | 0.250 | 0.960 | 2.005 | 0.297 | 0.298 | 0.250 | 0.958 | 2.117 |
| | MMA-S | 0.107 | 0.121 | 0.250 | 0.960 | 1.379 | 0.137 | 0.144 | 0.250 | 0.958 | 1.474 |
| | JMA-S | 0.106 | 0.120 | 0.250 | 0.970 | 1.379 | 0.112 | 0.121 | 0.250 | 0.972 | 1.458 |
| | MMA-B | 0.107 | 0.121 | 0.250 | 0.986 | 1.790 | 0.137 | 0.144 | 0.250 | 0.998 | 1.994 |
| | JMA-B | 0.106 | 0.120 | 0.250 | 0.984 | 1.773 | 0.112 | 0.121 | 0.250 | 0.994 | 1.973 |
| | JMA-M | 0.076 | 0.098 | 0.250 | 0.988 | 1.922 | 0.084 | 0.099 | 0.250 | 0.992 | 2.066 |
| $M = 11$ | JUST | 0.243 | 0.243 | 0.318 | 0.962 | 2.031 | 0.341 | 0.341 | 0.358 | 0.924 | 2.109 |
| | FULL | 0.275 | 0.275 | 0.349 | 0.962 | 2.168 | 0.380 | 0.380 | 0.384 | 0.922 | 2.145 |
| | AIC | 0.181 | 0.186 | 0.250 | 0.334 | 1.887 | 0.256 | 0.259 | 0.250 | 0.316 | 2.057 |
| | BIC | 0.049 | 0.095 | 0.250 | 0.022 | 1.764 | 0.103 | 0.134 | 0.250 | 0.046 | 2.073 |
| | ALASSO | 0.192 | 0.194 | 0.250 | 0.986 | 2.044 | 0.270 | 0.273 | 0.250 | 0.954 | 2.152 |
| | MMA-S | 0.094 | 0.105 | 0.250 | 0.982 | 1.423 | 0.139 | 0.149 | 0.250 | 0.944 | 1.431 |
| | JMA-S | 0.094 | 0.105 | 0.250 | 0.984 | 1.423 | 0.117 | 0.130 | 0.250 | 0.958 | 1.446 |
| | MMA-B | 0.094 | 0.105 | 0.250 | 0.996 | 1.853 | 0.139 | 0.149 | 0.250 | 0.988 | 2.037 |
| | JMA-B | 0.094 | 0.105 | 0.250 | 0.996 | 1.819 | 0.117 | 0.130 | 0.250 | 0.990 | 1.987 |
| | JMA-M | 0.065 | 0.083 | 0.250 | 0.994 | 2.031 | 0.087 | 0.107 | 0.250 | 0.994 | 2.109 |
| $M = 15$ | JUST | 0.269 | 0.269 | 0.367 | 0.968 | 2.098 | 0.391 | 0.392 | 0.370 | 0.916 | 2.180 |
| | FULL | 0.303 | 0.303 | 0.394 | 0.962 | 2.246 | 0.410 | 0.412 | 0.423 | 0.930 | 2.191 |
| | AIC | 0.183 | 0.190 | 0.250 | 0.314 | 1.926 | 0.260 | 0.263 | 0.250 | 0.300 | 2.094 |
| | BIC | 0.054 | 0.096 | 0.250 | 0.022 | 1.850 | 0.092 | 0.125 | 0.250 | 0.026 | 1.769 |
| | ALASSO | 0.190 | 0.195 | 0.250 | 0.974 | 2.076 | 0.288 | 0.293 | 0.250 | 0.942 | 2.177 |
| | MMA-S | 0.094 | 0.108 | 0.250 | 0.982 | 1.448 | 0.144 | 0.155 | 0.250 | 0.950 | 1.408 |
| | JMA-S | 0.096 | 0.110 | 0.250 | 0.978 | 1.448 | 0.116 | 0.130 | 0.250 | 0.968 | 1.444 |
| | MMA-B | 0.094 | 0.108 | 0.250 | 0.998 | 1.946 | 0.144 | 0.155 | 0.250 | 0.988 | 2.135 |
| | JMA-B | 0.096 | 0.110 | 0.250 | 0.994 | 1.879 | 0.116 | 0.130 | 0.250 | 0.988 | 2.070 |
| | JMA-M | 0.067 | 0.088 | 0.250 | 0.998 | 2.098 | 0.086 | 0.104 | 0.250 | 0.988 | 2.180 |

Table 4: Simulation results for different numbers of models for $n = 400$

| | | Homoskedastic setup | | | | | Heteroskedastic setup | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | Var | MSE | MAD | CP(95) | Len | Var | MSE | MAD | CP(95) | Len |
| $M = 7$ | JUST | 0.058 | 0.058 | 0.156 | 0.940 | 0.941 | 0.076 | 0.076 | 0.185 | 0.952 | 1.058 |
| | FULL | 0.066 | 0.066 | 0.176 | 0.928 | 0.999 | 0.077 | 0.077 | 0.173 | 0.960 | 1.085 |
| | AIC | 0.067 | 0.068 | 0.250 | 0.500 | 0.926 | 0.080 | 0.081 | 0.250 | 0.498 | 1.039 |
| | BIC | 0.040 | 0.075 | 0.250 | 0.062 | 0.905 | 0.055 | 0.083 | 0.250 | 0.086 | 1.046 |
| | ALASSO | 0.075 | 0.078 | 0.250 | 0.962 | 0.991 | 0.086 | 0.089 | 0.250 | 0.962 | 1.062 |
| | MMA-S | 0.041 | 0.048 | 0.198 | 0.954 | 0.749 | 0.051 | 0.056 | 0.222 | 0.958 | 0.847 |
| | JMA-S | 0.041 | 0.048 | 0.198 | 0.950 | 0.749 | 0.047 | 0.054 | 0.222 | 0.962 | 0.819 |
| | MMA-B | 0.041 | 0.048 | 0.198 | 0.982 | 0.900 | 0.051 | 0.056 | 0.222 | 0.988 | 0.996 |
| | JMA-B | 0.041 | 0.048 | 0.198 | 0.978 | 0.899 | 0.047 | 0.054 | 0.222 | 0.986 | 0.985 |
| | JMA-M | 0.035 | 0.047 | 0.237 | 0.992 | 0.941 | 0.040 | 0.053 | 0.250 | 0.990 | 1.058 |
| $M = 11$ | JUST | 0.063 | 0.063 | 0.185 | 0.962 | 0.992 | 0.084 | 0.084 | 0.200 | 0.940 | 1.104 |
| | FULL | 0.069 | 0.069 | 0.185 | 0.958 | 1.038 | 0.090 | 0.090 | 0.195 | 0.934 | 1.129 |
| | AIC | 0.067 | 0.068 | 0.250 | 0.544 | 0.934 | 0.088 | 0.089 | 0.250 | 0.518 | 1.058 |
| | BIC | 0.046 | 0.077 | 0.250 | 0.076 | 0.900 | 0.071 | 0.091 | 0.250 | 0.096 | 1.033 |
| | ALASSO | 0.072 | 0.076 | 0.250 | 0.982 | 1.032 | 0.092 | 0.095 | 0.250 | 0.952 | 1.076 |
| | MMA-S | 0.042 | 0.048 | 0.193 | 0.962 | 0.772 | 0.060 | 0.064 | 0.219 | 0.950 | 0.878 |
| | JMA-S | 0.042 | 0.048 | 0.198 | 0.962 | 0.772 | 0.056 | 0.062 | 0.224 | 0.948 | 0.850 |
| | MMA-B | 0.042 | 0.048 | 0.193 | 0.986 | 0.905 | 0.060 | 0.064 | 0.219 | 0.976 | 1.027 |
| | JMA-B | 0.042 | 0.048 | 0.198 | 0.986 | 0.905 | 0.056 | 0.062 | 0.224 | 0.980 | 1.014 |
| | JMA-M | 0.036 | 0.047 | 0.227 | 0.992 | 0.992 | 0.050 | 0.059 | 0.247 | 0.988 | 1.104 |
| $M = 15$ | JUST | 0.071 | 0.071 | 0.179 | 0.948 | 1.014 | 0.088 | 0.089 | 0.192 | 0.940 | 1.107 |
| | FULL | 0.075 | 0.076 | 0.182 | 0.948 | 1.054 | 0.092 | 0.092 | 0.208 | 0.940 | 1.127 |
| | AIC | 0.076 | 0.076 | 0.250 | 0.562 | 0.928 | 0.078 | 0.079 | 0.250 | 0.528 | 1.053 |
| | BIC | 0.067 | 0.084 | 0.250 | 0.126 | 0.891 | 0.061 | 0.084 | 0.250 | 0.116 | 1.040 |
| | ALASSO | 0.084 | 0.085 | 0.250 | 0.974 | 1.055 | 0.085 | 0.090 | 0.250 | 0.962 | 1.085 |
| | MMA-S | 0.050 | 0.052 | 0.208 | 0.956 | 0.779 | 0.052 | 0.057 | 0.219 | 0.948 | 0.864 |
| | JMA-S | 0.050 | 0.052 | 0.208 | 0.956 | 0.779 | 0.048 | 0.055 | 0.216 | 0.946 | 0.837 |
| | MMA-B | 0.050 | 0.052 | 0.208 | 0.990 | 0.927 | 0.052 | 0.057 | 0.219 | 0.980 | 1.027 |
| | JMA-B | 0.050 | 0.052 | 0.208 | 0.990 | 0.925 | 0.048 | 0.055 | 0.216 | 0.978 | 1.010 |
| | JMA-M | 0.045 | 0.050 | 0.229 | 0.990 | 1.014 | 0.042 | 0.053 | 0.240 | 0.990 | 1.107 |

a modified JMA estimator with asymptotic normality. The simulation results show that the coverage probabilities of the proposed methods generally achieve the nominal values, and both MMA and JMA estimators can provide the MSE reduction in the fixed parameter framework. However, we do not provide any theoretical justification of this finite sample improvement, and it would be greatly desirable to demonstrate the theoretical justification in a future study. Another possible extension would be to extend the proposed inference method to non-nested candidate models.[10]

# Notes

[1]In the case where there is no true model among all candidate models, i.e., all candidate models have omitted variables or irrelevant variables, the just-fitted model is the model that has no omitted variable and the smallest number of irrelevant variables, and the over-fitted model is the model that has no omitted variable but more irrelevant variables than the just-fitted model.

[2]It is possible that the MMA and JMA estimators are not asymptotically optimal in our framework. This is because the condition (15) of Hansen (2007) and the condition (A.7) of Hansen and Racine (2012) do not hold under the standard asymptotics with a finite number of regressions. These sufficient conditions require that there be no submodel $m$ for which the bias is zero, which does not hold in our framework since the just-fitted and over-fitted models have no bias.

[3]When the error term is heteroskedastic, it is possible that adding an irrelevant variable could decrease the estimation variance; see the example on pages 209–210 of Hansen (2017).

[4]Note that the value of $M_0$ is unknown in practice. As suggested by a referee, we average over all models when we simulate the asymptotic distribution. Based on Theorem 5, one would expect the modified JMA weights of under-fitted and over-fitted models should be small in the finite sample.

[5]The proposed simulation-based method can be easily extended to joint tests. Suppose that the parameter of interest is $\boldsymbol{\theta} = g(\boldsymbol{\beta})$ for some function $g : \mathbb{R}^K \to \mathbb{R}^L$. Let $\widehat{\boldsymbol{\theta}} = g(\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}_{\mathrm{MMA}}))$ be the estimate of $\boldsymbol{\theta}$. Applying the delta method to Theorem 2, we have $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to \sum_{s=1}^{S} \widetilde{\lambda}_{\mathrm{MMA},s} \mathbf{G}' \mathbf{V}_s \mathbf{Z}$ in distribution, where $\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\beta}} g(\boldsymbol{\beta})'$. Then we can conduct joint tests similarly to the proposed algorithm.

[6]Note that our asymptotic results are pointwise but not uniform. Although developing the uniform inference results is important, such an investigation is beyond the scope of this paper, and thus it is left for future research.

[7]In the simulations, we set $\gamma = 2$ and select the turning parameter $\lambda_n$ by the generalized cross-validation method.

[8]As an alternative, one could consider a residual bootstrap method to construct the confidence intervals for MMA and JMA. However, the simulation shows that the residual bootstrap method does not perform well as the pairs bootstrap method.

[9]Our simulations show that the MMA, JMA, and JMA-M methods often assign positive weights to under-fitted models, and these models generally have smaller variances than JUST and FULL. This may be the reason that MMA, JMA, and JMA-M achieve smaller MESs than JUST and FULL in finite samples. To eliminate the effects of under-fitted models, we also consider the case where the under-fitted models are not included in the candidate models. The simulation results show that the MSEs of MMA, JMA, and JMA-M are larger than those of JUST, but smaller than those of FULL in this case.

[10]It is not straightforward to extend our results to the non-nested models. This is because there is no simple relationship between the squared sum of residuals of the just-fitted or over-fitted model with the product of residual vectors of two non-nested under-fitted models.

# References


Andrews, D. W. K. (1991). Asymptotic optimality of generalized $C_L$, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics 47*, 359–377.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997). Model selection: An integral part of inference. *Biometrics 53*, 603–618.

Camponovo, L. (2015). On the validity of the pairs bootstrap for lasso estimators. *Biometrika 102*(4), 981–987.

Chatterjee, A. and S. N. Lahiri (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association 106*(494), 608–625.

Chatterjee, A. and S. N. Lahiri (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics 41*(3), 1232–1259.

Claeskens, G. and N. L. Hjort (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

DiTraglia, F. (2016). Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics 195*, 187–208.

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

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

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

Hansen, B. E. (2017). Econometrics. Unpublished Manuscript, University of Wisconsin.

Hansen, B. E. and J. Racine (2012). Jacknife model averaging. *Journal of Econometrics 167*, 38–46.

Hansen, P., A. Lunde, and J. Nason (2011). The model confidence set. *Econometrica 79*, 453–497.

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

Hjort, N. L. and G. Claeskens (2003b). Rejoinder to the focused information criterion and frequentist model average estimators. *Journal of the American Statistical Association 98*(464), 938–945.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*, 382–417.

Inoue, A. and L. Kilian (2008). How useful is bagging in forecasting economic time series? A case study of US consumer price inflation. *Journal of the American Statistical Association 103*, 511–522.

Kabaila, P. (1995). The effect of model selection on confidence regions and prediction regions. *Econometric Theory 11*, 537–537.

Kabaila, P. (1998). Valid confidence intervals in regression after variable selection. *Econometric Theory 14*(4), 463–482.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *The Annals of Statistics 18*, 191–219.

Leeb, H. and B. Pötscher (2003). The finite-sample distribution of post-model-selection estimators and uniform versus non-uniform approximations. *Econometric Theory 19*(1), 100–142.

Leeb, H. and B. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(1), 21–59.

Leeb, H. and B. Pötscher (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics 34*(5), 2554–2591.

Leeb, H. and B. Pötscher (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory 24*(02), 338–376.

Leeb, H. and B. Pötscher (2017). Testing in the presence of nuisance parameters: Some comments on tests post-model-selection and random critical values. In S. E. Ahmed (Ed.), *Big and Complex Data Analysis: Methodologies and Applications*, pp. 69–82. Springer International Publishing.

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

Liang, H., G. Zou, A. T. K. Wan, and X. Zhang (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association 106*, 1053–1066.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics 186*, 142–159.

Liu, Q. and R. Okui (2013). Heteroscedasticity-robust $C_p$ model averaging. *Econometrics Journal 16*, 462–473.

Lu, X. (2015). A covariate selection criterion for estimation of treatment effects. *Journal of Business and Economic Statistics 33*, 506–522.

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

Magnus, J., O. Powell, and P. Prüfer (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics 154*(2), 139–153.

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

Pötscher, B. (1991). Effects of model selection on inference. *Econometric Theory 7*(2), 163–185.

Pötscher, B. (2006). The distribution of model averaging estimators and an impossibility result regarding its estimation. *Lecture Notes-Monograph Series 52*, 113–129.

Pötscher, B. and H. Leeb (2009). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis 100*(9), 2065–2082.

Raftery, A. E. and Y. Zheng (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association 98*(464), 931–938.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.* Series B *58*, 267–288.

Van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes.* Springer Verlag.

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

Yang, Y. (2000). Combining different procedures for adaptive regression. *Journal of Multivariate Analysis 74*(1), 135–161.

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

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

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

Zhang, X., A. T. Wan, and S. Z. Zhou (2012). Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. *Journal of Business and Economic Statistics 30*, 132–142.

Zhang, X., A. T. K. Wan, and G. Zou (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics 174*, 82–94.

Zhang, X., G. Zou, and H. Liang (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika 101*, 205–218.

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

# Appendix

**Proof of Theorem 1.** Let $\mathbf{l} = (1, \ldots, 1)'$, an $M$-dimensional vector. Since $\sum_{m=1}^{M} w_m = 1$, we have $\mathbf{w}'\mathbf{K} = \mathbf{w}'\mathbf{K}\mathbf{l}'\mathbf{w} = \mathbf{w}'\mathbf{l}\mathbf{K}'\mathbf{w}$. Thus, $2\mathbf{w}'\mathbf{K} = \mathbf{w}'(K_m + K_j)_{m,j\in\{1,\ldots,M\}}\mathbf{w}$. Let $a_m = \mathbf{y}'(\mathbf{I}_n - \mathbf{P}_m)\mathbf{y}$ and $\mathbf{\Phi}$ be an $M \times M$ matrix with the $mj^{\text{th}}$ element

$$\Phi_{mj} = a_{\max\{m,j\}} + \widehat{\sigma}^2(K_m + K_j). \tag{A.1}$$

It is easy to verify that $\mathcal{C}(\mathbf{w}) = \mathbf{w}'\mathbf{\Phi}\mathbf{w}$ for any $\mathbf{w} \in \mathcal{W}$, and $a_m \leqslant a_j$ for $m > j$. Let $m$ be an index belonging to $\{1, \ldots, M_0\}$. Define

$$\widetilde{\mathbf{w}}_m = (\widehat{w}_{\text{MMA},1}, \ldots, \widehat{w}_{\text{MMA},m-1}, 0, \widehat{w}_{\text{MMA},m+1}, \ldots, \widehat{w}_{\text{MMA},M_0}, \ldots, \widehat{w}_{\text{MMA},M} + \widehat{w}_{\text{MMA},m})'.$$

Then it follows that

$$\begin{aligned}
0 &\leqslant \mathcal{C}(\widetilde{\mathbf{w}}_m) - \mathcal{C}(\widehat{\mathbf{w}}_{\text{MMA}}) \\
&= \widetilde{\mathbf{w}}_m'\mathbf{\Phi}\widetilde{\mathbf{w}}_m - \widehat{\mathbf{w}}_{\text{MMA}}'\mathbf{\Phi}\widehat{\mathbf{w}}_{\text{MMA}} \\
&= (\widetilde{\mathbf{w}}_m + \widehat{\mathbf{w}}_{\text{MMA}})'\mathbf{\Phi}(\widetilde{\mathbf{w}}_m - \widehat{\mathbf{w}}_{\text{MMA}}) \\
&= (2\widehat{\mathbf{w}}_{\text{MMA}}' + (0, \ldots, 0, -\widehat{w}_{\text{MMA},m}, 0, \ldots, 0, \widehat{w}_{\text{MMA},m}))\mathbf{\Phi}(0, \ldots, 0, -\widehat{w}_{\text{MMA},m}, 0, \ldots, 0, \widehat{w}_{\text{MMA},m})' \\
&= \widehat{w}_{\text{MMA},m}^2(a_m - a_M) + 2\widehat{\mathbf{w}}_{\text{MMA}}'\mathbf{\Phi}(0, \ldots, 0, -\widehat{w}_{\text{MMA},m}, 0, \ldots, 0, \widehat{w}_{\text{MMA},m})' \\
&= \widehat{w}_{\text{MMA},m}^2(a_m - a_M) + 2\widehat{w}_{\text{MMA},m}\widehat{\mathbf{w}}_{\text{MMA}}'(\Phi_{1M} - \Phi_{1m}, \ldots, \Phi_{MM} - \Phi_{Mm})' \\
&= \widehat{w}_{\text{MMA},m}^2(a_m - a_M) + 2\widehat{w}_{\text{MMA},m}\sum_{j=1}^{M}\widehat{w}_{\text{MMA},j}(\Phi_{Mj} - \Phi_{mj}) \\
&= \widehat{w}_{\text{MMA},m}^2(a_m - a_M) + 2\widehat{w}_{\text{MMA},m}\sum_{j=1}^{M}\widehat{w}_{\text{MMA},j}(a_M - a_{\max\{mj\}} + \widehat{\sigma}^2K_M - \widehat{\sigma}^2K_m) \\
&\leqslant \widehat{w}_{\text{MMA},m}^2(a_m - a_M) + 2\widehat{w}_{\text{MMA},m}^2(a_M - a_m) + 2\widehat{w}_{\text{MMA},m}\widehat{\sigma}^2\sum_{j=1}^{M}\widehat{w}_{\text{MMA},j}(K_M - K_m) \\
&= \widehat{w}_{\text{MMA},m}^2(a_m - a_M) + 2\widehat{w}_{\text{MMA},m}^2(a_M - a_m) + 2\widehat{w}_{\text{MMA},m}\widehat{\sigma}^2(K_M - K_m).
\end{aligned} \tag{A.2}$$

Thus, when $\widehat{w}_{\text{MMA},m} \neq 0$, we have

$$\widehat{w}_{\text{MMA},m} \leqslant (a_m - a_M)^{-1}2\widehat{\sigma}^2(K_M - K_m). \tag{A.3}$$

Let $\boldsymbol{\beta}_{m^c} = \boldsymbol{\Pi}_{m^c}\boldsymbol{\beta}$. It is straightforward to show that for any $m \in \{1, \ldots, M_0\}$,

$$\begin{aligned} & a_m - a_M \\ = \; & (\mathbf{e} + \mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{e} + \mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c}) - \mathbf{e}'(\mathbf{I}_n - \mathbf{P}_M)\mathbf{e} \\ = \; & (\mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\mathbf{I}_n - \mathbf{P}_s)(\mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c}) + 2\mathbf{e}'(\mathbf{I}_n - \mathbf{P}_m)\mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c} - \mathbf{e}'(\mathbf{P}_s - \mathbf{P}_M)\mathbf{e}. \quad \text{(A.4)} \end{aligned}$$

From Conditions (C.1)-(C.2), for any $j \in \{1, \ldots, M\}$, we obtain that

$$\mathbf{e}'\mathbf{P}_j\mathbf{e} = O_p(1), \qquad \mathbf{e}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{X}_{j^c}\boldsymbol{\beta}_{j^c} = O_p(1) \tag{A.5}$$

and

$$\widehat{\sigma}^2 = O_p(1). \tag{A.6}$$

It follows from Condition (C.1) that there exists a positive definite matrix $\widetilde{\mathbf{Q}}$ such that

$$n^{-1}[\mathbf{X}_m, \mathbf{X}_{m^c}]'[\mathbf{X}_m, \mathbf{X}_{m^c}] = n^{-1}\begin{bmatrix} \mathbf{X}_m'\mathbf{X}_m & \mathbf{X}_m'\mathbf{X}_{m^c} \\ \mathbf{X}_m'\mathbf{X}_m & \mathbf{X}_m'\mathbf{X}_{m^c} \end{bmatrix} \to \widetilde{\mathbf{Q}} = \begin{bmatrix} \widetilde{\mathbf{Q}}_{11} & \widetilde{\mathbf{Q}}_{12} \\ \widetilde{\mathbf{Q}}_{21} & \widetilde{\mathbf{Q}}_{22} \end{bmatrix}.$$

In addition, $\mathbf{X}$ is assumed to be of full column rank, and it is well known that

$$\begin{vmatrix} \widetilde{\mathbf{Q}}_{11} & \widetilde{\mathbf{Q}}_{12} \\ \widetilde{\mathbf{Q}}_{21} & \widetilde{\mathbf{Q}}_{22} \end{vmatrix} = |\widetilde{\mathbf{Q}}_{11}||\widetilde{\mathbf{Q}}_{22} - \widetilde{\mathbf{Q}}_{21}\widetilde{\mathbf{Q}}_{11}^{-1}\widetilde{\mathbf{Q}}_{12}|,$$

so $|\widetilde{\mathbf{Q}}_{22} - \widetilde{\mathbf{Q}}_{21}\widetilde{\mathbf{Q}}_{11}^{-1}\widetilde{\mathbf{Q}}_{12}| > 0$, which, along with Condition (C.1), implies

$$\begin{aligned} & n^{-1}(\mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\mathbf{I}_n - \mathbf{P}_s)(\mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c}) \\ = \; & n^{-1}\boldsymbol{\beta}_{m^c}'\{\mathbf{X}_m'\mathbf{X}_{m^c} - \mathbf{X}_m'\mathbf{X}_m(\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{X}_{m^c}\}\boldsymbol{\beta}_{m^c} \\ \to \; & \boldsymbol{\beta}_{m^c}'(\widetilde{\mathbf{Q}}_{22} - \widetilde{\mathbf{Q}}_{21}\widetilde{\mathbf{Q}}_{11}'\widetilde{\mathbf{Q}}_{12})\boldsymbol{\beta}_{m^c} > 0. \quad \text{(A.7)} \end{aligned}$$

From (A.4), (A.5) and (A.7), we have

$$n(a_m - a_M)^{-1} = O_p(1). \tag{A.8}$$

The result (10) is implied by (A.3), (A.6) and (A.8). ■

**Proof of Theorem 2.** Let $\boldsymbol{\Phi}^* = \boldsymbol{\Phi} - \|\mathbf{e}\|^2\mathbf{l}\mathbf{l}'$, where the second term $\|\mathbf{e}\|^2\mathbf{l}\mathbf{l}'$ is unrelated to $\mathbf{w}$. Therefore, we have

$$\widehat{\mathbf{w}}_{\text{MMA}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \mathbf{w}'\boldsymbol{\Phi}^*\mathbf{w}.$$

Rewrite $\widehat{\mathbf{w}}_{\text{MMA}} = (\widehat{\mathbf{w}}_1', \widehat{\mathbf{w}}_2')'$ such that $\widehat{\mathbf{w}}_1$ contains weights of under-fitted models. Correspondingly, we also rewrite $\boldsymbol{\Phi}^*$ as

$$\boldsymbol{\Phi}^* = \begin{pmatrix} \boldsymbol{\Phi}_{11}^* & \boldsymbol{\Phi}_{12}^* \\ \boldsymbol{\Phi}_{21}^* & \boldsymbol{\Phi}_{22}^* \end{pmatrix}.$$

From Conditions (C.1)-(C.2), we have $n^{-1}(a_m - \|\mathbf{e}\|^2) = O_p(1)$ for $1 \leqslant m \leqslant M_0$ and

$$a_m - \|\mathbf{e}\|^2 = O_p(1)$$

for $M_0 < m \leqslant M$. Thus, by (10) we have

$$\widehat{\tau}_1 \equiv \widehat{\mathbf{w}}_1' \mathbf{\Phi}_{11}^* \widehat{\mathbf{w}}_1 = o_p(1) \quad and \quad \widehat{\tau}_2 \equiv \widehat{\mathbf{w}}_1' \mathbf{\Phi}_{12}^* \widehat{\mathbf{w}}_2 = o_p(1), \tag{A.9}$$

where $\widehat{\tau}_1$ and $\widehat{\tau}_2$ are two scales. Let $S = M - M_0$ so that $\mathbf{\Phi}_{22}^*$ is an $S \times S$ matrix. From (A.1), we know that the $(s, j)$th element of $\mathbf{\Phi}_{22}^*$ can be written as

$$\Phi_{22,sj}^* = \widehat{\sigma}^2(K_{M_0+s} + K_{M_0+j}) - \mathbf{e}'\mathbf{P}_{M_0+\max\{s,j\}}\mathbf{e},$$

which converges to $\Gamma_{sj}$ (defined in (13)) in distribution. As in the proof of Theorem 3 of Liu (2015), by Theorem 3.2.2 of Van der Vaart and Wellner (1996) or Theorem 2.7 of Kim and Pollard (1990), we have $\widehat{\mathbf{w}}_2 \to \widetilde{\boldsymbol{\lambda}}_{\text{MMA}}$ in distribution. From Conditions (C.1)-(C.2), we know that for any $m \in \{1, \ldots, M_0\}$,

$$\widehat{\boldsymbol{\beta}}_m = \mathbf{\Pi}_m'(\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{y} = \mathbf{\Pi}_m'(\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{X}\boldsymbol{\beta} + \mathbf{\Pi}_m'(\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{e} = O_p(1). \tag{A.10}$$

Thus, from (10), we have

$$\begin{aligned}
& \sqrt{n}\left(\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}}_{\text{MMA}}) - \boldsymbol{\beta}\right) \\
= & \sum_{m=1}^{M_0} \widehat{\mathbf{w}}_{\text{MMA},m}\sqrt{n}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) + \sum_{m=M_0+1}^{M} \widehat{\mathbf{w}}_{\text{MMA},m}\sqrt{n}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) \\
= & \sum_{m=1}^{M_0} \widehat{\mathbf{w}}_{\text{MMA},m}\sqrt{n}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}) + \sum_{m=M_0+1}^{M} \widehat{\mathbf{w}}_{\text{MMA},m}\sqrt{n}\mathbf{\Pi}_m'(\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{e} \\
= & O_p(n^{-1/2}) + \sum_{m=M_0+1}^{M} \widehat{\mathbf{w}}_{\text{MMA},m}\mathbf{\Pi}_m'(\mathbf{\Pi}_m\mathbf{Q}_n\mathbf{\Pi}_m')^{-1}\mathbf{\Pi}_m\mathbf{Z}_n.
\end{aligned} \tag{A.11}$$

In addition, both $\widetilde{\lambda}_{\text{MMA},s}$ and $\mathbf{V}_s$ can be expressed in terms of $\mathbf{Z}$ and $\mathbf{Q}$. Thus, the result (12) holds. ■

**Proof of Theorem 3.** Denote $\mathbf{C}_m$ as an $n \times n$ diagonal matrix with the $i$th diagonal element

$$C_{m,ii} = h_{ii}^m/(1 - h_{ii}^m).$$

Therefore, we have $\mathbf{D}_m = \mathbf{C}_m + \mathbf{I}_n$. Let $\mathbf{\Psi}$ be an $M \times M$ matrix with the $(m, j)$th element

$$\Psi_{mj} = (\mathbf{e} + \mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{e} + \mathbf{X}_{j^c}\boldsymbol{\beta}_{j^c}) - 2K_m\widehat{\sigma}^2.$$

Therefore, it follows that

$$\mathbf{y}'(\mathbf{I}_n - \mathbf{P}_m)\mathbf{D}_m\mathbf{D}_j(\mathbf{I}_n - \mathbf{P}_j)\mathbf{y}$$

$$
\begin{aligned}
&= \mathbf{y}'(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{I}_n + \mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)\mathbf{y} \\
&= a_{\max\{m,j\}} + \mathbf{y}'(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)\mathbf{y} \\
&= a_{\max\{m,j\}} + (\mathbf{e} + \mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{e} + \mathbf{X}_{j^c}\boldsymbol{\beta}_{j^c}) \\
&= \Phi_{mj} + (\mathbf{e} + \mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{e} + \mathbf{X}_{j^c}\boldsymbol{\beta}_{j^c}) \\
&\quad -\widehat{\sigma}^2(K_m + K_j) \\
&= \Phi_{mj} + \Psi_{mj} + \widehat{\sigma}^2(K_m - K_j).
\end{aligned}
$$

Then we have

$$
\mathcal{J}(\mathbf{w}) = \mathcal{C}(\mathbf{w}) + \mathbf{w}'\boldsymbol{\Psi}\mathbf{w}. \tag{A.12}
$$

Let $m$ be an index belonging to $\{1, \ldots, M_0\}$. Define

$$
\overline{\mathbf{w}}_m = (\widehat{w}_{\text{JMA},1}, \ldots, \widehat{w}_{\text{JMA},m-1}, 0, \widehat{w}_{\text{JMA},m+1}, \ldots, \widehat{w}_{\text{JMA},M_0}, \ldots, \widehat{w}_{\text{JMA},M} + \widehat{w}_{\text{JMA},m})'.
$$

Using (A.12), we have

$$
\begin{aligned}
0 &\leqslant \mathcal{J}(\overline{\mathbf{w}}_m) - \mathcal{J}(\widehat{\mathbf{w}}_{\text{JMA}}) \\
&= \mathcal{C}(\overline{\mathbf{w}}_m) - \mathcal{C}(\widehat{\mathbf{w}}_{\text{JMA}}) + \overline{\mathbf{w}}_m'\boldsymbol{\Psi}\overline{\mathbf{w}}_m - \widehat{\mathbf{w}}_{\text{JMA}}'\boldsymbol{\Psi}\widehat{\mathbf{w}}_{\text{JMA}} \\
&= \mathcal{C}(\overline{\mathbf{w}}_m) - \mathcal{C}(\widehat{\mathbf{w}}_{\text{JMA}}) + \widehat{w}_{\text{JMA},m}^2(\Psi_{MM} + \Psi_{mm} - \Psi_{Mm} - \Psi_{mM}) \\
&\quad +2\widehat{w}_{\text{JMA},m}\sum_{j=1}^{M}\widehat{w}_{\text{JMA},j}(\Psi_{Mj} - \Psi_{mj}).
\end{aligned}
$$

So similar to (A.3), we know that when $\widehat{w}_{\text{JMA},m} \neq 0$,

$$
\begin{aligned}
\widehat{w}_{\text{JMA},m} &\leqslant (a_m - a_M)^{-1}\Big(2\widehat{\sigma}^2(K_M - K_m) + \widehat{w}_{\text{JMA},m}(\Psi_{MM} + \Psi_{mm} - \Psi_{Mm} - \Psi_{mM}) \\
&\quad +2\sum_{j=1}^{M}\widehat{w}_{\text{JMA},j}(\Psi_{Mj} - \Psi_{mj})\Big).
\end{aligned} \tag{A.13}
$$

Let $\mathcal{S}(\mathbf{A})$ be the largest singular value of a matrix $\mathbf{A}$. We know that any two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$,

$$
\mathcal{S}(\mathbf{AB}) \leqslant \mathcal{S}(\mathbf{A})\mathcal{S}(\mathbf{B}) \quad \textit{and} \quad \mathcal{S}(\mathbf{A} + \mathbf{B}) \leqslant \mathcal{S}(\mathbf{A}) + \mathcal{S}(\mathbf{B}),
$$

which, along with Conditions (C.1)-(C.3), implies that for any $m, j \in \{1, \ldots, M\}$,

$$
\begin{aligned}
&(\mathbf{e} + \mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c})'(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)(\mathbf{e} + \mathbf{X}_{j^c}\boldsymbol{\beta}_{j^c}) \\
&\leqslant \|\mathbf{e} + \mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c}\|\|\mathbf{e} + \mathbf{X}_{j^c}\boldsymbol{\beta}_{j^c}\|\mathcal{S}\{(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)\} \\
&\leqslant \|\mathbf{e} + \mathbf{X}_{m^c}\boldsymbol{\beta}_{m^c}\|\|\mathbf{e} + \mathbf{X}_{j^c}\boldsymbol{\beta}_{j^c}\|\{2\bar{h}_n + (\bar{h}_n)^2\} \\
&= o_p(n^{1/2}),
\end{aligned} \tag{A.14}
$$

From (A.6), (A.12) and (A.14), we know that $\Psi_{mj} = o_p(n^{1/2})$ for any $m, j \in \{1, \ldots, M\}$, which, along with (A.6), (A.8) and (A.13), implies (14). ■

**Proof of Theorem 4.** From (A.12), we need to focus on $\mathbf{\Psi}$. It is seen that for any $m \in \{1, \ldots, M\}$,

$$\begin{aligned}
\mathbf{e}'\mathrm{diag}(P_{11}^m, \ldots, P_{nn}^m)\mathbf{e} &= \sum_{i=1}^{n} e_i^2 \mathbf{x}'_{m,i}(\mathbf{X}'_m\mathbf{X}_m)^{-1}\mathbf{x}_{m,i} \\
&= \mathrm{tr}\Big((n^{-1}\mathbf{X}'_m\mathbf{X}_m)^{-1}n^{-1}\sum_{i=1}^{n} e_i^2\mathbf{x}_{m,i}\mathbf{x}'_{m,i}\Big) \\
&= \mathrm{tr}\left((\mathbf{\Pi}_m\mathbf{Q}_n\mathbf{\Pi}'_m)^{-1}\mathbf{\Pi}_m\mathbf{\Omega}_n\mathbf{\Pi}'_m\right).
\end{aligned}$$

From Condition (C.3), similar to (A.14), we have that for any $m \in \{1, \ldots, M\}$,

$$\begin{aligned}
\mathbf{e}'\mathbf{P}_m(\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)\mathbf{e} &\leqslant \|\mathbf{P}_m\mathbf{e}\|\mathcal{S}\left((\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)\right)\|\mathbf{e}\| \\
&\leqslant \|\mathbf{P}_m\mathbf{e}\|\bar{h}_n\|\mathbf{e}\| \\
&= o_p(1).
\end{aligned}$$

Similarly,

$$\mathbf{e}'\mathbf{P}_m(\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)\mathbf{P}_j\mathbf{e} = o_p(1).$$

From the above results, we have, for $m, j \notin \{1, \ldots, M_0\}$,

$$\begin{aligned}
\Psi_{mj} &= \mathbf{e}'(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{C}_m + \mathbf{C}_j + \mathbf{C}_m\mathbf{C}_j)(\mathbf{I}_n - \mathbf{P}_j)\mathbf{e} - 2K_m\sigma^2 \\
&= \mathrm{tr}\left((\mathbf{\Pi}_m\mathbf{Q}_n\mathbf{\Pi}'_m)^{-1}\mathbf{\Pi}_m\mathbf{\Omega}_n\mathbf{\Pi}'_m\right) + \mathrm{tr}\left(\mathbf{\Pi}_j\mathbf{Q}_n\mathbf{\Pi}'_j)^{-1}\mathbf{\Pi}_j\mathbf{\Omega}_n\mathbf{\Pi}'_j\right) + \widehat{r}_{mj} - 2K_m\widehat{\sigma}^2,
\end{aligned}$$

where $\widehat{r}_{mj} = o_p(1)$. Now, by Condition (C.4) and arguments similar to the proof of (12), we can obtain (15). ■

**Proof of Theorem 5.** Let $m$ be an index belonging to $\{M_0 + 2, \ldots, M\}$. Define

$$\breve{\mathbf{w}}_m = (\widetilde{w}_{\mathrm{JMA},1}, \ldots, \widetilde{w}_{\mathrm{JMA},M_0+1} + \widetilde{w}_{\mathrm{JMA},m}, \widetilde{w}_{\mathrm{JMA},M_0+2}, \ldots, \widetilde{w}_{\mathrm{JMA},m-1}, 0, \widetilde{w}_{\mathrm{JMA},m+1}, \ldots, \widetilde{w}_{\mathrm{JMA},M})'.$$

Using the derivation steps in (A.2), we have $\breve{\mathbf{w}}'_m\mathbf{\Phi}\breve{\mathbf{w}}_m - \widetilde{\mathbf{w}}'_{\mathrm{JMA}}\mathbf{\Phi}\widetilde{\mathbf{w}}_{\mathrm{JMA}} = O_p(1)$. Then, from (A.5), (A.6), (A.12) and (A.14), we know that

$$\begin{aligned}
&\widetilde{\mathcal{J}}(\breve{\mathbf{w}}_m) - \widetilde{\mathcal{J}}(\widetilde{\mathbf{w}}_{\mathrm{JMA}}) \\
&= \breve{\mathbf{w}}'_m\mathbf{\Phi}\breve{\mathbf{w}}_m - \widetilde{\mathbf{w}}'_{\mathrm{JMA}}\mathbf{\Phi}\widetilde{\mathbf{w}}_{\mathrm{JMA}} + (\phi_n - 2\widehat{\sigma}^2)\widetilde{w}_{\mathrm{JMA},m}(K_{M_0+1} - K_m) + O_p(1) \\
&= O_p(1) + (\phi_n - 2\widehat{\sigma}^2)\widetilde{w}_{\mathrm{JMA},m}(K_{M_0+1} - K_m) + O_p(1). \qquad \text{(A.15)}
\end{aligned}$$

Since $\widetilde{\mathbf{w}}_{\mathrm{JMA}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \widetilde{\mathcal{J}}(\mathbf{w})$, we have $\widetilde{\mathcal{J}}(\widetilde{\mathbf{w}}_{\mathrm{JMA}}) \leqslant \widetilde{\mathcal{J}}(\breve{\mathbf{w}}_m)$, which along with (A.15) and $K_{M_0+1} - K_m < 0$ implies that

$$(\phi_n - 2\widehat{\sigma}^2)\widetilde{w}_{\mathrm{JMA},m}(K_m - K_{M_0+1}) = O_p(1). \qquad \text{(A.16)}$$

The above result together with (A.6) and $\phi_n \to \infty$ implies that $\widetilde{w}_{\mathrm{JMA},m} = O_p(\phi_n^{-1})$, which is (20).

From the proofs of Theorem 1 and Theorem 3 and Conditions (C.1)-(C.3), it is straightforward to obtain that for any $m \in \{1, \ldots, M_0\}$, we have

$$\widetilde{w}_{\mathrm{JMA},m} = O_p(\phi_n/n) + o_p(n^{-1/2}). \tag{A.17}$$

Now, by (20), (A.10), (A.17), Conditions (C.1)-(C.2) and (C.4), and $\phi_n n^{-1/2} \to 0$, we have (21). ■