Econometric Theory, **39**, 2023, 412–441.
doi:10.1017/S0266466622000032

AN ASYMPTOTIC THEORY FOR LEAST SQUARES MODEL AVERAGING WITH NESTED MODELS

FANG FANG
East China Normal University

CHAOXIA YUAN
East China Normal University

WENLING TIAN
East China Normal University

Theoretical results of frequentist model averaging mainly focus on asymptotic optimality and asymptotic distribution of the model averaging estimator. However, even for basic least squares model averaging, many theoretical problems have not been well addressed yet. This article discusses asymptotic properties of a class of least squares model averaging methods with nested candidate models that includes the Mallows model averaging (MMA) of Hansen (2007, *Econometrica* 75, 1175–1189) as a special case. Two scenarios are considered: (i) all candidate models are under-fitted; and (ii) the true model is included in the candidate models. We find that in the first scenario, the least squares model averaging method asymptotically assigns weight one to the largest candidate model and the resulting model averaging estimator is asymptotically normal. In the second scenario with a slightly special weight space, if the penalty factor in the weight selection criterion is diverging with certain order, the model averaging estimator is asymptotically optimal by putting weight one to the true model. However, MMA with fixed model dimensions is not asymptotically optimal since it puts nonnegligible weights to over-fitted models. The theoretical results are clearly summarized with their restrictions, and some critical implications are discussed. Monte Carlo simulations confirm our theoretical results.

1. INTRODUCTION

Model averaging has attracted much attention from researchers in the past two decades as it has become a powerful forecasting tool in areas such as econometrics, social sciences, and medical studies. Without putting all our inferential eggs in

We would like to thank the Editor (Peter C.B. Phillips), the Co-Editor (Michael Jansson), and the anonymous referees for many constructive comments and suggestions that led to a much improved paper. Fang gratefully acknowledges the research support from National Key R&D Program of China (2021YFA1000100 and 2021YFA1000101) and the National Natural Science Foundation of China (12071143, 11831008, 11771146). Address correspondence to Fang Fang, Key Laboratory for Advanced Theory and Application in Statistics and Data Science—MOE, Faculty of Economics and Management, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, China; e-mail: ffang@sfs.ecnu.edu.cn.

one unevenly woven basket (Longford, 2005), model averaging combines results from different candidate models with effective weights to achieve better prediction results than a singly selected model. The model weights are selected either from a Bayesian perspective (see Hoeting et al., 1999 for a literature review) or from a frequentist perspective (Buckland, Burnham, and Augustin, 1997; Yang, 2001, 2003; Hjort and Claeskens, 2003a; Leung and Barron, 2006; Hansen, 2007; Hansen and Racine, 2012, among many others). In this article, we focus only on frequentist model averaging.

A substantial amount of frequentist model averaging methods have been proposed over the years. In the framework of parametric models, we have, for example, Mallows model averaging (MMA; Hansen, 2007), optimal mean squared error averaging (Liang et al., 2011), jackknife model averaging (JMA; Hansen and Racine, 2012), heteroskedasticity-robust C_p (Liu and Okui, 2013), optimal model averaging for linear mixed-effects models (Zhang, Zou, and Liang, 2014), multinomial and ordered logit models (Wan, Zhang, and Wang, 2014), Kullback–Leibler model averaging (Zhang, Zou, and Carroll, 2015), optimal model averaging for generalized linear models and generalized linear mixed-effects models (Zhang et al., 2016), model averaging for covariance matrix estimation (Zheng et al., 2017), and model averaging for high-dimensional data (Ando and Li, 2014, 2017; Zhang et al., 2020). There are also many semiparametric model averaging methods (see, for example, Zhang and Liang, 2011; Li, Linton, and Lu, 2015; Kitagawa and Muris, 2016; Li et al., 2018a, 2018b; Zhang and Wang, 2019; Zhu et al., 2019; Fang, Li, and Xia, 2020).

Theoretical results of frequentist model averaging mainly focus on asymptotic optimality and asymptotic distribution of the model averaging estimator. Hansen (2007) showed the asymptotic optimality of MMA, which means that the quadratic error obtained by MMA is asymptotically equivalent to the error of infeasible optimal weight vector. Most of the aforementioned model averaging methods prove the asymptotic optimality following Hansen (2007). However, on the other hand, theoretical results for the asymptotic distribution of the model averaging estimator are very limited in the literature. These asymptotic properties are not only critical to the inference after model averaging (Zhang and Liu, 2019), but also build a foundation for the study of choosing appropriate model averaging procedures. Some work has been done (Hjort and Claeskens, 2003a; Zhang and Liang, 2011; Hansen, 2014; Liu, 2015) in a local asymptotic framework where the regression coefficients are in a local $n^{-1/2}$ neighborhood of zero and n is the sample size. But there has been a discussion about the realism of the local asymptotic framework (Raftery and Zheng, 2003; Hjort and Claeskens, 2003b). Under the fixed parameter framework, the existing results most focus on model averaging for linear models. Two different scenarios are usually discussed in the literature: (i) all candidate models are under-fitted; and (ii) the true model is included in the candidate models.

In the first scenario in which all candidate models are under-fitted, Hansen (2007), Wan, Zhang, and Zou (2010), Hansen and Racine (2012), and Zhang (2021)

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

showed the asymptotic optimality of MMA and JMA. Zhang et al. (2020) considered a weight selection criterion defined in (2) of Section 2 and showed the asymptotic optimality when the penalty factor $\phi_n = \log(n)$. Fang and Liu (2020) discussed the limit of selected weight with nonnested candidate models. But the result is only applicable to a discrete weight set and fixed model dimensions. Liao et al. (2019) studied the convergence of the selected weight of a model averaging method for vector autoregressive models, and it includes MMA as a special case. Other than these, no asymptotic results for the limiting distribution of the model averaging estimator were presented in this scenario. Peng and Yang (2021) discussed a key question concerning when model averaging can outperform model selection. In the second scenario in which the true model is included in the candidate models, Zhang (2015) showed that MMA and JMA estimators are $\sqrt{n}$-consistent. Zhang and Liu (2019) further showed that both MMA and JMA estimators asymptotically assign zero weight to the under-fitted models, and MMA and JMA weights of true and over-fitted models are asymptotically random. Hence, the asymptotic distributions of MMA and JMA estimators are nonstandard. Zhang et al. (2020) showed that the model averaging method asymptotically assigns weight one to the true model when $\phi_n = \log(n)$. So the asymptotic distribution of the model averaging estimator is the same as the least-squares estimator of the true model. However, theoretical results of asymptotic optimality in this scenario were not available. The above results hold with more or less restrictions due to technical difficulties. We summarize some key results with their restrictions clearly marked in Table 1. Note that $\phi_n = 2$ means MMA.

Obviously, many theoretical problems of least squares model averaging have not been well addressed yet even for nested candidate models. Several critical questions still need to be answered:

(Q1) What is the asymptotic behavior of the selected weight and asymptotic distribution of the model averaging estimator in Scenario 1?
(Q2) How is the asymptotic optimality in Scenario 1 achieved?
(Q3) Is the least squares model averaging asymptotically optimal in Scenario 2? If yes, how is the optimality achieved? If no, what is the reason for it?
(Q4) Which ϕ_n is better, 2 or $\log(n)$?

The aim of this paper is to answer these questions to a certain extent and fill in the theoretical gap in the literature. In Section 2, we introduce a class of least squares model averaging methods which includes MMA as a special case. The main theoretical results are presented in Section 3. In the first scenario in which all candidate models are under-fitted, we show that the least squares model averaging method asymptotically assigns weight one to the largest candidate model and the resulting model averaging estimator is asymptotically normal. In the second scenario in which the true model is included in the candidate models, we show that if the penalty factor ϕ_n in the weight selection criterion (2) is diverging with certain order and a slightly special weight space is considered, the model averaging estimator is asymptotically optimal by putting weight one to the true model.

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

TABLE 1. Some key existing theoretical results for least squares model averaging estimator.

Scenario 1: All candidate models are under-fitted			
ϕ_n	Asy. optimality	$\widehat{\mathbf{w}}$	$\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$
2	Yes	$\|\widehat{\mathbf{w}}-\mathbf{w}^0\| \to_p 0$	
	Hansen (2007)*,†	Liao et al. (2019)*	
	Wan, Zhang, and Zou	$\widehat{w}_{M_n} \to_p 1$	
	(2010) Zhang (2021)	Fang and Liu (2020)$^{\dagger,\circ}$	
log(n)	Yes		
	Zhang et al. (2020)*		
Scenario 2: The true model is included in the candidate models			
	Asy. optimality	$\widehat{\mathbf{w}}$	$\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$
2		$\widehat{w}_u = O_p(1/n)$	Nonstandard
		$\widehat{w}_o \nrightarrow_p 0$	distribution$^{\star,\circ}$
		Zhang and Liu (2019)$^{\star,\circ}$	
log(n)		$\widehat{w}_u = O_p(\log(n)k_{M_0+1}/n)$	Asymptotically
		$P(\sum \widehat{w}_o = 0) \to 1$	normal*
		Zhang et al. (2020)*	

Notes: ϕ_n is given in (2); $\widehat{\mathbf{w}}$: the selected weight; $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$: the model averaging estimator; $\mathbf{w}^0$: the optimal weight minimizing the expected quadratic error; $\widehat{w}_{M_n}$: selected weight for the largest candidate model; $\widehat{w}_u$: selected weight for an under-fitted model; $\widehat{w}_o$: selected weight for an over-fitted model; k_{M_0+1}: model dimension for the true model. $\star$: nested candidate models; ‡: discrete weight set $\mathcal{H}_n(N)$; $\circ$: fixed model dimensions.

However, MMA with fixed model dimensions is not asymptotically optimal. We also provide some insights for the asymptotic distribution of the selected weight of MMA in this scenario. Section 4 examines the theoretical results by Monte Carlo simulations. Section 5 concludes the paper with some remarks. All the proofs are included in Appendixes A–G.

2. LEAST SQUARES MODEL AVERAGING

For $i = 1, \ldots, n$, let y_i be a scalar response and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip_n})'$ be a $p_n \times 1$ vector of covariates associated with y_i. We consider a homoskedastic linear regression model

$$y_i = \mu_i + e_i = \sum_{j=1}^{p_n} \beta_j x_{ij} + e_i, \tag{1}$$

where the random errors e_i are independent and identically distributed with mean 0 and variance σ^2 and β_j is an unknown regression coefficient for the jth covariate. Note that $p_n < n$, but it can be diverging, that is, it can increase when the sample size n increases.

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

We assume that only the first $K_n \le p_n$ covariates are available for model fitting. Consider a sequence of nested candidate models $\mathcal{M}_m$, $m = 1, \ldots, M_n$, where the mth model uses the first k_m elements of $\mathbf{x}_i$ and $0 < k_1 < \cdots < k_{M_n} = K_n$. Both K_n and M_n can be diverging. Denote $\mathcal{A}_m = \{1, \ldots, k_m\}$ and $\mathcal{A} = \{1, \ldots, K_n\}$.

For the mth model, let $\mathbf{X}_m$ be the $n \times k_m$ design matrix with ijth element x_{ij} and $\boldsymbol{\beta}_m = (\beta_1, \ldots, \beta_{k_m})'$ be the regression coefficient vector. The least-squares estimator is $\widehat{\boldsymbol{\beta}}_m = (\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{y}$, where $\mathbf{y} = (y_1, \ldots, y_n)'$ is the vector of response. Then the estimator of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ is $\widehat{\boldsymbol{\mu}}_m = \mathbf{X}_m\widehat{\boldsymbol{\beta}}_m = \mathbf{X}_m(\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{y} = \mathbf{P}_m\mathbf{y}$. The sum of squared error is $a_m = \|\mathbf{y} - \widehat{\boldsymbol{\mu}}_m\|^2 = \mathbf{y}'(\mathbf{I}_n - \mathbf{P}_m)\mathbf{y}$, where $\mathbf{I}_n$ is an $n \times n$ identity matrix.

Let w_m be the weight corresponding to the mth candidate model and $\mathbf{w} = (w_1, \ldots, w_{M_n})'$ be a weight vector belonging to the weight set $\mathcal{H}_n = \{\mathbf{w} \in [0,1]^{M_n} : \sum_{m=1}^{M_n} w_m = 1\}$. The least squares averaging estimator of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{K_n})'$ is

$$\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{m=1}^{M_n} w_m \begin{pmatrix} \widehat{\boldsymbol{\beta}}_m \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is a $(K_n - k_m) \times 1$ vector of 0. Denote $\mathbf{X} = \mathbf{X}_{M_n}$. The averaging estimator of $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{w}) = \sum_{m=1}^{M_n} w_m\widehat{\boldsymbol{\mu}}_m = \sum_{m=1}^{M} w_m\mathbf{P}_m\mathbf{y} = \mathbf{P}(\mathbf{w})\mathbf{y}$. The model weight $\mathbf{w}$ is selected by minimizing the criterion

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

where $\widehat{\sigma}^2 = a_{M_n}/(n - K_n)$ is an estimator of σ^2, $\mathbf{K} = (k_1, \ldots, k_{M_n})'$, and the penalty factor ϕ_n is a positive number which may depend on n. Note that this criterion was introduced by Zhang et al. (2020). When $\phi_n = 2$, (2) is the Mallows criterion of Hansen (2007). Although the theoretical results will be derived for a general ϕ_n, we are mainly interested in $\phi_n = 2$ and $\phi_n = \log(n)$. Note that the weight selection criterion (2) is similar to the generalized information criterion (GIC$_{\phi_n}$; Shao, 1997) in the area of model selection. Denote

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathcal{H}_n} \mathcal{G}_n(\mathbf{w}) \tag{3}$$

as the selected weight. Then $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ is the model averaging estimator that is of interest.

For each weight vector $\mathbf{w}$, we define the quadratic error as $L_n(\mathbf{w}) = \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w})\|^2$. Asymptotic optimality of the least squares model averaging is defined as

$$\frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w} \in \mathcal{H}_n} L_n(\mathbf{w})} \to_p 1.$$

Our target is to understand the asymptotic optimality and asymptotic distribution of $\widehat{\mathbf{w}}$ and $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ in two scenarios:

Scenario 1: At least one β_j is nonzero for $j > K_n$. So all the candidate models are under-fitted.

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

Scenario 2: There exists an M_0 such that $(\beta_j : j \in \mathcal{A}_{M_0+1} \setminus \mathcal{A}_{M_0})' \neq 0$ and $\beta_j = 0$ for all the $j > k_{M_0+1}$. So the first M_0 candidate models are under-fitted, the (M_0+1)th model is defined as the "true" model, and all the other models are over-fitted. Assume that there exists at least one over-fitted model. Note that M_0 can be diverging.

3. THEORETICAL RESULTS

The full $n \times p_n$ design matrix $\mathbf{X}_p$ is considered to be nonrandom and full rank. The results in this article are also valid in the almost sure sense when the $\mathbf{X}_p$ is random, provided that the required conditions involving $\mathbf{X}_p$ hold almost surely. All the limiting processes are with respect to $n \to \infty$. Let $\mathbf{X}_c$ be the $n \times (p_n - K_n)$ design matrix for the $p_n - K_n$ covariates that are not included in the candidate models so that $\mathbf{X}_p = (\mathbf{X}, \mathbf{X}_c)$ and $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_c\boldsymbol{\beta}_c$, where $\boldsymbol{\beta}_c = (\beta_{K_n+1}, \ldots, \beta_{p_n})'$.

3.1. Asymptotic Distribution with Under-Fitted Candidate Models

In this subsection, we explore the asymptotic behavior of $\widehat{\mathbf{w}}$ and asymptotic distribution of $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$ in Scenario 1 in which all the candidate models are under-fitted. Denote $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_{K_n}^*)' = E(\widehat{\boldsymbol{\beta}}_{M_n}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_c\boldsymbol{\beta}_c$. We assume the following conditions.

Condition (C1): $P(d_1 \leq \widehat{\sigma}^2/\sigma^2 \leq d_2) \to 1$ where d_1 and d_2 are two positive constants.

Condition (C2): $\lambda_{min}(n^{-1}\mathbf{X}'\mathbf{X}) \geq \kappa_1$ for a positive constant κ_1, where λ_{min} denotes the minimum eigenvalue.

Condition (C3): $\sum_{j \in \mathcal{A} \setminus \mathcal{A}_{M_n-1}} \beta_j^{*2} \geq c_\tau n^{-\tau}$ for positive constants c_τ and τ.

Conditions (C1) and (C2) are commonly used for theoretical results with diverging number of parameters (Zou and Zhang, 2009; Zhang et al., 2020). Note that Condition (C1) does not require that $\widehat{\sigma}^2$ is consistent and thus is easily satisfied. Condition (C3) indicates that the largest model and the second-largest model are separable, but the difference could converge to 0.

Theorem 1. *In Scenario 1, if Conditions (C1)–(C3) are satisfied and* $K_n/n^{1-\tau} \to 0$*, then, for any* $m < M_n$*, we have*

$$\widehat{w}_m = O_p\left(\frac{\phi_n K_n}{n^{1-\tau}}\right).$$

So if $\phi_n K_n M_n/n^{1-\tau} \to 0$*, we have* $\widehat{w}_{M_n} \to_p 1$.

Theorem 1 shows that $\widehat{\mathbf{w}}$ asymptotically puts all the weights to the largest model when all the nested candidate models are under-fitted. Based on this result, we can develop the asymptotical distribution of $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$, which intuitively should be the same

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

as the asymptotical distribution of $\widehat{\boldsymbol{\beta}}_{M_n}$. One more condition from Zou and Zhang (2009) is needed to establish the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{M_n}$ and $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$.

Condition (C4): (i) There exists a positive constant κ_2 such that $\lambda_{max}(n^{-1}\mathbf{X}'\mathbf{X}) \leq \kappa_2$, where λ_{max} denotes the maximum eigenvalue; (ii) $n^{-1}\max_{1\leq i\leq n}\sum_{j=1}^{K_n} x_{ij}^2 \to 0$; (iii) $E(|e_i|^{2+\varsigma}) < \infty$ for a positive constant ς; and (iv) $\log(K_n)/\log(n) \to \nu$ for a constant $\nu \in [0, 1)$.

Theorem 2. *In Scenario 1, if Conditions (C1)–(C4) are satisfied and* $\phi_n K_n M_n / n^{\frac{1}{2}-\tau} \to 0$, *then*

$$\boldsymbol{\alpha}'(\boldsymbol{X}'\boldsymbol{X})^{\frac{1}{2}}\left\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{w}}) - \boldsymbol{\beta}^*\right\} \to_d N(0, \sigma^2),$$

where $\boldsymbol{\alpha}$ *is a vector of norm 1.*

In Scenario 1, Hansen (2007), Wan, Zhang, and Zou (2010), Zhang et al. (2020), and Zhang (2021) showed the asymptotic optimality of $\widehat{\mathbf{w}}$ when $\phi_n = 2$ and $\log(n)$. To understand how the asymptotic optimality is achieved, we need to answer one question: Is it possible that a weighted combination of the candidate models has an even smaller quadratic error than the largest model? Denote $\mathbf{w}_m^0 = (0, \ldots, 1, \ldots, 0)'$ as a weight vector putting all weights on the mth model, and let

$$\mathbf{w}^* = (w_1^*, \ldots, w_{M_n}^*)' = \arg\min_{\mathbf{w}\in\mathcal{H}_n} L_n(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathcal{H}_n} \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w})\|^2$$

be the infeasible optimal weight. The following theorem tells us that the answer is asymptotically "No" in two perspectives. One more commonly used condition is needed.

Condition (C5): $\|\boldsymbol{\mu}\|^2 = O(n)$.

Theorem 3. *In Scenario 1, if Conditions (C2), (C3), and (C5) are satisfied and* $K_n/n^{1-2\tau} \to 0$, *then we have:*

(1) *For any non-random weight vector* $\boldsymbol{w} \neq \boldsymbol{w}_{M_n}^0$, $P(L_n(\boldsymbol{w}) > L_n(\boldsymbol{w}_{M_n}^0)) \to 1$.

(2) $\|\boldsymbol{w}^* - \boldsymbol{w}_{M_n}^0\| \to_p 0$, *i.e.,* $w_{M_n}^* \to_p 1$.

By Theorems 1 and 3, we are able to show the asymptotic optimality of $\widehat{\mathbf{w}}$ in Corollary 1 below in a different way from the literature. Two more regular conditions are needed.

Condition (C6): $\lambda_{min}\{n^{-1}\mathbf{X}_c'(\mathbf{I}_n - \mathbf{P}_{M_n})\mathbf{X}_c\} \geq \kappa_3$ for a positive constant κ_3.

Condition (C7): $\|\boldsymbol{\beta}_c\|^2 \geq d_\tau n^{-\tau}$ for a positive constant d_τ and τ from Condition (C3).

Corollary 1. *In Scenario 1, assume that Conditions (C1)–(C3) and (C5) are satisfied.*

(1) *If* $\phi_n K_n M_n / n^{1-\tau} \to 0$ *and* $K_n/n^{1-2\tau} \to 0$, *then* $\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\| \to_p 0$.

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

(2) *If Conditions (C6) and (C7) are also satisfied,* $\phi_n K_n M_n / n^{1-2\tau} \to 0$, *and* $K_n / n^{1-6\tau} \to 0$, *then* $1 - \widehat{w}_{M_n} = o_p(n^{-\tau})$, $1 - w^*_{M_n} = o_p(n^{-\tau})$, *and*

$$\frac{L_n(\widehat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{H}_n} L_n(\boldsymbol{w})} = \frac{L_n(\widehat{\boldsymbol{w}})}{L_n(\boldsymbol{w}^*)} \to_p 1,$$

i.e., $\widehat{\boldsymbol{w}}$ *is asymptotically optimal in Scenario 1.*

When all the nested candidate models are under-fitted, Theorems 1 and 3 and Corollary 1 reveal that both model averaging procedures with $\phi_n = 2$ or $\log(n)$ achieve asymptotic optimality by pushing all the weights to the largest candidate model, which has the minimum quadratic error among all the weighted candidate models. This conclusion holds even when both the difference between the largest two models and the difference between the largest model and the true model converge to 0 with a negative exponential order as long as the number of candidate models does not diverge too fast.

3.2. Asymptotic Optimality When True Model Is Included

In this subsection, we explore the asymptotic optimality of $\widehat{\mathbf{w}}$ in Scenario 2 in which the true model is included in the candidate models. Note that in this scenario, $\boldsymbol{\beta}_c = 0$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta}$. We need to modify Conditions (C2) and (C3) to the following ones, and one more Condition (C8) is assumed for Scenario 2.

Condition (C2'): $\lambda_{min}(n^{-1}\mathbf{X}'_{M_0+1}\mathbf{X}_{M_0+1}) \geq \kappa_0$ for a positive constant κ_0.
Condition (C3'): $\sum_{j\in\mathcal{A}_{M_0+1}\setminus\mathcal{A}_{M_0}} \beta_j^2 \geq c_{\tau_0} n^{-\tau_0}$ for positive constants c_{τ_0} and τ_0.
Condition (C8): $a_{M_0+1} > a_m$ for any $m > M_0 + 1$ in Scenario 2.

Since $a_{M_0+1} \geq a_m$, for $m > M_0 + 1$, Condition (C8) simply requires $a_{M_0+1} \neq a_m$, which is to make sure that the over-fitted models are distinguishable from the true model in terms of sum of squared error. It is satisfied if the projection of $\mathbf{y}$ on the column space of $\mathbf{X}_m$ is not the same as the projection on the column space of $\mathbf{X}_{M_0+1}$, which is usually true.

Theorem 4. *In Scenario 2, if Conditions (C2'), (C3'), (C5), and (C8) are satisfied and* $K_n / n^{1-2\tau_0} \to 0$, *then for any nonrandom weight vector* $\boldsymbol{w} \neq \boldsymbol{w}^0_{M_0+1}$, *we have*

$$P(L_n(\boldsymbol{w}) > L_n(\boldsymbol{w}^0_{M_0+1})) \to 1.$$

Theorem 4 tells us that the true model has the minimum quadratic error among all the weighted candidate models in Scenario 2. In this scenario, Zhang et al. (2020) showed that the weight of the true model will go to 1 in probability when $\phi_n = \log(n)$. Then intuitively $\widehat{\mathbf{w}}$ should be asymptotically optimal. On the other hand, when $\phi_n = 2$, Zhang and Liu (2019) showed that the weights of over-fitted models do not go to 0 when the model dimensions are fixed. Hence, the $\widehat{\mathbf{w}}$ should

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

not be asymptotically optimal. The following theorem rigorously proves the results if we restrict the weight space to

$$\mathcal{H}_n^{\delta} = \left\{ \mathbf{w} \in \mathcal{H}_n : \sum_{m<M_0+1} w_m = 0 \text{ or } \sum_{m<M_0+1} w_m \geq \delta n^{-\tau_0} \right\},$$

where δ is a positive constant which can be arbitrarily small and τ_0 is from Condition (C3'). Note that $\mathcal{H}_n^{\delta}$ contains the discrete weight space $\mathcal{H}_n(N)$ considered by Hansen (2007) as long as δ is small enough or n is large and it goes to $\mathcal{H}_n$ as n goes to infinity.

For notation convenience, we still use $\widehat{\mathbf{w}}$ and $\mathbf{w}^*$ to denote $\arg\min_{\mathbf{w}\in\mathcal{H}_n^{\delta}} \mathcal{G}_n(\mathbf{w})$ and $\arg\min_{\mathbf{w}\in\mathcal{H}_n^{\delta}} L_n(\mathbf{w})$, respectively.

Lemma 1. *In Scenario 2, if Conditions (C2'), (C3'), (C5), and (C8) are satisfied and* $K_n/n^{1-6\tau_0} \to 0$, *we have* $P(\boldsymbol{w}^* = \boldsymbol{w}^0_{M_0+1}) \to 1$.

Theorem 5. *In Scenario 2, assume that Conditions (C1), (C2'), (C3'), (C5), and (C8) are satisfied and* $K_n/n^{1-6\tau_0} \to 0$. *Then* $\max_{m<M_0+1} \widehat{w}_m = O_p\big(\frac{\phi_n k_{M_0+1}}{n^{1-\tau_0}}\big)$ *and* $\max_{m>M_0+1} \widehat{w}_m = O_p\big(\frac{K_n}{\phi_n}\big)$. *Furthermore, we have:*

(1) If $\phi_n \to \infty$ *and* $\phi_n^2 k_{M_0+1}^3 M_0^2/n^{1-2\tau_0} \to 0$, *then*

$$\frac{L_n(\widehat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{H}_n^{\delta}} L_n(\boldsymbol{w})} = \frac{L_n(\widehat{\boldsymbol{w}})}{L_n(\boldsymbol{w}^*)} \to_p 1.$$

(2) If $\phi_n = 2$ *and* $\phi_n^2 k_{M_0+1}^2 K_n M_0^2/n^{1-2\tau_0} \to 0$, *then as long as* $\sum_{m>M_0+1} \widehat{w}_m^2 (\frac{\boldsymbol{e}'\boldsymbol{P}_m\boldsymbol{e}}{\boldsymbol{e}'\boldsymbol{P}_{M_0+1}\boldsymbol{e}} - 1) \nrightarrow_p 0$, *we have*

$$\frac{L_n(\widehat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{H}_n^{\delta}} L_n(\boldsymbol{w})} = \frac{L_n(\widehat{\boldsymbol{w}})}{L_n(\boldsymbol{w}^*)} \nrightarrow_p 1.$$

When $\phi_n = \log(n)$, Theorem 5 shows that the least squares model averaging is asymptotically optimal by putting all the weights to the true model as long as the model dimensions do not diverge too fast.

When $\phi_n = 2$, the condition $\sum_{m>M_0+1} \widehat{w}_m^2 \big(\frac{\mathbf{e}'\mathbf{P}_m\mathbf{e}}{\mathbf{e}'\mathbf{P}_{M_0+1}\mathbf{e}} - 1\big) \nrightarrow_p 0$ is generally not established. However, in the special case that the model dimensions are fixed, we have $\tau_0 = 0$ and $\max_{m<M_0+1} \widehat{w}_m = O_p\big(\frac{\phi_n k_{M_0+1}}{n^{1-\tau_0}}\big) = O_p(n^{-1})$, which is the conclusion in Theorem 1 of Zhang and Liu (2019). If we assume $n^{-1}\mathbf{X}_p'\mathbf{X}_p \to \mathbf{Q}$ and $n^{-1/2}\mathbf{X}_p'\mathbf{e} \to_d \mathbf{Z} \sim N(0, \boldsymbol{\Omega})$ with positive definite matrices $\mathbf{Q}$ and $\boldsymbol{\Omega}$, then $\frac{\mathbf{e}'\mathbf{P}_m\mathbf{e}}{\mathbf{e}'\mathbf{P}_{M_0+1}\mathbf{e}} \nrightarrow_p 1$ is easy to check, and similar to Theorem 2 of Zhang and Liu (2019), we have $\{\widehat{w}_m, m \geq M_0+1\}$ converge to a nondegenerate random weight vector in distribution. Then $\sum_{m>M_0+1} \widehat{w}_m^2 \big(\frac{\mathbf{e}'\mathbf{P}_m\mathbf{e}}{\mathbf{e}'\mathbf{P}_{M_0+1}\mathbf{e}} - 1\big)$ does not converge to 0 in probability. More discussion is provided below. So Theorem 5 actually shows

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

that MMA ($\phi_n = 2$) is not asymptotically optimal in Scenario 2 with fixed model dimensions because it puts nonnegligible weights to over-fitted models. Note that this conclusion does not contradict the asymptotic optimality of MMA established in Hansen (2007) since a different scenario is considered here.

When $\phi_n = 2$ and the model dimensions are fixed, denote $M = M_n$ and $p = p_n$. Zhang and Liu (2019) showed that

$$(\widehat{w}_{M_0+1}, \ldots, \widehat{w}_M)' \to_d \widetilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}\in\mathcal{L}} \boldsymbol{\lambda}'\boldsymbol{\Gamma}\boldsymbol{\lambda},$$

where $\mathcal{L} = \left\{\boldsymbol{\lambda} \in [0,1]^{M-M_0} : \sum_{m=1}^{M-M_0} \lambda_m = 1\right\}$, $\boldsymbol{\Gamma}$ is an $(M - M_0) \times (M - M_0)$ matrix with the (m,l)th element $\Gamma_{ml} = 2\sigma^2 k_{M_0+m} - \mathbf{Z}'\mathbf{V}_{\max\{m,l\}}\mathbf{Z}$, $\mathbf{V}_m = \Pi'_{M_0+m}(\Pi_{M_0+m}\mathbf{Q}\Pi'_{M_0+m})^{-1}\Pi_{M_0+m}$, and $\Pi_m = (\mathbf{I}_{k_m}, \mathbf{0}_{k_m\times(p-k_m)})$. This is a very nice result, but the $\widetilde{\boldsymbol{\lambda}}$ is hard to understand. Here, we provide a nearly explicit form of $\widetilde{\boldsymbol{\lambda}}$ in a special case of $\mathbf{Q} = \mathbf{I}_p$. Actually, when $\mathbf{Q} = \mathbf{I}_p$, $\mathbf{V}_m$ is a p by p matrix that only the (i,i)th ($i \le k_{M_0+m}$) elements are 1 and the others are 0. Denote $\mathbf{Z} = (Z_1, \ldots, Z_p)'$. Then $\mathbf{Z}'\mathbf{V}_m\mathbf{Z} = Z_1^2 + \cdots + Z_{k_{M_0+m}}^2$. By ignoring the terms without $\boldsymbol{\lambda}$, we have

$$\begin{aligned}
\boldsymbol{\lambda}'\boldsymbol{\Gamma}\boldsymbol{\lambda} &\propto 2\sigma^2 \sum_{m=1}^{M-M_0} k_{M_0+m}\lambda_m + \sum_{m=k_{M_0+1}+1}^{k_{M_0+2}} Z_m^2\lambda_1^2 + \sum_{m=k_{M_0+2}+1}^{k_{M_0+3}} Z_m^2(\lambda_1+\lambda_2)^2 \\
&\quad + \cdots + \sum_{m=k_{M-1}+1}^{k_M} Z_m^2(\lambda_1 + \cdots + \lambda_{M-M_0-1})^2 \\
&\propto \sum_{m=k_{M_0+1}+1}^{k_{M_0+2}} Z_m^2\lambda_1^2 - 2(k_{M_0+2} - k_{M_0+1})\sigma^2\lambda_1 \\
&\quad + \sum_{m=k_{M_0+2}+1}^{k_{M_0+3}} Z_m^2(\lambda_1+\lambda_2)^2 - 2(k_{M_0+3} - k_{M_0+2})\sigma^2(\lambda_1+\lambda_2) \\
&\quad + \cdots + \sum_{m=k_{M-1}+1}^{k_M} Z_m^2(\lambda_1 + \cdots + \lambda_{M-M_0-1})^2 - 2(k_M - k_{M-1})\sigma^2(\lambda_1 + \cdots + \lambda_{M-M_0-1}).
\end{aligned}$$

If we release the constraint that λ_m should be between 0 and 1, the above equality is the sum of $M - M_0 - 1$ functions with free parameters λ_1, $\lambda_1+\lambda_2$, …, $\lambda_1+\cdots+\lambda_{M-M_0+1}$, respectively. Then we have

$$\widetilde{\lambda}_1 = \frac{(k_{M_0+2} - k_{M_0+1})\sigma^2}{\sum_{m=k_{M_0+1}+1}^{k_{M_0+2}} Z_m^2},$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

$$\widetilde{\lambda}_1+\widetilde{\lambda}_2=\frac{(k_{M_0+3}-k_{M_0+2})\sigma^2}{\sum_{m=k_{M_0+2}+1}^{k_{M_0+3}}Z_m^2},$$

$$\vdots$$

$$\widetilde{\lambda}_1+\cdots+\widetilde{\lambda}_{M-M_0-1}=\frac{(k_M-k_{M-1})\sigma^2}{\sum_{m=k_{M-1}+1}^{k_M}Z_m^2},$$

$$\widetilde{\lambda}_1+\cdots+\widetilde{\lambda}_{M-M_0-1}+\widetilde{\lambda}_{M-M_0}=1.$$

If we further restrict λ_m to $[0,1]$, the explicit form of $\widetilde{\boldsymbol{\lambda}}$ is hard to derive. But a clear message is that $P(\widetilde{\lambda}_1<1)>0$ since

$$P\left(\frac{(k_{M_0+2}-k_{M_0+1})\sigma^2}{\sum_{m=k_{M_0+1}+1}^{k_{M_0+2}}Z_m^2}<\frac{(k_{M_0+3}-k_{M_0+2})\sigma^2}{\sum_{m=k_{M_0+2}+1}^{k_{M_0+3}}Z_m^2}<\cdots<\frac{(k_M-k_{M-1})\sigma^2}{\sum_{m=k_{M-1}+1}^{k_M}Z_m^2}<1\right)>0.$$

Note that $\widetilde{\lambda}_1$ is the limit of $\widehat{w}_{M_0+1}$ which is the selected weight of the true model. So it indicates that least squares model averaging puts weight on the over-fitted models with a positive probability when $\phi_n=2$ and the model dimensions are fixed. In a simple case that $M-M_0=2$, i.e., there is one true model and one over-fitted model,

$$\widetilde{\lambda}_1=\min\left\{\frac{(k_{M_0+2}-k_{M_0+1})\sigma^2}{\sum_{m=k_{M_0+1}+1}^{k_{M_0+2}}Z_m^2},1\right\}\text{ and }\widetilde{\lambda}_2=1-\widetilde{\lambda}_1.$$

Specifically, $\widetilde{\lambda}_1$ has a truncated inverse chi-square distribution if $\boldsymbol{\Omega}=\sigma^2\mathbf{I}_p$. Note that $P(\widetilde{\lambda}_2>0)$ is positive.

3.3. Summary of the Theoretical Results

We summarize the main theoretical results in Table 2. The restrictions on the results are clearly marked.

In Scenario 1 when all the candidate models are under-fitted, the least squares model averaging with $\phi_n=2$ and $\log(n)$ both are asymptotically optimal by pushing all the weights to the largest candidate model which asymptotically has the smallest quadratic error among all the weighted models. And as we mentioned, this conclusion holds even when both the difference between the largest two models and the difference between the largest model and the true model converge to 0 with a negative exponential order as long as the number of candidate models does not diverge too fast. The model averaging estimator has the same asymptotical distribution as the least-squares estimator of the largest candidate model, which is normal. This answers Questions (Q1) and (Q2).

In Scenario 2 when the true model is included in the candidate models and a slightly special weight space is considered, the least squares model averaging

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

TABLE 2. Updated key theoretical results for least squares model averaging estimator.

Scenario 1: All candidate models are under-fitted			
ϕ_n	Asy. optimality	$\widehat{\mathbf{w}}$	$\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$
2	Yes Hansen (2007)*,† Wan, Zhang, and Zou (2010) Zhang (2021)	$\|\widehat{\mathbf{w}}-\mathbf{w}^0\| \to_p 0$ Liao et al. (2019)* $\widehat{w}_m = O_p(\frac{\phi_n K_n}{n^{1-\tau}})^{\star}$ $m < M_n$ $\widehat{w}_{M_n} \to_p 1^{\star}$	Asymptotically normal*
$\log(n)$	Yes Zhang et al. (2020)*	$\|\widehat{\mathbf{w}} - \mathbf{w}^*\| \to_p 0^{\star}$	
Scenario 2: The true model is included in the candidate models			
	Asy. optimality	$\widehat{\mathbf{w}}$	$\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$
2	No$^{\star,\circ,\diamond}$	$\widehat{w}_u = O_p(1/n)$ $\widehat{w}_o \not\to_p 0$ Zhang and Liu (2019)$^{\star,\circ}$	Nonstandard distribution$^{\star,\circ}$
$\log(n)$	Yes$^{\star,\diamond}$	$\widehat{w}_u = O_p(\log(n)k_{M_0+1}/n)$ $P(\sum \widehat{w}_o = 0) \to 1$ Zhang et al. (2020)*	Asymptotically normal*

Notes: ϕ_n is given in (2); $\widehat{\mathbf{w}}$: the selected weight; $\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})$: the model averaging estimator; $\mathbf{w}^0$: the optimal weight minimizing the expected quadratic error; $\mathbf{w}^*$: the optimal weight minimizing the quadratic error; $\widehat{w}_{M_n}$: selected weight for the largest candidate model; $\widehat{w}_u$: selected weight for an under-fitted model; $\widehat{w}_o$: selected weight for an over-fitted model; k_{M_0+1}: model dimension for the true model; K_n: model dimension for the largest model; $n^{-\tau}$ is the difference order between the largest candidate model and other models or the truth; $\star$: nested models; †: discrete weight set $\mathcal{H}_n(N)$; $\circ$: fixed model dimensions; $\diamond$: weight set $\mathcal{H}_n^{\delta}$.

methods with $\phi_n = 2$ and $\log(n)$ have different asymptotic properties. When $\phi_n = \log(n)$, the least squares model averaging is asymptotically optimal by pushing all the weights to the true model which asymptotically has the smallest quadratic error among all the weighted models. The model averaging estimator has the same asymptotic distribution as the least-squares estimator of the true model, which is normal. When $\phi_n = 2$ (MMA) and the model dimensions are fixed, asymptotic optimality does not hold since the model averaging method puts weight on the over-fitted models with a positive probability. The asymptotic distribution of the selected weight of the true model is closely related to the truncated inverse chi-square distribution. The model averaging estimator has a nonstandard asymptotic distribution. This partially answers Question (Q3).

The asymptotic behaviors of least squares model averaging with $\phi_n = 2$ and $\phi_n = \log(n)$ are very similar to the behaviors of Akaike information criterion (Akaike, 1973, AIC) and Bayesian information criterion (Schwarz, 1978, BIC;

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

Shao, 1997). In Scenario 1, AIC and BIC both select the largest model with probability going to 1. In Scenario 2, AIC cannot distinguish the true model and the over-fitted models, while BIC has model selection consistency, which means that the probability of selecting the true model goes to 1. We expect that the comparison of the finite-sample performances of $\phi_n = 2$ to $\phi_n = \log(n)$ should also be similar to the comparison of AIC and BIC. In Scenario 1, model averaging with $\phi_n = 2$ usually performs better than model averaging with $\phi_n = \log(n)$. On the other hand, $\phi_n = \log(n)$ is usually preferred to $\phi_n = 2$ in Scenario 2. However, with finite samples, the performances of model averaging methods with $\phi_n = 2$ and $\log(n)$ depend on many factors such as sample size, signal pattern, noise level, an so on. So it is hard to tell which one is better, 2 or $\log(n)$, for a specific set of data. In the area of model selection, we also have a similar dilemma in comparing the finite-sample performances of AIC and BIC (Zhang and Yang, 2015). This partially answers Question (Q4), and more discussion is provided in Section 5.

Remark. As suggested by the Editor, we make some cautionary mention for the use of theoretical results above. These results are based on model (1), and "true model" is regularly mentioned. However, as the aphorism "all models are wrong, but some are useful"[1] indicates, a "true model" may never exist because of the complexity of the data generating process in practice and difficulties of representing it in a probability space framework (Phillips, 2005). A related but nonetheless vital concern for valid inference is the multiplicity problem of data reuse or data snooping. We refer interested readers to Phillips (2005) and Leeb and Pötscher (2005) for an insightful discussion that has deep significance for applications.

4. SIMULATION STUDY

In this section, we conduct simulation studies to confirm the main theoretical results in Section 3.3. For each $i = 1, \ldots, n$, we generate y_i from

$$y_i = \sum_{j=1}^{p_n} \beta_j x_{ij} + e_i,$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip_n})'$ are independent and identically distributed as $N(\mathbf{0}, \Sigma)$ with $\Sigma_{jk} = 0.5^{|j-k|}$, e_i is $N(0,1)$ and independent of $\mathbf{x}_i$, $(\beta_1, \ldots, \beta_{p_n})' = c\boldsymbol{\beta}_0$, and $c = \sqrt{\frac{R^2}{(1-R^2)\boldsymbol{\beta}_0'\Sigma\boldsymbol{\beta}_0}}$, so that $R^2 = \mathrm{Var}\big(\sum_{j=1}^{p_n} \beta_j x_{ij}\big)/\mathrm{Var}(y_i)$ represents the signal strength.

In Scenario 1, $\boldsymbol{\beta}_0 = (1, 1, \ldots, 1)'$ or a random permutation of $(1, 2, \ldots, p_n)'$. In Scenario 2, $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}', 0, \ldots, 0)'$, where $\boldsymbol{\beta}_{01}$ is a $k_{M_0+1} \times 1$ vector that equals $(1, 1, \ldots, 1)'$ or a random permutation of $(1, 2, \ldots, k_{M_0+1})'$.

We consider $n = 100, 1,000, 10,000$, $R^2 = 0.3, 0.6, 0.9$, $p_n = \lceil 2n^{\frac{1}{2}} \rceil$, $K_n = \lceil n^{\frac{1}{2}} \rceil$, $M_n = \lceil n^{\frac{1}{3}} \rceil$, and $k_{M_0+1} = \lceil 2n^{\frac{1}{4}} \rceil$. The mth candidate model includes the first

[1] It is often attributed to Box (1976), but which has deeper historical origins.

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

TABLE 3. Simulation results for Scenario 1.

		Weight of the largest model			Loss ratio to $L_n(\mathbf{w}^*)$		
R^2	n	$\phi_n = 2$	$\phi_n = \log(n)$	$\mathbf{w}^*$	$\phi_n = 2$	$\phi_n = \log(n)$	$L_n(\mathbf{w}^0_{M_n})$
				$\boldsymbol{\beta}_0$ is $(1, 1, \ldots, 1)'$			
0.3	100	0.389	0.096	0.623	1.119	1.333	1.137
0.3	1,000	0.781	0.305	0.807	1.016	1.183	1.023
0.3	10,000	0.937	0.711	0.909	1.002	1.038	1.003
0.6	100	0.676	0.375	0.830	1.049	1.178	1.030
0.6	1,000	0.915	0.708	0.910	1.004	1.042	1.004
0.6	10,000	0.975	0.884	0.958	1.000	1.008	1.000
0.9	100	0.847	0.657	0.936	1.014	1.064	1.004
0.9	1,000	0.958	0.854	0.968	1.001	1.014	1.000
0.9	10,000	0.987	0.941	0.985	1.000	1.002	1.000
				$\boldsymbol{\beta}_0$ is a random permutation of $(1, 2, \ldots, p_n)'$			
0.3	100	0.466	0.142	0.688	1.102	1.337	1.113
0.3	1,000	0.724	0.241	0.796	1.017	1.165	1.024
0.3	10,000	0.891	0.505	0.887	1.002	1.037	1.003
0.6	100	0.753	0.472	0.858	1.036	1.156	1.024
0.6	1,000	0.893	0.632	0.903	1.005	1.038	1.005
0.6	10,000	0.957	0.800	0.947	1.001	1.007	1.001
0.9	100	0.879	0.723	0.947	1.012	1.057	1.003
0.9	1,000	0.948	0.820	0.964	1.001	1.011	1.001
0.9	10,000	0.978	0.900	0.981	1.000	1.002	1.000

$K_n - gap \times (M_n - m)$ covariates, where $gap = [K_n/M_n]$. In Scenario 2, the "true model" is defined as the smallest model that contains the first k_{M_0+1} covariates. For each scenario, there are $3 \times 3 \times 2 = 18$ parameter combinations. The simulation is conducted for 500 replications.

We compare the performances of least squares model averaging methods with $\phi_n = 2$ and $\phi_n = \log(n)$. The focus is on the weight limit and asymptotic optimality. Table 3 reports the averages (more than 500 simulation replications) of $\widehat{w}_{M_n}$, $w^*_{M_n}$, $L_n(\widehat{\mathbf{w}})/L_n(\mathbf{w}^*)$, and $L_n(\mathbf{w}^0_{M_n})/L_n(\mathbf{w}^*)$ for Scenario 1. Table 4 reports the averages of $\widehat{w}_{M_0+1}$, $w^*_{M_0+1}$, $L_n(\widehat{\mathbf{w}})/L_n(\mathbf{w}^*)$, and $L_n(\mathbf{w}^0_{M_0+1})/L_n(\mathbf{w}^*)$ for Scenario 2.

The simulation results can be summarized as follows.

(1) In Scenario 1, whether $\phi_n = 2$ or $\phi_n = \log(n)$, the weight of the largest candidate model keeps increasing as the sample size and approaches to 1 in most cases. In some cases, the weights are still not that large even with $n = 10,000$. For example, $\widehat{w}_{M_n} = 0.505$ when $R^2 = 0.3$, $n = 10,000$, $\phi_n = \log(n)$, and $\boldsymbol{\beta}_0$ is a

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

TABLE 4. Simulation results for Scenario 2.

		Weight of the true model			Loss ratio to $L_n(\mathbf{w}^*)$		
R^2	n	$\phi_n = 2$	$\phi_n = \log(n)$	$\mathbf{w}^*$	$\phi_n = 2$	$\phi_n = \log(n)$	$L_n(\mathbf{w}^0_{M_0+1})$
$\boldsymbol{\beta}_0$ is $(1, 1, \ldots, 1, 0, \ldots, 0)'$							
0.3	100	0.251	0.089	0.525	1.641	2.169	1.643
0.3	1,000	0.498	0.224	0.783	1.366	1.744	1.243
0.3	10,000	0.745	0.967	0.981	1.177	1.061	1.037
0.6	100	0.494	0.318	0.779	1.548	1.911	1.400
0.6	1,000	0.694	0.706	0.898	1.305	1.388	1.180
0.6	10,000	0.751	0.991	0.990	1.174	1.040	1.034
0.9	100	0.791	0.839	0.921	1.383	1.415	1.283
0.9	1,000	0.765	0.951	0.962	1.269	1.158	1.143
0.9	10,000	0.752	0.998	0.996	1.172	1.033	1.032
$\boldsymbol{\beta}_0 = (\boldsymbol{\beta}'_{01}, 0, \ldots, 0)'$, $\boldsymbol{\beta}_{01}$ is a random permutation of $(1, 2, \ldots, k_{M_0+1})'$							
0.3	100	0.301	0.134	0.619	1.680	2.160	1.660
0.3	1,000	0.617	0.535	0.865	1.343	1.540	1.199
0.3	10,000	0.751	0.983	0.987	1.174	1.047	1.035
0.6	100	0.586	0.461	0.821	1.568	1.829	1.439
0.6	1,000	0.715	0.864	0.934	1.304	1.232	1.157
0.6	10,000	0.753	0.995	0.993	1.172	1.036	1.033
0.9	100	0.795	0.882	0.931	1.424	1.373	1.326
0.9	1,000	0.747	0.976	0.975	1.282	1.131	1.131
0.9	10,000	0.754	0.999	0.997	1.170	1.032	1.032

random permutation. But if we increase the sample size to a larger number, the weight will get closer to 1. The optimal $\mathbf{w}^*$ also puts almost all the weights to the largest candidate model as the sample size goes large. The loss ratios all converge to 1 as expected, which confirms the asymptotic optimality of least squares model averaging in Scenario 1.

(2) In Scenario 2, when $\phi_n = 2$, the average weight of the true model stabilizes near 0.75 as the sample size goes large. Actually, even if we keep increasing the sample size, the average weight will not go to 1. Meanwhile, the weights of the true model for $\phi_n = \log(n)$ and the optimal weight $\mathbf{w}^*$ both go to 1. A direct consequence is that the loss ratios for $\phi_n = \log(n)$ and the true model converge to 1, indicating asymptotic optimality. But the loss ratio for $\phi_n = 2$ stabilizes near 1.17, which means that MMA is not asymptotically optimal in this scenario.

(3) For the comparison between $\phi_n = 2$ and $\phi_n = \log(n)$, we can see that in Scenario 1 $\phi_n = 2$ generally performs better than $\phi_n = \log(n)$ in terms of smaller

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

loss ratio although they both have asymptotic optimality. The advantage is more obvious when R^2 is small and the sample size is small. This result is expected since a larger penalty factor forces the weight to smaller models which usually have larger errors. Moreover, as n goes large, $\widehat{w}_{M_n}$ of $\phi_n = 2$ is approaching 1 much faster than $\widehat{w}_{M_n}$ of $\phi_n = \log(n)$. In Scenario 2, $\phi_n = \log(n)$ performs better than $\phi_n = 2$ when the sample size is large ($n = 10,000$) and the power of asymptotic theory is revealed. But when the sample size is small or moderate ($n = 100$ or 1,000), in some situations, $\phi_n = \log(n)$ outperforms $\phi_n = 2$, and in some other situations, $\phi_n = 2$ outperforms $\phi_n = \log(n)$. A typical trend is that $\phi_n = \log(n)$ performs worse than $\phi_n = 2$ when the sample size is relatively small and $\phi_n = \log(n)$ becomes better in comparison with $\phi_n = 2$ when the sample size goes large.

5. CONCLUDING REMARKS

In this article, we study the asymptotic behaviors of a class of least squares model averaging estimators with nested candidate models. The asymptotic results depend on the scenarios considered and the choice of the penalty factor ϕ_n in the weight selection criterion. The model averaging methods with $\phi_n = 2$ and $\phi_n = \log(n)$ have very similar asymptotic behaviors to the traditional model selection methods AIC and BIC, respectively. Hence, in large samples, model averaging and model selection are almost equivalent under the framework considered here, indicating that model averaging may be more valuable with small sample sizes compared to model selection. When the sample size becomes larger, the model selection uncertainty gets lower and model averaging may have no real advantage. This finding is consistent with the results from Yuan and Yang (2005) and Peng and Yang (2021).

A critical problem in practice is how to choose the most appropriate penalty factor ϕ_n in least squares model averaging methods. In Scenario 1 where all the candidate models are under-fitted, $\phi_n = 2$ is usually preferred. In Scenario 2 where the true model is included in the candidate models, $\phi_n = \log(n)$ has better theoretical results. However, with finite samples in practice, the performances of model averaging methods with $\phi_n = 2$ and $\log(n)$ depend on many factors such as sample size, signal pattern, noise level, and so on. What is more complicated is that we do not know which scenario is the truth. To deal with a similar dilemma in model selection, Zhang and Yang (2015) proposed to apply cross validation to choose between AIC and BIC. They showed that an adaptive selection by cross validation between AIC and BIC on a sequence of linear models leads to asymptotically optimal function estimation in both parametric and nonparametric scenarios. We may also apply cross validation to choose between $\phi_n = 2$ and $\phi_n = \log(n)$ in model averaging. The results established in this paper provide a foundation for the theoretical exploration of this cross validation task.

There are several limitations of our results. First, we only consider a homoskedastic error. With heteroskedastic errors, the Mallows-type criterion considered here is not optimal. However, by applying a jackknife criterion (Hansen

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

and Racine, 2012) plus an extra penalty term $\phi_n \mathbf{w}'\mathbf{K}$, we believe that similar results will be obtained. Actually, Zhang and Liu (2019) have already considered such a criterion and showed that it will force all the weights to the true model in Scenario 2 if $\phi_n = \log(n)$. Second, the results are confined to nested models. It is desirable to extend our results to the case of nonnested models, but it is quite challenging. Previous work, such as Hansen (2014) and Zhang and Liu (2019), also mentioned this challenge, but it seems not too many solutions are available so far. Third, our last theoretical result relies on a sightly special weight set due to technical difficulty. Developing asymptotic results without these limitations are important topics for future research.

APPENDIX

Appendix A. Proof of Theorem 1

Let $\mathbf{\Phi}$ be an $M_n \times M_n$ matrix with the ml^{th} element $\Phi_{ml} = a_{\max\{m,l\}}$. For nested candidate models, we have $(\mathbf{I}_n - \mathbf{P}_m)(\mathbf{I}_n - \mathbf{P}_l) = \mathbf{I}_n - \mathbf{P}_{\max\{m,l\}}$. Then

$$\mathcal{G}_n(\mathbf{w}) = \mathbf{w}'\mathbf{\Phi}\mathbf{w} + \phi_n \widehat{\sigma}^2 \mathbf{w}'\mathbf{K}, \qquad \text{for any } \mathbf{w} \in \mathcal{H}_n.$$

For any $m < M_n$, define

$$\widetilde{\mathbf{w}}_m = (\widehat{w}_1, \ldots, \widehat{w}_{m-1}, 0, \widehat{w}_{m+1}, \ldots, \widehat{w}_{M_n-1}, \widehat{w}_{M_n} + \widehat{w}_m)' = \widehat{\mathbf{w}} + (0, \ldots, 0, -\widehat{w}_m, 0, \ldots, 0, \widehat{w}_m)'.$$

Then

$$\begin{aligned}
0 &\le \mathcal{G}_n(\widetilde{\mathbf{w}}_m) - \mathcal{G}_n(\widehat{\mathbf{w}}) \\
&= [(\widetilde{\mathbf{w}}_m - \widehat{\mathbf{w}}) + 2\widehat{\mathbf{w}}]'\mathbf{\Phi}(\widetilde{\mathbf{w}}_m - \widehat{\mathbf{w}}) + \phi_n \widehat{\sigma}^2 (\widetilde{\mathbf{w}}_m - \widehat{\mathbf{w}})'\mathbf{K} \\
&= \widehat{w}_m^2 (a_m - a_{M_n}) + 2\widehat{w}_m \sum_{l=1}^{M_n} \widehat{w}_l (a_{M_n} - a_{\max\{m,l\}}) + \phi_n \widehat{\sigma}^2 \widehat{w}_m (K_n - k_m) \\
&\le \widehat{w}_m^2 (a_m - a_{M_n}) + 2\widehat{w}_m^2 (a_{M_n} - a_m) + \phi_n \widehat{\sigma}^2 \widehat{w}_m (K_n - k_m) \\
&= -\widehat{w}_m^2 (a_m - a_{M_n}) + \phi_n \widehat{\sigma}^2 \widehat{w}_m (K_n - k_m).
\end{aligned}$$

So when $\widehat{w}_m \neq 0$, we have

$$\widehat{w}_m \le (a_m - a_{M_n})^{-1} \phi_n \widehat{\sigma}^2 (K_n - k_m). \tag{A.1}$$

Note $E\{(\widehat{\boldsymbol{\beta}}_{M_n} - \boldsymbol{\beta}^*)'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}_{M_n} - \boldsymbol{\beta}^*)\} = E\{\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\} = \sigma^2 \mathrm{tr}\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} = \sigma^2 K_n$. So $n^{-1}(\widehat{\boldsymbol{\beta}}_{M_n} - \boldsymbol{\beta}^*)'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}_{M_n} - \boldsymbol{\beta}^*) = O_p(K_n/n)$. Then, similar to the proof of Lemma 2 of Zhang et al. (2020), we have

$$\begin{aligned}
&n^{-1}(a_m - a_{M_n}) \\
&\quad = n^{-1}\left\{\begin{pmatrix} \widehat{\boldsymbol{\beta}}_m \\ \mathbf{0} \end{pmatrix} - \widehat{\boldsymbol{\beta}}_{M_n}\right\}'\mathbf{X}'\mathbf{X}\left\{\begin{pmatrix} \widehat{\boldsymbol{\beta}}_m \\ \mathbf{0} \end{pmatrix} - \widehat{\boldsymbol{\beta}}_{M_n}\right\} \\
&\quad = n^{-1}\left\{\begin{pmatrix} \widehat{\boldsymbol{\beta}}_m \\ \mathbf{0} \end{pmatrix} - \boldsymbol{\beta}^* - (\widehat{\boldsymbol{\beta}}_{M_n} - \boldsymbol{\beta}^*)\right\}'\mathbf{X}'\mathbf{X}\left\{\begin{pmatrix} \widehat{\boldsymbol{\beta}}_m \\ \mathbf{0} \end{pmatrix} - \boldsymbol{\beta}^* - (\widehat{\boldsymbol{\beta}}_{M_n} - \boldsymbol{\beta}^*)\right\}
\end{aligned}$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

$$\geq \frac{1}{2}n^{-1}\left\{\begin{pmatrix}\widehat{\boldsymbol{\beta}}_m\\ \mathbf{0}\end{pmatrix}-\boldsymbol{\beta}^*\right\}'\mathbf{X}'\mathbf{X}\left\{\begin{pmatrix}\widehat{\boldsymbol{\beta}}_m\\ \mathbf{0}\end{pmatrix}-\boldsymbol{\beta}^*\right\}-n^{-1}(\widehat{\boldsymbol{\beta}}_{M_n}-\boldsymbol{\beta}^*)'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}_{M_n}-\boldsymbol{\beta}^*)$$

$$\geq \frac{1}{2}\lambda_{min}(n^{-1}\mathbf{X}'\mathbf{X})\left\|\begin{pmatrix}\widehat{\boldsymbol{\beta}}_m\\ \mathbf{0}\end{pmatrix}-\boldsymbol{\beta}^*\right\|^2+O_p(K_n/n)$$

$$\geq \frac{1}{2}\kappa_1\times\sum_{j\in\mathcal{A}\setminus\mathcal{A}_{M_n-1}}\beta_j^{*2}+O_p(K_n/n)$$

$$\geq \frac{1}{2}\kappa_1 c_\tau n^{-\tau}+O_p(K_n/n)$$

$$=\frac{1}{2}\kappa_1 c_\tau n^{-\tau}+o_p(n^{-\tau}). \tag{A.2}$$

Combining (A.1), (A.2), and Condition (C1), we complete the proof of Theorem 1.

Appendix B. Proof of Theorem 2

Note that $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{X}_c\boldsymbol{\beta}_c+\mathbf{e}=\mathbf{X}\boldsymbol{\beta}^*+(\mathbf{I}_n-\mathbf{P}_{M_n})\mathbf{X}_c\boldsymbol{\beta}_c+\mathbf{e}$. So $\widehat{\boldsymbol{\beta}}_{M_n}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}=\boldsymbol{\beta}^*+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$ and

$$\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{M_n}-\boldsymbol{\beta}^*)=\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{e}=\sum_{i=1}^n r_i e_i,$$

where $r_i=\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_i'$ is the ith row of $\mathbf{X}$. It is easy to see that

$$\sum_{i=1}^n r_i^2=\sum_{i=1}^n\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\boldsymbol{\alpha}=\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\sum_{i=1}^n\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\boldsymbol{\alpha}$$
$$=\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}\boldsymbol{\alpha}=\boldsymbol{\alpha}'\boldsymbol{\alpha}=1.$$

Following the proof of Theorem 3.3 of Zou and Zhang (2009), under Conditions (C2) and (C4), Lyapunov's condition for the central limit theorem can be established. Thus, $\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{M_n}-\boldsymbol{\beta}^*)\to_d N(0,\sigma^2)$. For $m<M_n$, by (A.14) of Zhang et al. (2020), we have $\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}\left\{\begin{pmatrix}\widehat{\boldsymbol{\beta}}_m\\ \mathbf{0}\end{pmatrix}-\boldsymbol{\beta}^*\right\}=O_p(n^{\frac{1}{2}})$. Therefore, when $\phi_n K_n M_n/n^{\frac{1}{2}-\tau}\to 0$, we have

$$\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}\{\widehat{\boldsymbol{\beta}}(\widehat{\mathbf{w}})-\boldsymbol{\beta}^*\}$$
$$=\sum_{m=1}^{M_n-1}\widehat{w}_m\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}\left\{\begin{pmatrix}\widehat{\boldsymbol{\beta}}_m\\ \mathbf{0}\end{pmatrix}-\boldsymbol{\beta}^*\right\}+\widehat{w}_{M_n}\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{M_n}-\boldsymbol{\beta}^*)$$
$$=O_p\left(\frac{\phi_n K_n M_n}{n^{\frac{1}{2}-\tau}}\right)+\boldsymbol{\alpha}'(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{M_n}-\boldsymbol{\beta}^*)(1+o_p(1))\to_d N(0,\sigma^2).$$

Appendix C. Proof of Theorem 3

By direct calculation, we have

$$L_n(\mathbf{w})=\sum_{m=1}^{M_n}w_m^2 a_m+2\sum_{m<l}w_m w_l a_l+\|\mathbf{e}\|^2-2\sum_{m=1}^{M_n}w_m\mathbf{e}'(\mathbf{I}_n-\mathbf{P}_m)\mathbf{y} \tag{A.3}$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

and $L_n(\mathbf{w}^0_{M_n}) = a_{M_n} + \|\mathbf{e}\|^2 - 2\mathbf{e}'(\mathbf{I}_n - \mathbf{P}_{M_n})\mathbf{y}$. So

$$L_n(\mathbf{w}) - L_n(\mathbf{w}^0_{M_n}) = \sum_{m=1}^{M_n} w_m^2 a_m + 2\sum_{m<l} w_m w_l a_l - a_{M_n} + A_{n,1} + A_{n,2} - A_{n,3} - A_{n,4},$$

where $A_{n,1} = 2\sum_{m=1}^{M_n} w_m \mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}$, $A_{n,2} = 2\sum_{m=1}^{M_n} w_m \mathbf{e}'\mathbf{P}_m\mathbf{e}$, $A_{n,3} = 2\mathbf{e}'\mathbf{P}_{M_n}\boldsymbol{\mu}$, and $A_{n,4} = 2\mathbf{e}'\mathbf{P}_{M_n}\mathbf{e}$.

For any $\epsilon > 0$,

$$\begin{aligned}
&P\left(\max_{1\le m\le M_n} |\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}| > n^{1-\tau}\epsilon\right) \\
&\le \sum_{m=1}^{M_n} P(|\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}| > n^{1-\tau}\epsilon) \le \sum_{m=1}^{M_n} n^{-2+2\tau}\epsilon^{-2} E\{\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}\}^2 \\
&\le \sum_{m=1}^{M_n} n^{-2+2\tau}\epsilon^{-2} C\|\mathbf{P}_m\boldsymbol{\mu}\|^2 \qquad \textbf{(A.4)} \\
&\le \sum_{m=1}^{M_n} n^{-2+2\tau}\epsilon^{-2} C\|\boldsymbol{\mu}\|^2 = O\left(\frac{M_n}{n^{1-2\tau}}\right) \le O\left(\frac{K_n}{n^{1-2\tau}}\right) \to 0,
\end{aligned}$$

where (A.4) holds by Theorem 2 of Whittle (1960) and C is a positive constant unrelated to m and n. So $\max_{1\le m\le M_n} |\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}| = o_p(n^{1-\tau})$. Since $|A_{n,1}| \le 2\sum_{m=1}^{M_n} w_m |\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}| \le 2\sum_{m=1}^{M_n} w_m \max_{1\le m\le M_n} |\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}| = 2\max_{1\le m\le M_n} |\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}| = o_p(n^{1-\tau})$, we have $A_{n,1}/n = o_p(n^{-\tau})$. Similarly, $A_{n,3}/n = o_p(n^{-\tau})$. Moreover,

$$\begin{aligned}
P\left(\max_{1\le m\le M_n} \mathbf{e}'\mathbf{P}_m\mathbf{e} > n^{1-\tau}\epsilon\right) &= P(\mathbf{e}'\mathbf{P}_{M_n}\mathbf{e} > n^{1-\tau}\epsilon) \\
&\le \frac{E(\mathbf{e}'\mathbf{P}_{M_n}\mathbf{e})}{n^{1-\tau}\epsilon} = \frac{\sigma^2 K_n}{n^{1-\tau}\epsilon} = O\left(\frac{K_n}{n^{1-\tau}}\right) \to 0.
\end{aligned}$$

So $\max_{1\le m\le M_n} \mathbf{e}'\mathbf{P}_m\mathbf{e} = o_p(n^{1-\tau})$. Furthermore, we have $A_{n,2}/n = o_p(n^{-\tau})$ and $A_{n,4}/n = o_p(n^{-\tau})$.

Denote $\eta_n = n^{-1}(\widehat{\boldsymbol{\beta}}_{M_n} - \boldsymbol{\beta}^*)'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}_{M_n} - \boldsymbol{\beta}^*) = O_p(K_n/n)$. From (A.2), we have $n^{-1}(a_m - a_{M_n}) \ge \frac{1}{2}\kappa_1 c_\tau n^{-\tau} - \eta_n$. So

$$\begin{aligned}
&n^{-1}\{L_n(\mathbf{w}) - L_n(\mathbf{w}^0_{M_n})\} \\
&= \sum_{m=1}^{M_n} w_m^2 \frac{a_m}{n} + 2\sum_{m<l} w_m w_l \frac{a_l}{n} - \frac{a_{M_n}}{n} + o_p(n^{-\tau}) \\
&\ge \sum_{m=1}^{M_n-1} w_m^2\left(\frac{a_{M_n}}{n} + \frac{1}{2}\kappa_1 c_\tau n^{-\tau} - \eta_n\right) + w_{M_n}^2 \frac{a_{M_n}}{n}
\end{aligned}$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

$$
\begin{aligned}
&+2\sum_{m<l\le M_n-1} w_m w_l\left(\frac{a_{M_n}}{n}+\frac{1}{2}\kappa_1 c_\tau n^{-\tau}-\eta_n\right)\\
&+2\sum_{m<M_n} w_m w_{M_n}\frac{a_{M_n}}{n}-\frac{a_{M_n}}{n}+o_p(n^{-\tau})\\
&=\sum_{m=1}^{M_n-1} w_m^2\left(\frac{1}{2}\kappa_1 c_\tau n^{-\tau}-\eta_n\right)+2\sum_{m<l\le M_n-1} w_m w_l\left(\frac{1}{2}\kappa_1 c_\tau n^{-\tau}-\eta_n\right)+o_p(n^{-\tau})\\
&=\left(\sum_{m=1}^{M_n-1} w_m\right)^2\left(\frac{1}{2}\kappa_1 c_\tau n^{-\tau}-\eta_n\right)+o_p(n^{-\tau})\\
&=(1-w_{M_n})^2\left(\frac{1}{2}\kappa_1 c_\tau n^{-\tau}-O_p(K_n/n)\right)+o_p(n^{-\tau})\\
&=n^{-\tau}\left\{(1-w_{M_n})^2\frac{1}{2}\kappa_1 c_\tau+o_p(1)\right\},
\end{aligned}
\tag{A.5}
$$

which shows the first part of Theorem 3.

For the second part, we just need to show that $P(1-w^*_{M_n}>\epsilon)\to 0$ for any given $\epsilon>0$. Actually, when $1-w^*_{M_n}>\epsilon$, following exactly the same arguments as above, we can show that

$$
0\ge n^{-1+\tau}\{L_n(\mathbf{w}^*)-L_n(\mathbf{w}^0_{M_n})\}\ge(1-w^*_{M_n})^2\frac{1}{2}\kappa_1 c_\tau+o_p(1)>\epsilon^2\frac{1}{2}\kappa_1 c_\tau+o_p(1).
$$

So $P(1-w^*_{M_n}>\epsilon)\le P(\epsilon^2\frac{1}{2}\kappa_1 c_\tau+o_p(1)\le 0)\to 0$.

Appendix D. Proof of Corollary 1

Under the assumed conditions, the first part is just a direct conclusion from Theorems 1 and 3. So we focus on the second part.

Note that

$$
\begin{aligned}
n^{-1}L_n(\mathbf{w}^0_{M_n})&=\frac{1}{n}\boldsymbol{\mu}'(\mathbf{I}_n-\mathbf{P}_{M_n})\boldsymbol{\mu}+\frac{1}{n}\mathbf{e}'\mathbf{P}_{M_n}\mathbf{e}\\
&=\boldsymbol{\beta}_c'\frac{\mathbf{X}_c'(I_n-\mathbf{P}_{M_n})\mathbf{X}_c}{n}\boldsymbol{\beta}_c+o_p(n^{-\tau})\\
&\ge\lambda_{min}\{n^{-1}\mathbf{X}_c'(I_n-\mathbf{P}_{M_n})\mathbf{X}_c\}\|\boldsymbol{\beta}_c\|^2+o_p(n^{-\tau})\\
&\ge\kappa_3 d_\tau n^{-\tau}+o_p(n^{-\tau}).
\end{aligned}
$$

So we just need to show that

$$
n^{-1}L_n(\widehat{\mathbf{w}})=n^{-1}L_n(\mathbf{w}^0_{M_n})+o_p(n^{-\tau})\text{ and }n^{-1}L_n(\mathbf{w}^*)=n^{-1}L_n(\mathbf{w}^0_{M_n})+o_p(n^{-\tau}).
$$

We can write

$$
L_n(\widehat{\mathbf{w}})=\sum_{m=1}^{M_n}\widehat{w}_m^2\|\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_m\|^2+2\sum_{m<l}\widehat{w}_m\widehat{w}_l<\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_m,\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_l>.
$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

In the proof of Theorem 3, we showed that $\max_{1\le m\le M_n}|\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}| = o_p(n^{1-\tau})$ and $\max_{1\le m\le M_n}\mathbf{e}'\mathbf{P}_m\mathbf{e} = o_p(n^{1-\tau})$. So

$$\max_{m<l}|<\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_m,\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_l> - L_n(\mathbf{w}_l^0)| = \max_{m<l}|\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}-\mathbf{e}'\mathbf{P}_l\boldsymbol{\mu}+\mathbf{e}'\mathbf{P}_m\mathbf{e}-\mathbf{e}'\mathbf{P}_l\mathbf{e}| = o_p(n^{1-\tau}).$$

Then

$$\begin{aligned}
&n^{-1}L_n(\widehat{\mathbf{w}})\\
&\quad=\sum_{m=1}^{M_n}\widehat{w}_m^2\frac{L_n(\mathbf{w}_m^0)}{n}+2\sum_{m<l}\widehat{w}_m\widehat{w}_l\frac{L_n(\mathbf{w}_l^0)}{n}+o_p(n^{-\tau})\\
&\quad=\sum_{m<M_n}\widehat{w}_m^2\frac{L_n(\mathbf{w}_m^0)}{n}+\widehat{w}_{M_n}^2\frac{L_n(\mathbf{w}_{M_n}^0)}{n}+2\sum_{m<l<M_n}\widehat{w}_m\widehat{w}_l\frac{L_n(\mathbf{w}_l^0)}{n}\\
&\qquad+2\sum_{m<M_n}\widehat{w}_m\widehat{w}_{M_n}\frac{L_n(\mathbf{w}_{M_n}^0)}{n}+o_p(n^{-\tau})\\
&\quad:=B_{n,1}+B_{n,2}+B_{n,3}+B_{n,4}+o_p(n^{-\tau}).
\end{aligned}$$

Note that when $\phi_nK_nM_n/n^{1-2\tau}\to 0$, we have $1-\widehat{w}_{M_n}=\sum_{m<M_n}\widehat{w}_m = O_p\left(\frac{\phi_nK_nM_n}{n^{1-\tau}}\right) = o_p(n^{-\tau})$ by Theorem 1. So

$$\begin{aligned}
B_{n,1}+B_{n,3} &\le \left(\sum_{m<M_n}\widehat{w}_m\right)^2\max_{m<M_n}\frac{L_n(\mathbf{w}_m^0)}{n}\\
&=(1-\widehat{w}_{M_n})^2\max_{m<M_n}\left\{\frac{1}{n}\boldsymbol{\mu}'(\mathbf{I}_n-\mathbf{P}_m)\boldsymbol{\mu}+\frac{1}{n}\mathbf{e}'\mathbf{P}_m\mathbf{e}\right\}\\
&\le(1-\widehat{w}_{M_n})^2\left(\frac{\|\boldsymbol{\mu}\|^2}{n}+\max_{m<M_n}\frac{1}{n}\mathbf{e}'\mathbf{P}_m\mathbf{e}\right)\\
&=o_p(n^{-2\tau})\{O(1)+o_p(n^{-\tau})\}=o_p(n^{-2\tau}),
\end{aligned}$$

and

$$\begin{aligned}
B_{n,2}+B_{n,4} &=\left(\widehat{w}_{M_n}^2+2\sum_{m<M_n}\widehat{w}_m\widehat{w}_{M_n}\right)n^{-1}L_n(\mathbf{w}_{M_n}^0)\\
&=\widehat{w}_{M_n}(2-\widehat{w}_{M_n})n^{-1}L_n(\mathbf{w}_{M_n}^0)\\
&=(1+o_p(n^{-\tau}))n^{-1}L_n(\mathbf{w}_{M_n}^0).
\end{aligned}$$

Furthermore, notice that

$$n^{-1}L_n(\mathbf{w}_{M_n}^0)=\frac{1}{n}\boldsymbol{\mu}'(\mathbf{I}_n-\mathbf{P}_{M_n})\boldsymbol{\mu}+o_p(n^{-\tau})\le\frac{\|\boldsymbol{\mu}\|^2}{n}+o_p(n^{-\tau})=O_p(1).$$

Then we have $n^{-1}L_n(\widehat{\mathbf{w}})=n^{-1}L_n(\mathbf{w}_{M_n}^0)+o_p(n^{-\tau})$.

When $K_n/n^{1-6\tau}\to 0$, by similar arguments to the proof of Theorem 3, we can show that $\max_{1\le m\le M_n}|\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}| = o_p(n^{1-3\tau})$ and $\max_{1\le m\le M_n}|\mathbf{e}'\mathbf{P}_m\mathbf{e}| = o_p(n^{1-3\tau})$. Then, similar to

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

(A.5), we can show that

$$n^{-1}\{L_n(\mathbf{w}^*)-L_n(\mathbf{w}^0_{M_n})\}\geq(1-w^*_{M_n})^2\left(\frac{1}{2}\kappa_1 c_\tau n^{-\tau}-O_p(K_n/n)\right)+o_p(n^{-3\tau}).$$

For any given $\epsilon>0$, when $1-w^*_{M_n}>n^{-\tau}\epsilon$, we have

$$0\geq n^{-1}\{L_n(\mathbf{w}^*)-L_n(\mathbf{w}^0_{M_n})\}\geq n^{-2\tau}\epsilon^2\left(\frac{1}{2}\kappa_1 c_\tau n^{-\tau}-O_p(K_n/n)\right)+o_p(n^{-3\tau})$$
$$=\frac{1}{2}\epsilon^2\kappa_1 c_\tau n^{-3\tau}+o_p(n^{-3\tau}).$$

So $P(1-w^*_{M_n}>n^{-\tau}\epsilon)\leq P(\frac{1}{2}\epsilon^2\kappa_1 c_\tau+o_p(1)\leq 0)\to 0$, i.e., $1-w^*_{M_n}=o_p(n^{-\tau})$. Then, following exactly the same proof as $n^{-1}L_n(\widehat{\mathbf{w}})=n^{-1}L_n(\mathbf{w}^0_{M_n})+o_p(n^{-\tau})$, we can show that $n^{-1}L_n(\mathbf{w}^*)=n^{-1}L_n(\mathbf{w}^0_{M_n})+o_p(n^{-\tau})$ and finish the proof.

Appendix E. Proof of Theorem 4

Denote $\zeta_n=n^{-1}(\widehat{\boldsymbol{\beta}}_{M_0+1}-\boldsymbol{\beta}_{M_0+1})'\mathbf{X}'_{M_0+1}\mathbf{X}_{M_0+1}(\widehat{\boldsymbol{\beta}}_{M_0+1}-\boldsymbol{\beta}_{M_0+1})=O_p(k_{M_0+1}/n)$. Under Conditions (C2') and (C3'), similar to (A.2), we can show that

$$n^{-1}(a_m-a_{M_0+1})\geq\frac{1}{2}\kappa_0 c_{\tau_0}n^{-\tau_0}-\zeta_n \text{ when } m<M_0+1. \tag{A.6}$$

When $K_n/n^{1-2\tau_0}\to 0$, similar to Theorem 3, we can show that $\max_{1\leq m\leq M_n}|\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}|=o_p(n^{1-\tau_0})$ and $\max_{1\leq m\leq M_n}\mathbf{e}'\mathbf{P}_m\mathbf{e}=o_p(n^{1-\tau_0})$.

Denote $\tilde{\mathbf{w}}=(\tilde{w}_1,\ldots,\tilde{w}_{M_n})'=(0,\ldots,0,\sum_{m=1}^{M_0+1}w_m,w_{M_0+2},\ldots,w_{M_n})'$. Then, by (A.3), we have

$$n^{-1}\{L_n(\mathbf{w})-L_n(\tilde{\mathbf{w}})\}$$
$$=\sum_{m=1}^{M_n}w_m^2\frac{a_m}{n}+2\sum_{m<l}w_m w_l\frac{a_l}{n}-\sum_{m=1}^{M_n}\tilde{w}_m^2\frac{a_m}{n}-2\sum_{m<l}\tilde{w}_m\tilde{w}_l\frac{a_l}{n}+o_p(n^{-\tau_0})$$
$$=\sum_{m\leq M_0+1}w_m^2\frac{a_m}{n}-\left(\sum_{m=1}^{M_0+1}w_m\right)^2\frac{a_{M_0+1}}{n}$$
$$+2\sum_{m<l<M_0+1}w_m w_l\frac{a_l}{n}+2w_{M_0+1}\left(\sum_{m<M_0+1}w_m\right)\frac{a_{M_0+1}}{n}+o_p(n^{-\tau_0})$$
$$\geq\sum_{m<M_0+1}w_m^2\left(\frac{a_{M_0+1}}{n}+\frac{1}{2}\kappa_0 c_{\tau_0}n^{-\tau_0}-\zeta_n\right)+w_{M_0+1}^2\frac{a_{M_0+1}}{n}-\left(\sum_{m=1}^{M_0+1}w_m\right)^2\frac{a_{M_0+1}}{n}$$
$$+2\sum_{m<l<M_0+1}w_m w_l\left(\frac{a_{M_0+1}}{n}+\frac{1}{2}\kappa_0 c_{\tau_0}n^{-\tau_0}-\zeta_n\right)$$
$$+2w_{M_0+1}\left(\sum_{m<M_0+1}w_m\right)\frac{a_{M_0+1}}{n}+o_p(n^{-\tau_0})$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

$$
\begin{aligned}
&= \left(\sum_{m<M_0+1} w_m\right)^2 \left(\frac{1}{2}\kappa_0 c_{\tau_0} n^{-\tau_0} - \zeta_n\right) + o_p(n^{-\tau_0}) \\
&= \left(\sum_{m<M_0+1} w_m\right)^2 \left(\frac{1}{2}\kappa_0 c_{\tau_0} n^{-\tau_0} - O_p(k_{M_0+1}/n)\right) + o_p(n^{-\tau_0}) \\
&= n^{-\tau_0}\left\{\left(\sum_{m<M_0+1} w_m\right)^2 \frac{1}{2}\kappa_0 c_{\tau_0} + o_p(1)\right\}. \qquad \textbf{(A.7)}
\end{aligned}
$$

On the other hand, for any $M_0+1 \le m < l$, since model m and l are correct models, we have $(\mathbf{I}_n - \mathbf{P}_m)\boldsymbol{\mu} = 0$ and $(\mathbf{I}_n - \mathbf{P}_l)\boldsymbol{\mu} = 0$. Then

$$
\begin{aligned}
< \mathbf{P}_m\mathbf{y} - \boldsymbol{\mu}, \mathbf{P}_l\mathbf{y} - \boldsymbol{\mu} > &= < \mathbf{P}_m\mathbf{e} - (\mathbf{I}_n - \mathbf{P}_m)\boldsymbol{\mu}, \mathbf{P}_l\mathbf{e} - (\mathbf{I}_n - \mathbf{P}_l)\boldsymbol{\mu} > \\
&= < \mathbf{P}_m\mathbf{e}, \mathbf{P}_l\mathbf{e} > = \mathbf{e}'\mathbf{P}_m\mathbf{P}_l\mathbf{e} = \mathbf{e}'\mathbf{P}_m\mathbf{e} = L_n(\mathbf{w}_m^0).
\end{aligned}
$$

For any $m > M_0+1$, $L_n(\mathbf{w}_m^0) = \|\mathbf{e}\|^2 - a_m > \|\mathbf{e}\|^2 - a_{M_0+1} = L_n(\mathbf{w}_{M_0+1}^0)$. So

$$
\begin{aligned}
L_n(\tilde{\mathbf{w}}) &= \| \sum_{m=M_0+1}^{M_n} \tilde{w}_m(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_m)\|^2 \\
&= \sum_{m=M_0+1}^{M_n} \tilde{w}_m^2 L_n(\mathbf{w}_m^0) + 2 \sum_{M_0+1\le m<l\le M_n} \tilde{w}_m\tilde{w}_l < \mathbf{P}_m\mathbf{y} - \boldsymbol{\mu}, \mathbf{P}_l\mathbf{y} - \boldsymbol{\mu} > \\
&= \sum_{m=M_0+1}^{M_n} \tilde{w}_m^2 L_n(\mathbf{w}_m^0) + 2 \sum_{M_0+1\le m<l\le M_n} \tilde{w}_m\tilde{w}_l L_n(\mathbf{w}_m^0) \\
&\ge \sum_{m=M_0+1}^{M_n} \tilde{w}_m^2 L_n(\mathbf{w}_{M_0+1}^0) + 2 \sum_{M_0+1\le m<l\le M_n} \tilde{w}_m\tilde{w}_l L_n(\mathbf{w}_{M_0+1}^0) \\
&= \left(\sum_{m=M_0+1}^{M_n} \tilde{w}_m\right)^2 L_n(\mathbf{w}_{M_0+1}^0) \\
&= L_n(\mathbf{w}_{M_0+1}^0), \qquad \textbf{(A.8)}
\end{aligned}
$$

where the equation holds if and only if $\tilde{w}_{M_0+1} = 1$, i.e., $\sum_{m=1}^{M_0+1} w_m = 1$.

For any $\mathbf{w} \ne \mathbf{w}_{M_0+1}^0$, if $\sum_{m<M_0+1} w_m > 0$, then by (A.7) and (A.8), we have

$$
\begin{aligned}
n^{-1+\tau_0}\{L_n(\mathbf{w}) - L_n(\mathbf{w}_{M_0+1}^0)\} &\ge n^{-1+\tau_0}\{L_n(\mathbf{w}) - L_n(\tilde{\mathbf{w}})\} \\
&\ge \left(\sum_{m<M_0+1} w_m\right)^2 \frac{1}{2}\kappa_0 c_{\tau_0} + o_p(1).
\end{aligned}
$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

Then

$$P(n^{-1+\tau_0}\{L_n(\mathbf{w})-L_n(\mathbf{w}^0_{M_0+1})\}>0)\geq P\left(\left(\sum_{m<M_0+1}w_m\right)^2\frac{1}{2}\kappa_0\tau_0+o_p(1)>0\right)\to 1,$$

i.e., $P(L_n(\mathbf{w})-L_n(\mathbf{w}^0_{M_0+1})>0)\to 1$.

If $\sum_{m<M_0+1}w_m=0$, then $\mathbf{w}=\tilde{\mathbf{w}}$ and $\sum_{m=1}^{M_0+1}w_m\neq 1$ since $\mathbf{w}\neq\mathbf{w}^0_{M_0+1}$. Then, by (A.8), we have

$$L_n(\mathbf{w})-L_n(\mathbf{w}^0_{M_0+1})=L_n(\tilde{\mathbf{w}})-L_n(\mathbf{w}^0_{M_0+1})>0.$$

Appendix F. Proof of Lemma 1

Denote $\tilde{\mathbf{w}}^*=(0,\ldots,0,\sum_{m=1}^{M_0+1}w^*_m,w^*_{M_0+2},\ldots,w^*_{M_n})'\in\mathcal{H}_n^\delta$. When $K_n/n^{1-6\tau_0}\to 0$, by similar arguments to the proof of Theorem 3, we can show that $\max_{1\leq m\leq M_n}|\mathbf{e}'\mathbf{P}_m\boldsymbol{\mu}|=o_p(n^{1-3\tau_0})$ and $\max_{1\leq m\leq M_n}|\mathbf{e}'\mathbf{P}_m\mathbf{e}|=o_p(n^{1-3\tau_0})$. When $\sum_{m<M_0+1}w^*_m\geq\delta n^{-\tau_0}$, similar to (A.7), we can show that

$$0\geq n^{-1}\{L_n(\mathbf{w}^*)-L_n(\tilde{\mathbf{w}}^*)\}\geq\left(\sum_{m<M_0+1}w^*_m\right)^2\frac{1}{2}\kappa_0c_{\tau_0}n^{-\tau_0}+o_p(n^{-3\tau_0})$$

$$\geq\frac{1}{2}\delta^2\kappa_0c_{\tau_0}n^{-3\tau_0}+o_p(n^{-3\tau_0}).$$

So

$$P\left(\sum_{m<M_0+1}w^*_m\geq\delta n^{-\tau_0}\right)\leq P\left(0\geq\frac{1}{2}\delta^2\kappa_0c_{\tau_0}+o_p(1)\right)\to 0.$$

Then, by the definition of $\mathcal{H}_n^\delta$, we have $P(\sum_{m<M_0+1}w^*_m=0)\to 1$.

When $\sum_{m<M_0+1}w^*_m=0$, following the same arguments as (A.8), we have $L_n(\mathbf{w}^*)\geq L_n(\mathbf{w}^0_{M_0+1})$, where the equation holds if and only if $w^*_{M_0+1}=1$. On the other hand, by the definition of $\mathbf{w}^*$, $L_n(\mathbf{w}^*)\leq L_n(\mathbf{w}^0_{M_0+1})$. So $L_n(\mathbf{w}^*)=L_n(\mathbf{w}^0_{M_0+1})$ and $w^*_{M_0+1}=1$. Then

$$P(w^*_{M_0+1}=1)\geq P\left(\sum_{m<M_0+1}w^*_m=0\right)\to 1,$$

i.e., $P(\mathbf{w}^*=\mathbf{w}^0_{M_0+1})\to 1$.

Appendix G. Proof of Theorem 5

By Lemma 1, we have $P(L_n(\mathbf{w}^*)=L_n(\mathbf{w}^0_{M_0+1}))\to 1$. So we just need to consider whether $L_n(\widehat{\mathbf{w}})/L_n(\mathbf{w}^0_{M_0+1})$ converges to 1 in probability or not.

For any $m<M_0+1$, when $\widehat{w}_m\neq 0$, similar to (A.1), we can show that

$$\widehat{w}_m\leq(a_m-a_{M_0+1})^{-1}\phi_n\widehat{\sigma}^2(k_{M_0+1}-k_m).$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

Then, by (A.6) and Condition (C1), we have

$$\max_{m<M_0+1} \widehat{w}_m = O_p\left(\frac{\phi_n k_{M_0+1}}{n^{1-\tau_0}}\right). \tag{A.9}$$

For any $m > M_0+1$, define

$$\begin{aligned}\tilde{\mathbf{w}}_m &= (\widehat{w}_1, \ldots, \widehat{w}_{M_0}, \widehat{w}_{M_0+1} + \widehat{w}_m, \widehat{w}_{M_0+2}, \ldots, \widehat{w}_{m-1}, 0, \widehat{w}_{m+1}, \ldots, \widehat{w}_{M_n})' \\ &= \widehat{\mathbf{w}} + (0, \ldots, 0, \widehat{w}_m, 0, \ldots, 0, -\widehat{w}_m, 0, \ldots, 0)'.\end{aligned}$$

Then

$$\begin{aligned}0 &\le \mathcal{G}_n(\widetilde{\mathbf{w}}_m) - \mathcal{G}_n(\widehat{\mathbf{w}}) \\ &= [(\widetilde{\mathbf{w}}_m - \widehat{\mathbf{w}}) + 2\widehat{\mathbf{w}}]' \boldsymbol{\Phi} (\widetilde{\mathbf{w}}_m - \widehat{\mathbf{w}}) + \phi_n \widehat{\sigma}^2 (\widetilde{\mathbf{w}}_m - \widehat{\mathbf{w}})' \mathbf{K} \\ &= \widehat{w}_m^2 (a_{M_0+1} - a_m) + 2\widehat{w}_m \sum_{l=1}^{M_n} \widehat{w}_l (a_{\max\{l, M_0+1\}} - a_{\max\{l,m\}}) + \phi_n \widehat{\sigma}^2 \widehat{w}_m (k_{M_0+1} - k_m) \\ &= \widehat{w}_m^2 (a_{M_0+1} - a_m) + 2\widehat{w}_m \sum_{l=1}^{M_0+1} \widehat{w}_l (a_{M_0+1} - a_m) \\ &\quad + 2\widehat{w}_m \sum_{l=M_0+2}^{m-1} \widehat{w}_l (a_l - a_m) + \phi_n \widehat{\sigma}^2 \widehat{w}_m (k_{M_0+1} - k_m) \\ &\le \widehat{w}_m^2 (a_{M_0+1} - a_m) + 2\widehat{w}_m \sum_{l=1}^{m-1} \widehat{w}_l (a_{M_0+1} - a_m) + \phi_n \widehat{\sigma}^2 \widehat{w}_m (k_{M_0+1} - k_m) \\ &\le \widehat{w}_m^2 (a_{M_0+1} - a_m) + 2\widehat{w}_m (a_{M_0+1} - a_m) + \phi_n \widehat{\sigma}^2 \widehat{w}_m (k_{M_0+1} - k_m).\end{aligned}$$

So, when $\widehat{w}_m \neq 0$, we have

$$\widehat{w}_m \le (k_m - k_{M_0+1})^{-1} \widehat{\sigma}^{-2} \phi_n^{-1} (\widehat{w}_m^2 + 2\widehat{w}_m)(a_{M_0+1} - a_m). \tag{A.10}$$

Note that $E((a_{M_0+1} - a_{M_n})/(K_n - k_{M_0+1})) = \sigma^2$, so

$$\frac{a_{M_0+1} - a_m}{k_m - k_{M_0+1}} \le \frac{a_{M_0+1} - a_{M_n}}{K_n - k_{M_0+1}} \times \frac{K_n - k_{M_0+1}}{k_{M_0+2} - k_{M_0+1}} = O_p(K_n).$$

Then, by (A.10) and Condition (C1), we have

$$\max_{m>M_0+1} \widehat{w}_m = O_p\left(\frac{K_n}{\phi_n}\right). \tag{A.11}$$

Moreover, we have

$$\begin{aligned}\max_{m<M_0+1} L_n(\mathbf{w}_m^0) &= \max_{m<M_0+1} \{\boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{P}_m)\boldsymbol{\mu} + \mathbf{e}'\mathbf{P}_m\mathbf{e}\} \\ &\le \|\boldsymbol{\mu}\|^2 + \max_{1\le m\le M_n} \mathbf{e}'\mathbf{P}_m\mathbf{e} \\ &= O(n) + o_p(n) = O_p(n),\end{aligned} \tag{A.12}$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

and

$$\max_{m\geq M_0+1} L_n(\mathbf{w}_m^0) = \max_{m\geq M_0+1} \mathbf{e}'\mathbf{P}_m\mathbf{e} \leq \mathbf{e}'\mathbf{P}_{M_n}\mathbf{e} = O_p(K_n). \tag{A.13}$$

(1) When $\phi_n \to \infty$, following Lemma 1 of Zhang et al. (2020), we can show that $P(\sum_{m>M_0+1} \widehat{w}_m = 0) \to 1$. Then $\widehat{w}_{M_0+1} \to_p 1$, and with probability going to 1, by (A.9) and (A.11)–(A.13), we have

$$\begin{aligned}
&L_n(\widehat{\mathbf{w}}) \\
&= \| \sum_{m\leq M_0+1} \widehat{w}_m(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_m)\|^2 \\
&\leq \sum_{m\leq M_0+1} \widehat{w}_m^2 L_n(\mathbf{w}_m^0) + 2 \sum_{m<l\leq M_0+1} \widehat{w}_m\widehat{w}_l\sqrt{L_n(\mathbf{w}_m^0)L_n(\mathbf{w}_l^0)} \\
&= \widehat{w}_{M_0+1}^2 L_n(\mathbf{w}_{M_0+1}^0) + \sum_{m<M_0+1} \widehat{w}_m^2 L_n(\mathbf{w}_m^0) + 2 \sum_{m<l<M_0+1} \widehat{w}_m\widehat{w}_l\sqrt{L_n(\mathbf{w}_m^0)L_n(\mathbf{w}_l^0)} \\
&\quad + 2 \sum_{m<M_0+1} \widehat{w}_m\widehat{w}_{M_0+1}\sqrt{L_n(\mathbf{w}_m^0)L_n(\mathbf{w}_{M_0+1}^0)} \\
&= \widehat{w}_{M_0+1}^2 L_n(\mathbf{w}_{M_0+1}^0) + O_p\left(\frac{\phi_n^2 k_{M_0+1}^2 M_0}{n^{2-2\tau_0}}\right) O_p(n) + O_p\left(\frac{\phi_n^2 k_{M_0+1}^2 M_0^2}{n^{2-2\tau_0}}\right) O_p(n) \\
&\quad + O_p\left(\frac{\phi_n k_{M_0+1} M_0}{n^{1-\tau_0}}\right) O_p(\sqrt{n}) O_p(\sqrt{k_{M_0+1}}) \\
&= \widehat{w}_{M_0+1}^2 L_n(\mathbf{w}_{M_0+1}^0) + O_p\left(\frac{\phi_n^2 k_{M_0+1}^2 M_0}{n^{1-2\tau_0}}\right) + O_p\left(\frac{\phi_n^2 k_{M_0+1}^2 M_0^2}{n^{1-2\tau_0}}\right) + O_p\left(\frac{\phi_n k_{M_0+1}^{3/2} M_0}{n^{\frac{1}{2}-\tau_0}}\right) \\
&= \widehat{w}_{M_0+1}^2 L_n(\mathbf{w}_{M_0+1}^0) + o_p(1).
\end{aligned}$$

So, with probability going to 1,

$$1 \leq \frac{L_n(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}_n^\delta} L_n(\mathbf{w})} = \frac{L_n(\widehat{\mathbf{w}})}{L_n(\mathbf{w}^*)} = \frac{L_n(\widehat{\mathbf{w}})}{L_n(\mathbf{w}_{M_0+1}^0)} \leq \widehat{w}_{M_0+1}^2 + o_p(1) \to_p 1.$$

(2) When $\phi_n = 2$, $\sum_{m\geq M_0+1} \widehat{w}_m \to_p 1$ and

$$\begin{aligned}
L_n(\widehat{\mathbf{w}}) &= \| \sum_{m=1}^{M_n} \widehat{w}_m(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_m)\|^2 \\
&= \| \sum_{m=M_0+1}^{M_n} \widehat{w}_m(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_m)\|^2 + \| \sum_{m<M_0+1} \widehat{w}_m(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_m)\|^2 \\
&\quad + 2 < \sum_{m<M_0+1} \widehat{w}_m(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_m), \sum_{l\geq M_0+1} \widehat{w}_l(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_l) > ,
\end{aligned}$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

where the second term is bounded by

$$\sum_{m<M_0+1} \widehat{w}_m^2 L_n(\mathbf{w}_m^0) + 2 \sum_{m<l<M_0+1} \widehat{w}_m \widehat{w}_l \sqrt{L_n(\mathbf{w}_m^0) L_n(\mathbf{w}_l^0)}$$
$$= O_p\left(\frac{\phi_n^2 k_{M_0+1}^2 M_0}{n^{2-2\tau_0}}\right) O_p(n) + O_p\left(\frac{\phi_n^2 k_{M_0+1}^2 M_0^2}{n^{2-2\tau_0}}\right) O_p(n) = o_p(1),$$

and the third term is bounded by

$$2 \sum_{m<M_0+1} \sum_{l\geq M_0+1} \widehat{w}_m \widehat{w}_l \sqrt{L_n(\mathbf{w}_m^0) L_n(\mathbf{w}_l^0)}$$
$$= 2 \sum_{m<M_0+1} \widehat{w}_m \sqrt{L_n(\mathbf{w}_m^0)} \times \sum_{l\geq M_0+1} \widehat{w}_l \sqrt{L_n(\mathbf{w}_l^0)}$$
$$= O_p\left(\frac{\phi_n k_{M_0+1} M_0}{n^{1-\tau_0}}\right) O_p(\sqrt{n}) \times O_p(\sqrt{K_n})$$
$$= O_p\left(\frac{\phi_n k_{M_0+1} \sqrt{K_n} M_0}{n^{\frac{1}{2}-\tau_0}}\right) = o_p(1).$$

So

$$L_n(\widehat{\mathbf{w}}) = \Big\| \sum_{m=M_0+1}^{M_n} \widehat{w}_m (\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_m) \Big\|^2 + o_p(1)$$
$$= \sum_{m=M_0+1}^{M_n} \widehat{w}_m^2 L_n(\mathbf{w}_m^0) + 2 \sum_{M_0+1\leq m<l\leq M_n} \widehat{w}_m \widehat{w}_l \langle \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_m, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_l \rangle + o_p(1)$$
$$= \sum_{m=M_0+1}^{M_n} \widehat{w}_m^2 L_n(\mathbf{w}_m^0) + 2 \sum_{M_0+1\leq m<l\leq M_n} \widehat{w}_m \widehat{w}_l L_n(\mathbf{w}_m^0) + o_p(1)$$
$$= \sum_{m=M_0+1}^{M_n} \left(\widehat{w}_m^2 + 2\widehat{w}_m \sum_{l=m+1}^{M_n} \widehat{w}_l \right) L_n(\mathbf{w}_m^0) + o_p(1)$$
$$= \sum_{m=M_0+1}^{M_n} \left(\widehat{w}_m^2 + 2\widehat{w}_m \sum_{l=m+1}^{M_n} \widehat{w}_l \right) (\|\mathbf{e}\|^2 - a_m) + o_p(1)$$
$$= \sum_{m=M_0+1}^{M_n} \left(\widehat{w}_m^2 + 2\widehat{w}_m \sum_{l=m+1}^{M_n} \widehat{w}_l \right) (\|\mathbf{e}\|^2 - a_{M_0+1})$$
$$+ \sum_{m=M_0+1}^{M_n} \left(\widehat{w}_m^2 + 2\widehat{w}_m \sum_{l=m+1}^{M_n} \widehat{w}_l \right) (a_{M_0+1} - a_m) + o_p(1)$$
$$= \left(\sum_{m=M_0+1}^{M_n} \widehat{w}_m \right)^2 L_n(\mathbf{w}_{M_0+1}^0) + \sum_{m=M_0+1}^{M_n} \left(\widehat{w}_m^2 + 2\widehat{w}_m \sum_{l=m+1}^{M_n} \widehat{w}_l \right) (a_{M_0+1} - a_m) + o_p(1).$$

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

Then

$$\frac{L_n(\widehat{\mathbf{w}})}{L_n(\mathbf{w}^0_{M_0+1})} = 1 + o_p(1) + \frac{\sum_{m=M_0+1}^{M_n}\left(\widehat{w}_m^2 + 2\widehat{w}_m\sum_{l=m+1}^{M_n}\widehat{w}_l\right)(a_{M_0+1}-a_m)+o_p(1)}{L_n(\mathbf{w}^0_{M_0+1})}$$
$$\geq 1 + o_p(1) + \frac{\sum_{m>M_0+1}\widehat{w}_m^2(a_{M_0+1}-a_m)+o_p(1)}{L_n(\mathbf{w}^0_{M_0+1})}.$$

If $L_n(\widehat{\mathbf{w}})/\inf_{\mathbf{w}\in\mathcal{H}_n^\delta} L_n(\mathbf{w}) \to_p 1$, then by the fact that $P(\inf_{\mathbf{w}\in\mathcal{H}_n^\delta} L_n(\mathbf{w}) = L_n(\mathbf{w}^0_{M_0+1})) \to 1$ and the above equality, we have $\sum_{m>M_0+1}\widehat{w}_m^2(a_{M_0+1}-a_m)/L_n(\mathbf{w}^0_{M_0+1}) \to_p 0$. Note that $L_n(\mathbf{w}^0_m) = \|\mathbf{e}\|^2 - a_m = \mathbf{e}'\mathbf{P}_m\mathbf{e}$, for $m \geq M_0+1$. Therefore, $\sum_{m>M_0+1}\widehat{w}_m^2\left(\frac{\mathbf{e}'\mathbf{P}_m\mathbf{e}}{\mathbf{e}'\mathbf{P}_{M_0+1}\mathbf{e}} - 1\right) \to_p 0$.

REFERENCES

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B. Petroc and F. Csake (eds.), *Second International Symposium on Information Theory*, pp. 267–281. Akademiai Kiado.

Ando, T. & K.-C. Li (2014) A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* 109, 254–265.

Ando, T. & K.-C. Li (2017) A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics* 45, 2654–2679.

Box, G.E.P. (1976) Science and statistics. *Journal of the American Statistical Association* 71, 791–799.

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

Fang, F., J. Li, & X. Xia (2020) Semiparametric model averaging prediction for dichotomous response. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2020.09.008.

Fang, F. & M. Liu (2020) Limit of the optimal weight in least squares model averaging with non-nested models. *Economics Letters* 196, 109586.

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

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

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

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

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

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

Kitagawa, T. & C. Muris (2016) Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics* 193, 271–289.

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

Leung, G. & A.R. Barron (2006) Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* 52, 3396–3410.

Li, C., Q. Li, J.S. Racine, & D. Zhang (2018a) Optimal model averaging of varying coefficient models. *Statistica Sinica* 28, 2795–2809.

https://doi.org/10.1017/S0266466622000032 Published online by Cambridge University Press

Li, D., O. Linton, & Z. Lu (2015) A flexible semiparametric forecasting model for time series. *Journal of Econometrics* 187, 345–357.
Li, J., X. Xia, W.K. Wong, & D. Nott (2018b) Varying-coefficient semiparametric model averaging prediction. *Biometrics* 74, 1417–1428.
Liang, H., G. Zou, A.T.K. Wan, & X. Zhang (2011) Optimal weight choice for frequentist model averaging estimators. *Journal of the American Statistical Association* 106, 1053–1066.
Liao, J., X. Zong, X. Zhang, & G. Zou (2019) Model averaging based on leave-subject-out cross-validation for vector autoregressions. *Journal of Econometrics* 209, 35–60.
Liu, C.-A. (2015) Distribution theory of the least squares averaging estimator. *Journal of Econometrics* 186, 142–159.
Liu, Q. & R. Okui (2013) Heteroskedasticity-robust C_p model averaging. *The Econometrics Journal* 16, 463–472.
Longford, N.T. (2005) Editorial: Model selection and efficiency—Is "which model…?" the right question? *Journal of the Royal Statistical Society, Series A* 168, 469–472.
Peng, J. & Y. Yang (2021) On improvability of model selection by model averaging. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2020.12.003.
Phillips, P.C.B. (2005) Automated discovery in econometrics. *Econometric Theory* 21, 3–20.
Raftery, A.E. & Y. Zheng (2003) Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association* 98, 931–938.
Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics* 6, 461–464.
Shao, J. (1997) An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* 7, 221–264.
Wan, A.T.K., X. Zhang, & S. Wang (2014) Frequentist model averaging for multinomial and ordered logit models. *International Journal of Forecasting* 30, 118–128.
Wan, A.T.K., X. Zhang, & G. Zou (2010) Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156, 277–283.
Whittle, P. (1960) Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and Its Applications* 5, 302–305.
Yang, Y. (2001) Adaptive regression by mixing. *Journal of the American Statistical Association* 96, 574–586.
Yang, Y. (2003) Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica* 13, 783–809.
Yuan, Z. & Y. Yang (2005) Combining linear regression models: When and how? *Journal of the American Statistical Association* 100, 1202–1214.
Zhang, X. (2015) Consistency of model averaging estimators. *Economics Letters* 130, 120–123.
Zhang, X. (2021) A new study on asymptotic optimality of least squares model averaging. *Econometric Theory* 37, 388–407.
Zhang, X. & H. Liang (2011) Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics* 39, 174–200.
Zhang, X. & C.-A. Liu (2019) Inference after model averaging in linear regression models. *Econometric Theory* 35, 816–841.
Zhang, X. & W. Wang (2019) Optimal model averaging estimation for partially linear models. *Statistica Sinica* 29, 693–718.
Zhang, X., D. Yu, G. Zou, & H. Liang (2016) Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association* 111, 1775–1790.
Zhang, X., G. Zou, & R.J. Carroll (2015) Model averaging based on Kullback–Leibler distance. *Statistica Sinica* 25, 1583–1598.
Zhang, X., G. Zou, & H. Liang (2014) Model averaging and weight choice in linear mixed effects models. *Biometrika* 101, 205–218.
Zhang, X., G. Zou, H. Liang, & R.J. Carroll (2020) Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association* 115, 972–984.

Zhang, Y. & Y. Yang (2015) Cross-validation for selecting a model selection procedure. *Journal of Econometrics* 187, 95–112.

Zheng, H., K.-W. Tsui, X. Kang, & X. Deng (2017) Cholesky-based model averaging for covariance matrix estimation. *Statistical Theory and Related Fields* 1, 48–58.

Zhu, R., A.T.K. Wan, X. Zhang, & G. Zou (2019) A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association* 114, 882–892.

Zou, H. & H. Zhang (2009) On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics* 37, 1733–1751.