Econometric Theory, **0**, 2025, 1–33.
doi:10.1017/S0266466624000355

FROM MODEL SELECTION TO MODEL AVERAGING: A COMPARISON FOR NESTED LINEAR MODELS

WENCHAO XU
Shanghai University of International Business and Economics

XINYU ZHANG
Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Model selection (MS) and model averaging (MA) are two popular approaches when many candidate models exist. Theoretically, the estimation risk of an oracle MA is not larger than that of an oracle MS because the former is more flexible, but a foundational issue is this: Does MA offer a substantial improvement over MS? Recently, seminal work by Peng and Yang (2022) has answered this question under nested models with linear orthonormal series expansion. In the current paper, we further respond to this question under linear nested regression models. A more general nested framework, heteroscedastic and autocorrelated random errors, and sparse coefficients are allowed in the current paper, giving a scenario that is more common in practice. A remarkable implication is that MS can be significantly improved by MA under certain conditions. In addition, we further compare MA techniques with different weight sets. Simulation studies illustrate the theoretical findings in a variety of settings.

1. INTRODUCTION

In the past two decades, model selection (MS) has received growing attention in statistics and econometrics. When a list of candidate models is considered, MS attempts to select a single best model. A large number of MS criteria have been proposed in the literature, including the Akaike information criterion (AIC; Akaike, 1973), Mallows' C_p (Mallows, 1973), Bayesian information criterion (BIC; Schwarz, 1978), and cross-validation (CV; Allen, 1974; Stone, 1974). Model

We thank the Editor (Peter C.B. Phillips), the Co-Editor (Liangjun Su), two anonymous referees, Jingfu Peng, Yundong Tu, and Yuhong Yang for many constructive comments and suggestions. Xu's research was partially supported by the National Natural Science Foundation of China (12101591). Zhang's research was partially supported by the National Natural Science Foundation of China (71925007, 72091212, and 71988101), Beijing Natural Science Foundation (Z240004), and the CAS Project for Young Scientists in Basic Research (YSBR-008). Address correspondence to Xinyu Zhang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; e-mail: xinyu@amss.ac.cn.

averaging (MA) is an alternative to MS and operates by taking a weighted average of the estimators or predictions from candidate models; Thus, MA is a smoothed extension of MS and can potentially reduce risk relative to MS (Magnus et al., 2010; Yuan and Yang, 2005).

Asymptotic efficiency (or asymptotic optimality) is a key theoretical goal pursued in both MS and MA research. It states that the risk (or loss) of MS (or MA) is equivalent to that of the infeasible oracle candidate model (or averaged estimator/prediction). For MS methods, AIC is asymptotically efficient, but BIC is not in a nonparametric framework (Shao, 1997; Shibata, 1983). Li (1987) established the asymptotic efficiency of C_p and leave-one-out CV (LOO-CV) for the homoscedastic nonparametric regression. Andrews (1991) extended the results of Li (1987) to the case of heteroscedastic errors. See Ding et al. (2018) for a recent review of the properties of MS methods.

For the asymptotic optimality of MA, Hansen (2007) established the asymptotic optimality for Mallows model averaging (MMA) when the candidate models are nested and the weights are restricted to a discrete set. Wan et al. (2010) and Zhang (2021) extended the result of Hansen (2007) to a non-nested model setting with continuous weights. Hansen and Racine (2012) established the asymptotic optimality of Jackknife model averaging (JMA) for heteroscedastic errors with the weights contained in a discrete set. Zhang et al. (2013) broadened Hansen and Racine (2012)'s scope of analysis to dependent data and a continuous weight set. Liu and Okui (2013) proposed a heteroscedasticity-robust MA with asymptotic optimality. Ando and Li (2014, 2017) removed the conventional MA restriction that the sum of weights equals one and established the asymptotic optimality of JMA for high-dimensional linear models and generalized linear models. Zhao et al. (2016) broadened Ando and Li (2014)'s scope of analysis to dependent data. Zhang and Wang (2019), Fang et al. (2019), and Feng et al. (2022) established the asymptotic optimality of MA in nonparametric, missing data, and nonlinear models, respectively.

However, most literature focuses on optimal properties (e.g., asymptotic efficiency) of MS or MA in their own terms. Although many successful empirical advancements in MA have been demonstrated (see, e.g., Moral-Benito, 2015; Magnus and Luca, 2016; Lehrer and Xie, 2017; Steel, 2020), theoretical investigations comparing MS and MA are still lacking. Recently, Peng and Yang (2022) made a seminal contribution to filling this gap. They studied the foundational matter of comparing the oracle/optimal MS with MA procedures in a nested model setting with series expansion and have some remarkable findings. However, a limit of Peng and Yang (2022) is that their study was built with three restrictions: orthonormal design, homoscedastic and independent random errors, and non-sparse coefficients, which could make the study not applicable in real-world scenarios. The goal of the current paper is to broaden the scope of analysis of Peng and Yang (2022) to a general model setting and to answer additional important questions. Specifically, our main contributions are as follows.

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

(i) Without the three aforementioned restrictions, we answer the questions: Does MA offer a significant improvement over MS? If so, when? Moreover, we partition the predictor variables into groups, and the group size is allowed to be larger than 1, which leads to a more general nested model framework than that in Peng and Yang (2022). We define a sequence of indices (say $\theta_{n,m}, m = 1, \ldots, d_n$) for grouped variables and find that the decaying order of this sequence determines when MA is substantially better than MS. Specifically, when the number of candidate models is large enough and $\theta_{n,m}$ decays slowly in m, the benefit of MA over MS is real. However, when either the number of candidate models is too small or it is large enough and $\theta_{n,m}$ decays fast in m, MA has no real advantage over MS. As a result, the analysis of Peng and Yang (2022) becomes a special case of ours. In finite simulation studies, we compare the performance of MMA or JMA with several MS methods, including AIC, BIC, and LOO-CV. The simulation results support our theoretical findings.

(ii) In the MA literature, MA weights can be selected from three different weight sets, namely the unit simplex (Liu, 2015; Wan et al., 2010; Zhang et al., 2013), the unit hypercube (Ando and Li, 2014, 2017), and the discrete weight set (Hansen, 2007; Hansen and Racine, 2012). In our work, we broaden the scope to compare MAs with weights belonging to these three weight sets. Two main findings emerge. First, relaxing the weight set of unit simplex to unit hypercube does not reduce the risk of MA asymptotically. Second, when the number of candidate models is large enough and $\theta_{n,m}$ decays slowly in m, discretizing the unit simplex can enlarge the risk of MA by a substantial multiple.

Extending the work of Peng and Yang (2022) presents several theoretical challenges. First, our study relies on deriving explicit expressions for the optimal risks of MS and MA, which become challenging when addressing non-orthonormal predictors and heteroscedastic and autocorrelated error terms. Second, Peng and Yang (2022)'s work requires a large enough number of candidate models, whereas our work includes scenarios where it is small. This introduces additional complexities for our analysis. Last, the expression of optimal MA risk with the discrete weight set is much more complicated, making the comparison of MAs with different weight sets challenging.

The rest of the paper is organized as follows. Section 2 provides the model setting and four important questions to be answered. Section 3 presents the main results from the comparison of MS and MA. Section 4 considers the comparison of MAs with three different weight sets. Section 5 provides two examples to verify the theoretical results, and Section 6 presents the results of finite sample simulations. Section 7 concludes the paper. The proofs of our theoretical results are contained in the Appendix and the Supplementary Material. The Supplementary Material also contains some additional theoretical results and simulation studies.

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

2. MODEL SETTING AND QUESTIONS

Consider the model

$$y_i = \mu_i + \varepsilon_i = \sum_{j=1}^{p_n} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are random errors, $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top, j = 1, \ldots, p_n$ are predictor variables, and $p_n (p_n < n)$ is the number of the predictors. In matrix notation, (1) can be written as $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. Furthermore, we assume that $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$ so that $\boldsymbol{\mu} = E(\mathbf{y}|\mathbf{X})$, where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{p_n}\}$. Note that the condition $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$ rules out dynamic models in which lagged dependent variables work as regressors. We denote $\mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is a positive definite matrix with the (i,j)th element being $E(\varepsilon_i \varepsilon_j|\mathbf{X})$. We permit $\boldsymbol{\Omega}$ to be dependent on $\mathbf{X}$ and nondiagonal, thus allowing the errors to be both conditionally heteroscedastic and autocorrelated. In the current paper, we use bold forms to denote vectors or matrices.

Following Hansen (2007, 2014), Feng and Liu (2020), and Zhang et al. (2020), we consider nested models, where the mth model uses the first ν_m predictor variables, such that $0 = \nu_0 < \nu_1 < \nu_2 < \cdots < \nu_{q_n - 1} < \nu_{q_n} = p_n$, and q_n is a positive integer. This nested framework essentially requires that the predictor variables are partitioned into q_n groups and the grouped predictors are ordered, where the mth group of predictors is $\{x_{i,\nu_{m-1}+1}, \ldots, x_{i,\nu_m}\}$, and its size is $\nu_m - \nu_{m-1}$ for $m = 1, \ldots, q_n$.

We consider the first M_n $(2 \le M_n \le q_n)$ nested candidate models for MS and MA. Let $\mathbf{X}_m$ be the $n \times \nu_m$ design matrix of the mth candidate model. We assume $\mathbf{X}_m$ is of full rank for any $m \in \{1, \ldots, M_n\}$. Then, under the mth model, the estimator of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}}_m = \mathbf{X}_m (\mathbf{X}_m^\top \mathbf{X}_m)^{-1} \mathbf{X}_m^\top \mathbf{y} \equiv \mathbf{P}_m \mathbf{y},$$

where $\mathbf{P}_m = \mathbf{X}_m (\mathbf{X}_m^\top \mathbf{X}_m)^{-1} \mathbf{X}_m^\top$ is the hat matrix. An MS method selects an index (say $m^\star$) from the index set $\mathcal{H}_n = \{1, \ldots, M_n\}$ and estimates $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}}_{m^\star}$ using the selected model. Let $\mathbf{w} = (w_1, \ldots, w_{M_n})^\top$ be a weight vector belonging to the unit simplex of $\mathbb{R}^{M_n}$:

$$\mathcal{W}_n = \left\{ \mathbf{w} \in [0,1]^{M_n} : \sum_{m=1}^{M_n} w_m = 1 \right\}.$$

Then, the MA estimator of $\boldsymbol{\mu}$ with weight vector $\mathbf{w}$ is

$$\hat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{m=1}^{M_n} w_m \hat{\boldsymbol{\mu}}_m = \sum_{m=1}^{M_n} w_m \mathbf{P}_m \mathbf{y} \equiv \mathbf{P}(\mathbf{w}) \mathbf{y},$$

where $\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M_n} w_m \mathbf{P}_m$. The measurement of estimation accuracy is the squared prediction risk, which is defined as $R_n^{\mathrm{MS}}(m) = E\{L_n^{\mathrm{MS}}(m)|\mathbf{X}\}$ and

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

$R_n^{\mathrm{MA}}(\mathbf{w}) = E\{L_n^{\mathrm{MA}}(\mathbf{w})|\mathbf{X}\}$ for MS and MA, respectively. In these,

$$L_n^{\mathrm{MS}}(m) = \|\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 \quad \text{and} \quad L_n^{\mathrm{MA}}(\mathbf{w}) = \|\hat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2$$

are the squared prediction loss for MS and MA, respectively, and $\|\cdot\|^2$ is the squared Euclidean norm.

Let m_n^* be the oracle optimal model index that minimizes $R_n^{\mathrm{MS}}(m)$ in $\mathcal{H}_n$, and let $\mathbf{w}_n^*$ be the oracle optimal weight vector that minimizes $R_n^{\mathrm{MA}}(\mathbf{w})$ in $\mathcal{W}_n$. Since m_n^* and $\mathbf{w}_n^*$ are unknown, we cannot apply them in practice. However, using the asymptotically optimal MS and MA procedures mentioned in Section 1, we can select a model index $\hat{m}_n$ and weights $\hat{\mathbf{w}}_n$ in the sense that

$$\frac{R_n^{\mathrm{MS}}(\hat{m}_n)}{R_n^{\mathrm{MS}}(m_n^*)} \to_p 1 \quad \text{and} \quad \frac{R_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n)}{R_n^{\mathrm{MA}}(\mathbf{w}_n^*)} \to_p 1, \tag{2}$$

where $\to_p$ denotes convergence in probability. All limiting processes are studied with respect to $n \to \infty$. Note that in (2), $\hat{m}_n$ and $\hat{\mathbf{w}}_n$ are directly plugged into the expressions for $R_n^{\mathrm{MS}}(m)$ and $R_n^{\mathrm{MA}}(\mathbf{w})$. Yang (1999) and Zhang et al. (2020) introduced a new type of asymptotic optimality for MS and MA as follows:

$$\frac{E\{L_n^{\mathrm{MS}}(\hat{m}_n)\}}{R_n^{\mathrm{MS}}(m_n^*)} \to_{a.s.} 1 \quad \text{and} \quad \frac{E\{L_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n)\}}{R_n^{\mathrm{MA}}(\mathbf{w}_n^*)} \to_{a.s.} 1, \tag{3}$$

where $\to_{a.s.}$ denotes convergence almost surely (a.s.). Compared to (2), (3) takes into account the randomness of $\hat{m}_n$ and $\hat{\mathbf{w}}_n$.

Since MS is a special case of MA with weights concentrating on a single model, it is obvious that $R_n^{\mathrm{MS}}(m_n^*) \geq R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$. The first task of this paper is essentially to explore the improvability of the oracle regression model m_n^* by the oracle MA. Let $\Delta_n = R_n^{\mathrm{MS}}(m_n^*) - R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ denote the potential risk reduction of the oracle MA compared to the oracle MS. We first consider the following two key questions:

Q1. Can $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ bring in a smaller order than $R_n^{\mathrm{MS}}(m_n^*)$? That is, can $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)/R_n^{\mathrm{MS}}(m_n^*) = o(1)$ happen a.s.?

Q2. Is Δ_n a substantial reduction relative to $R_n^{\mathrm{MS}}(m_n^*)$ or actually negligible? If both can happen, when is MA substantially better than MS?

Remark 1. The vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{p_n}$ are said to be orthonormal if

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}x_{ik} = 0, \quad 1 \leq j \neq k \leq p_n. \tag{4}$$

Note that (4) is satisfied for a nonparametric regression with orthonormal series expansion. Peng and Yang (2022) have answered Questions Q1 and Q2 under the assumption that the predictors are orthonormal (i.e., (4) holds), and the error ε_i's are homoscedastic and independent, which can be restrictive for some applications. In the current paper, we answer the same questions in a more general model setting without the above assumptions. Moreover, Peng and Yang (2022) used the typical nested framework $\nu_m = m$, which is also relaxed in the current paper.

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

In addition to the weight set $\mathcal{W}_n$, two other weight sets are popular in the MA literature: the unit hypercube

$$\mathcal{Q}_n = \left\{\mathbf{w} \in [0,1]^{M_n} : 0 \le w_m \le 1\right\},$$

and the discrete weight set

$$\mathcal{W}_n(N) = \left\{\mathbf{w} : w_m \in \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\right\}, \sum_{m=1}^{M} w_m = 1\right\},$$

for a fixed positive integer N. The weight set $\mathcal{Q}_n$ removes the restriction that weights add up to 1 in $\mathcal{W}_n$. Ando and Li (2014) used this weight set to study the optimal MA for the first time in a high-dimensional linear regression setting. More recently, this weight set was further applied to various regression models, including high-dimensional generalized linear model (Ando and Li, 2017), high-dimensional quantile regression (Wang et al., 2023), and high-dimensional survival analysis (He et al., 2020; Yan et al., 2021). In the MA literature, the discrete set $\mathcal{W}_n(N)$ is often only considered to establish some asymptotic theories of MA for technical convenience; see, e.g., Hansen (2007), Hansen and Racine (2012), and Fang and Liu (2020). In practice, different weight sets can produce different results, hence the comparison of MA techniques with different weight sets is very important. In the literature, they are compared by numerical examples; see, for example, Ando and Li (2014) and Wang et al. (2023). The next task of the current paper is on the theoretical comparison of MA with the weights belonging to different weight sets.

Let $\tilde{\mathbf{w}}_n^*$ and $\mathbf{w}_{n,N}^*$ be the oracle optimal weights that minimize $R_n^{\mathrm{MA}}(\mathbf{w})$ in $\mathcal{Q}_n$ and $\mathcal{W}_n(N)$, respectively. Since $\mathcal{W}_n(N) \subset \mathcal{W}_n \subset \mathcal{Q}_n$, we know $R_n^{\mathrm{MA}}(\tilde{\mathbf{w}}_n^*) \le R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \le R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*)$. This result implies that the weight relaxation could bring a smaller optimal risk of MA, and the restriction of the weight set $\mathcal{W}_n$ to $\mathcal{W}_n(N)$ could lead to a larger optimal risk of MA. However, it is unclear whether the risk reduction of optimal MA by relaxing the weight set $\mathcal{W}_n$ to $\mathcal{Q}_n$ is substantial and whether the risk increment of optimal MA by restricting the weight set $\mathcal{W}_n$ to $\mathcal{W}_n(N)$ is substantial. Since $\mathcal{W}_n$ is widely used, we use $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ as a benchmark for the comparisons. Therefore, we consider the other two key issues as follows:

Q3. Is $R_n^{\mathrm{MA}}(\mathbf{w}_n^*) - R_n^{\mathrm{MA}}(\tilde{\mathbf{w}}_n^*)$ a substantial reduction relative to $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ or actually negligible? If both can happen, when is $\tilde{\mathbf{w}}_n^*$ substantially better than $\mathbf{w}_n^*$?

Q4. Is the risk increment $R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) - R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ substantial relative to $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ or actually negligible? If both can happen, when $\mathbf{w}_n^*$ is substantially better than $\mathbf{w}_{n,N}^*$?

The answers to Questions Q1 and Q2 broaden the scope of Peng and Yang (2022)'s work on the advantages of MA over MS. The answers to Questions Q3 and Q4 provide a previously unavailable insight on the relative strengths of MA with these three weight sets.

Throughout this paper, we use the following symbols. For two positive sequences a_n and b_n, $a_n \succeq b_n$ means $b_n = O(a_n)$, and $a_n \asymp b_n$ means both

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

$a_n \succeq b_n$ and $b_n \succeq a_n$. Also, $a_n \sim b_n$ means that $a_n/b_n \to 1$. For two stochastic sequences a_n and b_n, $a_n \asymp_p b_n$ means that there exist $0 < \underline{c} \le \bar{c} < \infty$ such that $\underline{c}\{1+o_p(1)\} \le |a_n/b_n| \le \bar{c}\{1+o_p(1)\}$ for all sufficiently large n; $a_n \sim_p b_n$ means that $a_n/b_n \to_p 1$. Let $\lfloor a \rfloor$ be the greatest integer less than or equal to a. Let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ be the minimum and maximum eigenvalues of a matrix $\mathbf{M}$, respectively.

3. COMPARISONS OF MS AND MA PROCEDURES

In this section, we aim to answer Questions Q1 and Q2. In Section 3.1, we introduce some important notation and assumptions. Then, in Section 3.2, we theoretically investigate the comparison of the oracle optimal model m_n^* and the oracle MA. In Section 3.3, we extend the obtained results to compare two specific asymptotically optimal MS and MA procedures.

We first introduce some notation. Let $\mathbf{X}_m^c$ be the $n \times (p_n - \nu_m)$ design matrix that consists of the predictors excluded from the mth model. Let $\boldsymbol{\beta}_m = (\beta_1, \ldots, \beta_{\nu_m})^\top$ and $\boldsymbol{\beta}_m^c = (\beta_{\nu_m+1}, \ldots, \beta_{p_n})^\top$. Then, $\boldsymbol{\mu} = \mathbf{X}_m\boldsymbol{\beta}_m + \mathbf{X}_m^c\boldsymbol{\beta}_m^c$. For convenience, we further assume $\mathbf{X}_m$ is of full rank for any $m \in \{M_n+1, \ldots, q_n\}$.

3.1. Grouped Variable Importance

In this subsection, we introduce notation used to measure the importance of each group of predictors, whose decaying order determines when MA is substantially better than MS. For the mth model ($m = 1, \ldots, q_n$), define

$$\theta_{n,m} = \frac{\boldsymbol{\beta}_{m-1}^{c\top}\mathbf{X}_{m-1}^{c\top}(\mathbf{I}_n - \mathbf{P}_{m-1})\mathbf{X}_{m-1}^c\boldsymbol{\beta}_{m-1}^c - \boldsymbol{\beta}_m^{c\top}\mathbf{X}_m^{c\top}(\mathbf{I}_n - \mathbf{P}_m)\mathbf{X}_m^c\boldsymbol{\beta}_m^c}{n\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}}, \tag{5}$$

where $\mathbf{P}_0 = \mathbf{0}$. Two remarks are worth noticing. First, $\theta_{n,m}$ is obtained during the derivation of the oracle optimal model index m_n^* and weights $\mathbf{w}_n^*$; see Equations (A.4) and (A.6) in the Appendix for details. Second, the numerator in (5) has a simpler expression $\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}$, as mentioned in Remark 2 below, and the complex form is presented here for ease of explanation. The two terms in the numerator of (5) measure the importance of the remaining predictors after excluding from predictors from models $m-1$ and m, respectively. Consequently, $\theta_{n,m}$ can be regarded as the importance of the mth group of variables in some sense. Therefore, we refer to $\theta_{n,m}$ as the grouped variable importance (GVI). In the following remark, we provide another explanation for $\theta_{n,m}$.

Remark 2 (Another explanation for $\theta_{n,m}$). By some simple calculations, we can get a simpler form for $\theta_{n,m}$ as follows

$$\theta_{n,m} = \frac{n^{-1}\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}}{\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}}. \tag{6}$$

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

In the first formula of the Appendix, we demonstrate that the risk of $\hat{\boldsymbol{\mu}}_m$ is given by

$$\begin{aligned} R_n^{\mathrm{MS}}(m) &= E\left\{\|\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 | \mathbf{X}\right\} \\ &= \mathrm{tr}\left[\{E(\hat{\boldsymbol{\mu}}_m|\mathbf{X}) - \boldsymbol{\mu}\}\{E(\hat{\boldsymbol{\mu}}_m|\mathbf{X}) - \boldsymbol{\mu}\}^\top\right] + \mathrm{tr}\left\{\mathrm{Var}(\hat{\boldsymbol{\mu}}_m|\mathbf{X})\right\} \\ &= \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_m)\boldsymbol{\mu} + \mathrm{tr}(\mathbf{P}_m\boldsymbol{\Omega}). \end{aligned}$$

Here, $\boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_m)\boldsymbol{\mu}$ is the trace of the squared bias term of $\hat{\boldsymbol{\mu}}_m$, and $\mathrm{tr}(\mathbf{P}_m\boldsymbol{\Omega})$ is the trace of the conditional variance term of $\hat{\boldsymbol{\mu}}_m$; these two traces are nonincreasing and increasing in m, respectively, because of the nested framework of candidate models. Therefore, by (6), the numerator and denominator of $\theta_{n,m}$ are the decrement of the squared bias scaled by n and the increment of the variance of risks, respectively, when adding the mth group of predictors to the $(m-1)$th model. When $\boldsymbol{\Omega} = \sigma^2\mathbf{I}_n$ and the size of groups is fixed (say v^*), the increment of the variance of risks is fixed to be $v^*\sigma^2$.

Next, we impose additional assumptions for the model (1) to obtain simple forms of GVI as follows.

Case 1 (Homoscedastic and uncorrelated errors). If the error terms are homoscedastic and uncorrelated with variance $\sigma^2 > 0$, it is easy to see that $\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\} = \sigma^2(v_m - v_{m-1})$. Then, $\theta_{n,m}$ reduces to

$$\theta_{n,m} = \frac{\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}}{n\sigma^2(v_m - v_{m-1})}.$$

Case 2 (Orthonormal design). If the orthonormal design assumption (4) is satisfied, it is easy to show that $\boldsymbol{\beta}_m^{c\top}\mathbf{X}_m^{c\top}(\mathbf{I}_n - \mathbf{P}_m)\mathbf{X}_m^c\boldsymbol{\beta}_m^c = n\|\boldsymbol{\beta}_m^c\|^2$. Then, $\theta_{n,m}$ reduces to

$$\theta_{n,m} = \frac{\sum_{j=v_{m-1}+1}^{v_m}\beta_j^2}{\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}}.$$

Furthermore, we consider $\boldsymbol{\Omega} = \sigma^2(\rho^{|k-l|})_{k,l=1,\ldots,n}$ for some $|\rho| < 1$ and $\sigma^2 > 0$, i.e., the error terms follow a first-order autoregressive (AR) process with the autocorrelation coefficient ρ. From Trench (1999), $\frac{1-|\rho|}{1+|\rho|}\sigma^2 \le \lambda_{\min}(\boldsymbol{\Omega}) \le \lambda_{\max}(\boldsymbol{\Omega}) \le \frac{1+|\rho|}{1-|\rho|}\sigma^2$. Then, $\frac{1-|\rho|}{1+|\rho|}\sigma^2(v_m - v_{m-1}) \le \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\} \le \frac{1+|\rho|}{1-|\rho|}\sigma^2(v_m - v_{m-1})$. As a result, $\theta_{n,m}$ satisfies

$$\frac{1-|\rho|}{1+|\rho|}\frac{\sum_{j=v_{m-1}+1}^{v_m}\beta_j^2}{\sigma^2(v_m - v_{m-1})} \le \theta_{n,m} \le \frac{1+|\rho|}{1-|\rho|}\frac{\sum_{j=v_{m-1}+1}^{v_m}\beta_j^2}{\sigma^2(v_m - v_{m-1})}. \tag{7}$$

When $\rho = 0$, it further reduces to the following Case 3.

Case 3 (Orthonormal design and homoscedastic and uncorrelated errors). If the orthonormal design assumption (4) is satisfied and the error terms are homoscedastic and uncorrelated with variance $\sigma^2 > 0$, then $\theta_{n,m}$ has a simple form

$$\theta_{n,m} = \frac{\sum_{j=v_{m-1}+1}^{v_m}\beta_j^2}{\sigma^2(v_m - v_{m-1})}.$$

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

Case 4 (Model setting of Peng and Yang (2022)). In the model setting of Peng and Yang (2022) (i.e., under the assumptions of orthonormal design (4), homoscedastic and uncorrelated errors with variance $\sigma^2 > 0$, and the typical nested framework $\nu_m = m$), $\theta_{n,m}$ reduces to

$$\theta_{n,m} = \beta_m^2/\sigma^2,$$

where β_m is the coefficient of x_{im}.

From these cases (especially Cases 2–4), the numerator of $\theta_{n,m}$ is determined by the coefficients of the variables in the mth group. This further implies that $\theta_{n,m}$ can serve as a measure of the grouped variable importance.

We now introduce two assumptions for the model (1), which are commonly used in the MA literature.

Assumption 1. $\|\boldsymbol{\mu}\|^2/n = O(1)$ a.s.

Assumption 2. There are constants $0 < c_1 \le c_2 < \infty$ such that $c_1 \le \lambda_{\min}(\boldsymbol{\Omega}) \le \lambda_{\max}(\boldsymbol{\Omega}) \le c_2$ a.s.

Assumption 1 requires the average of μ_i^2 to be bounded. Assumption 2 excludes the degeneracy and divergence of the error terms. As shown in Case 2, when $\boldsymbol{\Omega} = \sigma^2(\rho^{|k-l|})_{k,l=1,\ldots,n}$, we can set $c_1 = \frac{1-|\rho|}{1+|\rho|}\sigma^2$ and $c_2 = \frac{1+|\rho|}{1-|\rho|}\sigma^2$. Similar assumptions can be found in previous works such as Wan et al. (2010), Zhang et al. (2013), and Liu et al. (2016). It is worth noting that Assumption 1 has limitations. For instance, assuming $\lambda_{\min}(n^{-1}\mathbf{X}_{q_n}^\top \mathbf{X}_{q_n}) \ge c_0$ a.s. for some constant $c_0 > 0$, we have $\|\boldsymbol{\mu}\|^2/n \ge c_0 \sum_{j=1}^{p_n} \beta_j^2$ a.s. Consequently, when $\sum_{j=1}^{p_n} \beta_j^2 \to \infty$, we have $\|\boldsymbol{\mu}\|^2/n \to \infty$ a.s., indicating that Assumption 1 does not hold in this case.

Given the property $\mathbf{P}_m\mathbf{P}_l = \mathbf{P}_{\min(m,l)}$ in the nested model setting, it can be easily verified that $\mathbf{P}_m - \mathbf{P}_{m-1}$ is a symmetric idempotent matrix. Therefore, the numerator of (6) satisfies $0 \le n^{-1}\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} \le \|\boldsymbol{\mu}\|^2/n$, indicating that the numerator of (5) is nonnegative. Additionally, using Assumption 2, we have

$$c_1 \le c_1(\nu_m - \nu_{m-1}) \le \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\} \le c_2(\nu_m - \nu_{m-1}) \text{ a.s.}$$

Combining the above results, (6), and Assumption 1, we can conclude that $\sum_{m=1}^{q_n} \theta_{n,m} \le \|\boldsymbol{\mu}\|^2/(nc_1) < \infty$ a.s. and

$$0 \le \theta_{n,m} < \infty \text{ for any } m = 1, \ldots, q_n \text{ a.s.}$$

Let K_0 be a sufficiently large constant. We further impose an assumption on $\theta_{n,m}$ as follows.

Assumption 3. For each $n \ge K_0$, $\{\theta_{n,m}\colon m = 1, \ldots, q_n\}$ is a nonincreasing sequence a.s.

Assumption 3 is an extension of Assumption 1 in Peng and Yang (2022). The nonincreasing ordering of $\{\theta_{n,m}\}_{m=1}^{q_n}$ allows us to conveniently characterize the unknown optimal model index m_n^* and weights $\mathbf{w}_n^*$, $\tilde{\mathbf{w}}_n^*$, and $\mathbf{w}_{n,N}^*$, and it essentially requires the predictors to be groupwise ordered from the most important to the least

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

important, i.e, the ordering of grouped variables is "correct". However, unlike Peng and Yang (2022), where the sequence $\{\theta_{n,m}\colon m = 1, \ldots, q_n\}$ must be positive, the current paper allows for some components in the sequence to be zeros, i.e., we allow some totally unimportant variables. For the aforementioned Cases 3 and 4, this means we allow some kind of sparsity of coefficients, which is an important property, especially for high-dimensional problems. For Case 2, from (7), we know that when the error terms follow a first-order AR process, Assumption 3 can be satisfied with certain constraints on β_j, ν_m, and ρ. Under Assumption 3, let $d_n = \max\{m \in \{1, \ldots, q_n\}\colon \theta_{n,m} > 0\}$ be the number of important groups of predictors. If $d_n < q_n$, the mth group of predictors is not important for $m = d_n + 1, \ldots, q_n$.

Next, we make the following assumptions.

Assumption 4. There exists a constant $V \geq 1$ independent of both m and n, satisfying $\max_{1 \leq m \leq d_n}(\nu_m - \nu_{m-1}) \leq V$ uniformly for $n \geq K_0$.

Assumption 5. There exists a nonstochastic positive sequence $\bar{\theta}_m, m = 1, \ldots, d_n$ such that for each $n \geq K_0$, $\min_{1 \leq m \leq d_n}(\theta_{n,m} - \bar{\theta}_m) \geq 0$ a.s.

Assumption 6. For each $n \geq K_0$, $R_n^{\mathrm{MS}}(m)$ first decreases and then increases as m varies from 1 to d_n a.s.

Assumption 4 means that when the predictors are partitioned into groups as described in Section 2, the sizes of all important groups do not grow to infinity as n increases. Hansen (2014) and Zhang et al. (2016a) also made assumptions on group sizes when comparing the risks of estimators using the full model and the MMA. Assumption 5 basically eliminates the case that $\theta_{n,m}$ goes to zero as n increases for any fixed m, excluding the local-to-zero asymptotic framework with fixed dimensions considered by Hjort and Claeskens (2003) and Liu (2015). We allow for the possibility that $\bar{\theta}_m$ tends to zero as $m \to \infty$. Note that in the model setting of Peng and Yang (2022), Assumptions 4 and 5 are obviously satisfied.

Assumption 6 requires that d_n be reasonably large. For example, when $d_n \equiv d_0$ is fixed and Assumptions 2 and 5 hold, by Equation (A.1) in the Appendix,

$$\begin{aligned} R_n^{\mathrm{MS}}(m) - R_n^{\mathrm{MS}}(m-1) &= n\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}\left(\frac{1}{n} - \theta_{n,m}\right) \\ &\leq n\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}\left(\frac{1}{n} - \bar{\theta}_m\right) < 0 \text{ a.s.} \end{aligned}$$

for $m = 2, \ldots, d_0$ and each sufficiently large n. This result implies that $R_n^{\mathrm{MS}}(m)$ is decreasing on $\{1, \ldots, d_0\}$ for each sufficiently large n a.s., and thus Assumption 6 is not satisfied. Under Assumptions 3 and 5, Assumption 6 is satisfied when $\theta_{n,d_n} < 1/n$ a.s. for each sufficiently large n, as derived below. Given Assumptions 3 and 5, and the condition $\theta_{n,d_n} < 1/n$ a.s., there exist an event $\mathcal{E}$ with $\Pr(\mathcal{E}) = 1$ and an $m_n' \in \{2, \ldots, d_n - 1\}$ such that on $\mathcal{E}$, $\theta_{n,m_n'} > 1/n \geq \theta_{n,m_n'+1}$. Consequently,

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

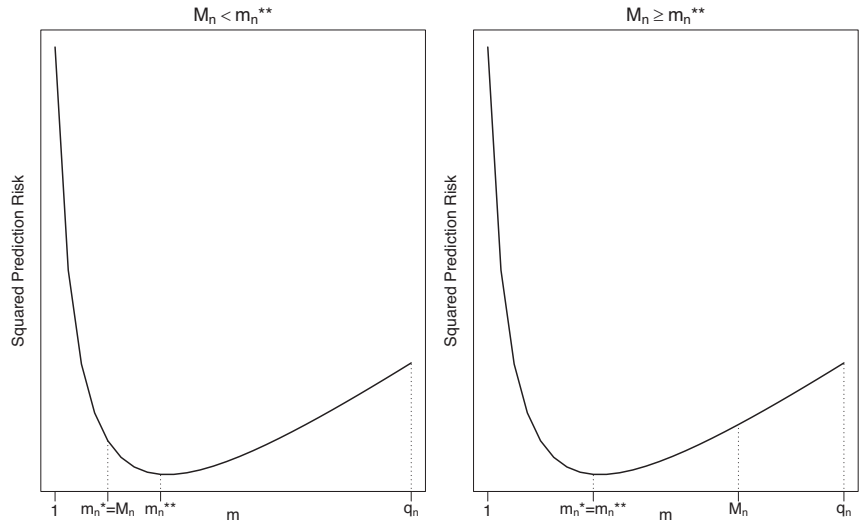

FIGURE 1. Two typical situations of the squared prediction risk $R_n^{\mathrm{MS}}(m)$ and the relationship between M_n and m_n^{**} under Assumption 6. In left panel: $M_n < m_n^{**}$. In right panel: $M_n \geq m_n^{**}$.

on $\mathcal{E}$, for $m = 2, \ldots, m_n'$,

$$R_n^{\mathrm{MS}}(m) - R_n^{\mathrm{MS}}(m-1) \leq n\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\mathbf{\Omega}\}\left(\frac{1}{n} - \theta_{n,m_n'}\right) < 0;$$

and for $m = m_n' + 1, \ldots, d_n$,

$$R_n^{\mathrm{MS}}(m) - R_n^{\mathrm{MS}}(m-1) \geq n\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\mathbf{\Omega}\}\left(\frac{1}{n} - \theta_{n,m_n'+1}\right) \geq 0;$$

and $R_n^{\mathrm{MS}}(d_n) - R_n^{\mathrm{MS}}(d_n - 1) > 0$. These results together imply that $R_n^{\mathrm{MS}}(m)$ is decreasing in $m \in \{1, \ldots, m_n'\}$ and increasing in $m \in \{m_n', \ldots, d_n\}$ a.s. Assumption 6 also implies that, for any sufficiently large n, $R_n^{\mathrm{MS}}(m)$ first decreases and then increases as m increases from 1 to q_n a.s.; refer to Figure 1. In the derivation of (A.2) in Peng and Yang (2022), they also used this assumption.

Let $m_n^{**} = \arg\min_{m \in \{1,\ldots,q_n\}} R_n^{\mathrm{MS}}(m)$ be the global optimal model index. We assume that m_n^{**} is unique. Note that m_n^{**} may not equal m_n^* because the number of candidate models M_n may be too small to include the m_n^{**}th model. In fact, under Assumption 6,

$$m_n^* = \min\{M_n, m_n^{**}\} = \begin{cases} M_n, & \text{if } M_n < m_n^{**}, \\ m_n^{**}, & \text{if } M_n \geq m_n^{**}. \end{cases}$$

See Figure 1 that shows two typical situations for m_n^*. Peng and Yang (2022) compared MS and MA under the assumption that M_n is large enough to include

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

m_n^{**}, i.e., $M_n \geq m_n^{**}$. However, that paper did not compare MS and MA when $M_n < m_n^{**}$. In this paper, we also relax this assumption and investigate the impact of M_n on the comparison of MS and MA.

In Sections 3.2 and 3.3, we shall show that the number of candidate models M_n and the decaying order of $\{\theta_{n,m}\}_{m=1}^{d_n}$ determine when MA is substantially better than MS.

3.2. A Comparison of Oracle Optimal MS and MA

In this subsection, we address Questions Q1 and Q2, specifically comparing the risks of the optimal MS and MA estimators. We begin with the following theorem on the order relationship of $R_n^{\mathrm{MS}}(m_n^*)$ and $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$, which provides an answer to Question Q1.

THEOREM 1 (Answer to Question Q1). *Suppose that Assumptions 1–3 and 6 hold. Then, for any sufficiently large n,*

$$R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \geq \frac{1}{2} R_n^{\mathrm{MS}}(m_n^*) \ a.s. \tag{8}$$

Moreover, the risks using m_n^* and $\mathbf{w}_n^*$ have the same order a.s., i.e.,

$$R_n^{\mathrm{MS}}(m_n^*) \asymp R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \ a.s. \tag{9}$$

Theorem 1 places no restriction on the number of candidate models M_n, whereas Peng and Yang (2022) imposed the condition $M_n \geq m_n^{**}$. Inequality (8) implies $\Delta_n \leq R_n^{\mathrm{MS}}(m_n^*)/2$ for sufficiently large n, suggesting that the potential risk reduction of MA compared to MS does not exceed half of optimal risk of MS. Equation (9) indicates that while MA has a smaller optimal risk than MS, it actually cannot reduce the increasing rate (or improve the decreasing rate) of risk by the optimal MS. Thus, even if the oracle model based on MS can be improved by MA, the potential advantage of MA in risk reduction is limited in terms of the increasing or decreasing rate.

We now turn our attention to Question Q2. We first present the following theorem on some elementary properties of the global optimal model index m_n^{**}, $R_n^{\mathrm{MS}}(m_n^*)$, and $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$. These properties are important for the subsequent analysis.

THEOREM 2. *Suppose that Assumptions 1–6 are satisfied. Then,*

(i) $m_n^{**} \to \infty$ *a.s.*
(ii) $R_n^{\mathrm{MS}}(m_n^*) \to \infty$ *and* $R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \to \infty$ *a.s.*

Theorem 2(i) suggests that, under certain mild assumptions, the index of the global optimal model diverges to infinity almost surely as $n \to \infty$. For practical applications, this result tells us that a diverging dimension should be utilized to achieve promising MS performance. Theorem 2(ii) reveals that the smallest risks

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

of MS and MA grow to infinity almost surely as the sample size n increases. In the proof of this theorem provided in the Supplementary Material, we demonstrate that $R_n^{\mathrm{MS}}(m_n^*)$ and $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ diverge at a rate no slower than m_n^{**}. It is worth noting Theorem 2(ii) does not require the restriction $M_n \geq m_n^{**}$, which was imposed by Peng and Yang (2022).

To investigate the impact of the number of candidate models M_n on the comparison of MS and MA, we consider the following two conditions for M_n.

Condition M1 (M_n Too Small). $\lim_{n\to\infty} M_n/m_n^{**} = 0$ a.s.
Condition M2 (M_n Large Enough). There exists a constant $\underline{c} > 0$ such that $M_n/m_n^{**} \geq \underline{c}$ holds a.s. for sufficiently large n.

By Theorem 2(i), under Assumptions 1–6, Condition M1 is satisfied when M_n is either fixed or diverges to infinity at a rate slower than m_n^{**}. In practice, we do not know m_n^{**}, so Condition M1 may happen. Condition M2 is met either when $M_n \geq m_n^{**}$ a.s. (as considered by Peng and Yang (2022)) or when $M_n < m_n^{**}$ a.s., but M_n has the same order as m_n^{**} almost surely. Next, we explore the degree of improvement $\Delta_n/R_n^{\mathrm{MS}}(m_n^*)$ by the following theorems under Conditions M1 and M2.

THEOREM 3 (Answer to Question Q2 under Condition M1). *Suppose that Assumptions 1–6 hold. Under Condition M1,*

$$\Delta_n = o\left\{R_n^{\mathrm{MS}}(m_n^*)\right\} \text{ a.s.}$$

Theorem 3 suggests that when the number of candidate models M_n is fixed or diverges to infinity at a slower rate than m_n^{**}, MA has no essential advantage over MS. This again indicates that in practical applications, a large enough number of candidate models should be utilized. It is important to note that the result in Theorem 3 holds for both the slowly and fast decaying $\{\theta_{n,m}\}_{m=1}^{d_n}$, as defined in Conditions A1–A2 below.

Furthermore, when Condition M2 holds, we explore the degree of improvement $\Delta_n/R_n^{\mathrm{MS}}(m_n^*)$ by the following theorem under sensible conditions on $\theta_{n,m}$, which provides an answer to Question Q2 under Condition M2. The answer depends on the decaying order of $\{\theta_{n,m}\}_{m=1}^{d_n}$.

Condition A1 (Slowly Decaying $\{\theta_{n,m}\}_{m=1}^{d_n}$). There exist constants $k > 1$, $0 < \delta \leq \eta < 1$ with $k\eta < 1$, and $K > 0$ such that for every integer sequence $\{l_n\}$ satisfying $\lim_{n\to\infty} l_n = \infty$,

$$\delta \leq \theta_{n,\lfloor kl_n \rfloor}/\theta_{n,l_n} \leq \eta \text{ for any } n \geq K \text{ a.s.}$$

Condition A2 (Fast Decaying $\{\theta_{n,m}\}_{m=1}^{d_n}$). For every constant $k > 1$ and every integer sequence $\{l_n\}$ satisfying $\lim_{n\to\infty} l_n = \infty$,

$$\lim_{n\to\infty} \theta_{n,\lfloor kl_n \rfloor}/\theta_{n,l_n} = 0 \text{ a.s.}$$

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

In the model setting of Peng and Yang (2022), since $\theta_{n,m} = \beta_m^2/\sigma^2$, Conditions A1 and A2 are equivalent to the conditions 1 and 2 of Peng and Yang (2022), respectively.

THEOREM 4 (Answer to question Q2 under condition M2). *Suppose that Assumptions 1–6 and Condition M2 holds. Under Condition A1, we have*

$$\Delta_n \asymp R_n^{\mathrm{MS}}(m_n^*) \text{ a.s.}$$

Under Condition A2, we have

$$\Delta_n = o\{R_n^{\mathrm{MS}}(m_n^*)\} \text{ a.s.}$$

From Theorems 1 and 4, under Condition A1, we have

$$\frac{1}{2} \le \liminf_{n\to\infty} \frac{R_n^{\mathrm{MA}}(\mathbf{w}_n^*)}{R_n^{\mathrm{MS}}(m_n^*)} \le \limsup_{n\to\infty} \frac{R_n^{\mathrm{MA}}(\mathbf{w}_n^*)}{R_n^{\mathrm{MS}}(m_n^*)} \le 1 - c^* \text{ a.s.}$$

for some $c^* \in (0, 1/2]$; and under Condition A2, $R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \sim R_n^{\mathrm{MS}}(m_n^*)$ a.s. Therefore, when the number of candidate models M_n is large enough, there is a phase transition in the advantage of MA over MS. When $\theta_{n,m}$ decays slowly in m, the oracle MA can reduce the optimal risk of MS by a substantial fraction; however, when $\theta_{n,m}$ decays fast in m, MA has no real advantage over MS.

To gain a better understanding of Conditions A1 and A2, we consider the following more simple conditions (i.e., Assumption 7 and Conditions B1 and B2), which imply Conditions A1 and A2 by Lemma 1 below.

Assumption 7. There exists a nonstochastic positive sequence $\theta_m^*, m = 1, \ldots, d_n$ such that

$$\max_{1\le m\le d_n} \left| \frac{\theta_{n,m}}{\theta_m^*} - 1 \right| \to_{a.s.} 0.$$

Condition B1 (Slowly Decaying θ_m^*). There exist constants $k > 1$ and $0 < \delta^* \le \eta^* < 1$ with $k\eta^* < 1$ such that $\delta^* \le \theta_{\lfloor km \rfloor}^*/\theta_m^* \le \eta^*$ when m is large enough.

Condition B2 (Fast Decaying θ_m^*). For every constant $k > 1$, $\lim_{m\to\infty} \theta_{\lfloor km \rfloor}^*/\theta_m^* = 0$.

LEMMA 1. *Suppose that Assumption 7 holds. Then, Conditions B1 and B2 imply Conditions A1 and A2, respectively.*

Assumption 7 implies that $\lim_{n\to\infty} \theta_{n,m} = \theta_m^*$ for any fixed m a.s. Thus, Assumption 7 can lead to Assumption 5 by taking $\bar{\theta}_m = \theta_m^*/2$. Condition B1 is satisfied for $\theta_m^* \sim m^{-2\alpha}$ or slightly more generally for $\theta_m^* \sim m^{-2\alpha}(\log m)^\beta$ with constants $\alpha > 1/2$ and $\beta \in \mathbb{R}$. Condition B2 is satisfied for the exponential-decay case; that is, $\theta_m^* \sim \exp(-cm)$ for some $c > 0$. These two types of decaying rates are commonly seen in the literature. For example, in the research of infinite-order AR models, Ing and Wei (2005), Ing (2007), and Liao et al. (2021) considered the exponential-

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

decay and algebraic-decay cases for the AR coefficients, which are described in our context as follows:

(i) Exponential-decay case: $C_1 m^{-\tau_1} e^{-cm} \le \theta_m^* \le C_2 m^{\tau_1} e^{-cm}$, where C_1, C_2, τ_1, and c are constants with $C_2 \ge C_1 > 0$, $\tau_1 \ge 0$, and $c > 0$.
(ii) Algebraic-decay case: $(C_3 - C_4 m^{-\tau_2}) m^{-\bar{\alpha}} \le \theta_m^* \le (C_3 + C_4 m^{-\tau_2}) m^{-\bar{\alpha}}$, where C_3, C_4, τ_2, and $\bar{\alpha} > 1$ are positive constants.

It can be easily verified that the exponential-decay case (i) and algebraic-decay case (ii) satisfy Conditions B2 and B1, respectively.

3.3. A Comparison of Two Specific MS and MA Procedures

Up to now, the theoretical results of Theorems 1, 3 and 4 mainly focus on the comparison of oracle optimal MS and MA, not directly on the comparison of two specific MS and MA procedures. Fortunately, by using (2) and (3), we can do the latter comparison by connecting the feasible risks (when using a selected model index or weights from some methods) and infeasible risks (when using the oracle model index or weights). In the literature, the proof of the asymptotic efficiency (or optimality) of MS and MA requires the smallest risks of MS and MA (i.e., $R_n^{\mathrm{MS}}(m_n^*)$ and $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ in our notation) to grow to infinity as the sample size increases. Both growth results have been verified in Theorem 2(ii) under Assumptions 1–6.

Let $\hat{m}_n$ and $\hat{\mathbf{w}}_n$ be the selected model index and chosen weights based on asymptotically optimal MS and MA methods, respectively. Then, we have the following two consequences.

COROLLARY 1. *Suppose that Assumptions 1–6 hold, and $\hat{m}_n$ and $\hat{\mathbf{w}}_n$ are asymptotically optimal in the sense of (2), i.e., $R_n^{\mathrm{MS}}(\hat{m}_n)/R_n^{\mathrm{MS}}(m_n^*) \to_p 1$ and $R_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n)/R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \to_p 1$. Then,*

(i) *the risks using $\hat{m}_n$ and $\hat{\mathbf{w}}_n$ have the same order, i.e., $R_n^{\mathrm{MS}}(\hat{m}_n) \asymp_p R_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n)$;*
(ii) *under Conditions M2 and A1, MA using $\hat{\mathbf{w}}_n$ essentially improves over MS using $\hat{m}_n$, i.e., $R_n^{\mathrm{MS}}(\hat{m}_n) - R_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n) \asymp_p R_n^{\mathrm{MS}}(\hat{m}_n)$;*
(iii) *under either Condition M1 or Conditions M2 and A2, $\hat{m}_n$ and $\hat{\mathbf{w}}_n$ are asymptotically equivalent in risk, i.e., $R_n^{\mathrm{MS}}(\hat{m}_n) \sim_p R_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n)$.*

COROLLARY 2. *Suppose that Assumptions 1–6 hold, and $\hat{m}_n$ and $\hat{\mathbf{w}}_n$ are asymptotically optimal in the sense of (3), i.e., $E\{L_n^{\mathrm{MS}}(\hat{m}_n)\}/R_n^{\mathrm{MS}}(m_n^*) \to_{a.s.} 1$ and $E\{L_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n)\}/R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \to_{a.s.} 1$. Then, the results of Corollary 1 hold a.s. when $R_n^{\mathrm{MS}}(\hat{m}_n)$, $R_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n)$, $\asymp_p$, and $\sim_p$ are replaced by $E\{L_n^{\mathrm{MS}}(\hat{m}_n)\}$, $E\{L_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n)\}$, $\asymp$, and $\sim$, respectively.*

Since the proofs of Corollaries 1 and 2 are similar, we provide only the proof of Corollary 1, which appears in Section S.1.5 of the Supplementary Material. Alhorn et al. (2019); (2021) also contributed to the discussion of the superiority of MA over MS, but they used fixed weights in their theoretical studies and other

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

weights with explicit forms in their numerical studies; these weights supply no support to prove asymptotic optimality.

4. COMPARISONS OF MAS WITH DIFFERENT WEIGHT SETS

This section compares the optimal risks of MAs when the weights belong to three weight sets: $\mathcal{W}_n$, $\mathcal{Q}_n$, and $\mathcal{W}_n(N)$. These comparisons provide answers to Questions Q3 and Q4.

4.1. A Comparison of MAs with Weight Sets $\mathcal{W}_n$ and $\mathcal{Q}_n$

In this subsection, our focus is on comparing the risks of MA estimators when the weights come from $\mathcal{W}_n$ and $\mathcal{Q}_n$. We first present the following theorem, which answers Question Q3.

THEOREM 5 (Answer to Question Q3). *Suppose that Assumptions 1–6 hold. Then,*

$$R_n^{\mathrm{MA}}(\mathbf{w}_n^*) - R_n^{\mathrm{MA}}(\tilde{\mathbf{w}}_n^*) = o\left\{R_n^{\mathrm{MA}}(\mathbf{w}_n^*)\right\} \ a.s., \tag{10}$$

i.e., $\mathbf{w}_n^*$ *and* $\tilde{\mathbf{w}}_n^*$ *are asymptotically equivalent in risk.*

Equation (10) indicates that while the weight relaxation could lead to a smaller optimal risk of MA, it does not provide any substantial benefit asymptotically. Note that Theorem 5 does not require any assumptions about the number of candidate models M_n or the decaying order of $\{\theta_{n,m}\}_{m=1}^{d_n}$.

Furthermore, we compare two specific asymptotically optimal MS and MA procedures, where MA weights are chosen from the weight set $\mathcal{Q}_n$. Let $\hat{\mathbf{w}}_n^Q$ denote the chosen weights based on a specific MA method that satisfies the asymptotic optimality (2) or (3), without imposing the total weight constraint $\sum_{m=1}^{M_n} w_m = 1$. For example, the asymptotic optimality (2) of JMA without the total weight constraint has been established by Ando and Li (2014) and Zhao et al. (2016) for independent data and dependent data, respectively. Using Theorem 5, we can easily derive the following corollary, which compares MA without the total weight constraint with MS.

COROLLARY 3. *Suppose that Assumptions 1–6 hold.*

(i) *Assume that* $\hat{m}_n$ *and* $\hat{\mathbf{w}}_n^Q$ *satisfy* $R_n^{\mathrm{MS}}(\hat{m}_n)/R_n^{\mathrm{MS}}(m_n^*) \to_p 1$ *and* $R_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n^Q)/R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \to_p 1$. *Then the results of Corollary 1 hold when* $\hat{\mathbf{w}}_n$ *is replaced by* $\hat{\mathbf{w}}_n^Q$.

(ii) *Assume that* $\hat{m}_n$ *and* $\hat{\mathbf{w}}_n^Q$ *satisfy* $E\{L_n^{\mathrm{MS}}(\hat{m}_n)\}/R_n^{\mathrm{MS}}(m_n^*) \to_{a.s.} 1$ *and* $E\{L_n^{\mathrm{MA}}(\hat{\mathbf{w}}_n^Q)\}/R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \to_{a.s.} 1$. *Then the results of Corollary 2 hold when* $\hat{\mathbf{w}}_n$ *is replaced by* $\hat{\mathbf{w}}_n^Q$.

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

4.2. A Comparison of MAs with Weight Sets $\mathcal{W}_n$ and $\mathcal{W}_n(N)$

In this subsection, we focus on the comparison of the optimal risks of MA estimators with weights belonging to the weight sets $\mathcal{W}_n$ and $\mathcal{W}_n(N)$, respectively. We present the following theorem on an upper bound of $R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) - R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ and an answer to Question Q4.

THEOREM 6 (Answer to Question Q4). *Suppose that Assumptions 1–6 hold. Then, for any sufficiently large n,*

$$R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) - R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \le \frac{1}{2N} R_n^{\mathrm{MS}}(m_n^*) \ a.s. \tag{11}$$

Furthermore, under Conditions M2 and A1, we have

$$R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) - R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \asymp R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \ a.s.;$$

and under either Condition M1 or Conditions M2 and A2, we have

$$R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) - R_n^{\mathrm{MA}}(\mathbf{w}_n^*) = o\left\{R_n^{\mathrm{MA}}(\mathbf{w}_n^*)\right\} \ a.s.,$$

i.e.,

$$R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \sim R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \ a.s.$$

Observing that $R_n^{\mathrm{MA}}(\mathbf{w}_{n,1}^*) = R_n^{\mathrm{MS}}(m_n^*)$, we can note that Theorems 1 and 3–4 are special cases of Theorem 6 with $N = 1$. The upper bound in (11) implies that for a fixed and sufficiently large sample size, $R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*)$ can be made arbitrarily close to $R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ when N is large enough. This result is expected since $\mathcal{W}_n(N)$ approaches $\mathcal{W}_n$ as closely as desired by making N sufficiently large. From Theorem 6, when M_n is large enough and $\theta_{n,m}$ decays slowly in m, restricting the weight set $\mathcal{W}_n$ to $\mathcal{W}_n(N)$ can enlarge the optimal risk of MA by a substantial multiple. When either M_n is too small or M_n is large enough and $\theta_{n,m}$ decays fast in m, MA restricted to the discrete weight set has no real disadvantage over MA with $\mathcal{W}_n$. In Section S.3 of the Supplementary Material, we provide a further comparison of MA techniques with nested discrete weight sets, including a comparison of MS and MA with a discrete weight set.

5. TWO EXAMPLES

In this section, we provide two examples to validate the theoretical results in Theorems 3, 4 and 6. Detailed derivations can be found in Section S.2 of the Supplementary Material. Given $a, b > 0$, the incomplete beta function is defined as $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}\,dt$ for $0 \le x \le 1$.

In both examples, we follow the model setting of Peng and Yang (2022): the model (1) with the orthonormal design assumption (4), homoscedastic and uncorrelated error terms, and $\nu_m = m$ for $m = 1, \ldots, p_n$. Under this configuration, $\theta_{n,m} = \beta_m^2/\sigma^2$.

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

Example 5.1 (Slowly decaying $\theta_{n,m}$). The true coefficients are set to $\beta_m = m^{-\alpha}$ with $\alpha > 1/2$. Then, $\theta_{n,m} = m^{-2\alpha}/\sigma^2$ and Condition A1 is satisfied. By some simple calculations, we can get $m_n^{**} \sim (\frac{n}{\sigma^2})^{\frac{1}{2\alpha}}$. We consider the following two situations regarding the number of candidate models M_n.

(i) When $\lim_{n\to\infty} M_n/m_n^{**} = 0$, we have

$$\frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \sim \frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \sim \begin{cases} \sum_{m=M+1}^{\infty} m^{-2\alpha}, & \text{if } M_n \equiv M \text{ is fixed as } n \to \infty, \\ \frac{M_n^{-2\alpha+1}}{2\alpha-1}, & \text{if } \lim_{n\to\infty} M_n = \infty \text{ but } M_n = o(m_n^{**}). \end{cases}$$

This verifies that $R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \sim R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ under Condition M1, which accords with Theorem 3 and the third conclusion of Theorem 6.

(ii) When $M_n/m_n^{**} \geq \underline{c}$ for some $\underline{c} > 0$, we have

$$\frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \asymp \frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \asymp n^{-\frac{2\alpha-1}{2\alpha}}.$$

By a simple calculation, we know that $R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) - R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ is lower bounded by $\frac{\varpi\sigma^2}{2^{2\alpha+1}\varpi^{-2\alpha+2}}\left(\frac{n}{\sigma^2}\right)^{\frac{1}{2\alpha}}$, where $\varpi = \min\{\underline{c}, (2N-1)^{-\frac{1}{2\alpha}}\}$. Moreover, if $\lim_{n\to\infty} M_n/m_n^{**} = \kappa$, $\kappa \in (0,\infty]$ and $M_n = o(p_n)$, we have

$$\lim_{n\to\infty} \frac{R_n^{\mathrm{MA}}(\mathbf{w}_n^*)}{R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*)} = \frac{1}{\psi_N^* + \frac{\kappa^{-2\alpha+1}}{2\alpha}} \left[\frac{2\alpha-1}{4\alpha^2}\left\{\frac{\pi}{\sin(\frac{\pi}{2\alpha})} - B\left(\frac{1}{1+\kappa^{2\alpha}}; 1-\frac{1}{2\alpha}, \frac{1}{2\alpha}\right)\right\} + \frac{\kappa^{-2\alpha+1}}{2\alpha}\right],$$

where ψ_N^* is defined in (S.6) of the Supplementary Material. Furthermore, we can show that

$$\lim_{n\to\infty} \frac{R_n^{\mathrm{MA}}(\mathbf{w}_n^*)}{R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*)} < 1,$$

which verifies that $R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) - R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \asymp R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$, which accords with the first conclusion of Theorem 4 and the second conclusion of Theorem 6. Figure 2(a) plots $\lim_{n\to\infty} R_n^{\mathrm{MA}}(\mathbf{w}_n^*)/R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*)$ against $N \in \{1,\ldots,10\}$ for $\kappa = 0.5, 1, 2$, and Figure 2(b) plots $\lim_{n\to\infty} R_n^{\mathrm{MA}}(\mathbf{w}_n^*)/R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*)$ against $\kappa \in (0,8)$ for $N = 1, 2, 4$, where $\alpha = 0.8$, which further verifies that $\lim_{n\to\infty} R_n^{\mathrm{MA}}(\mathbf{w}_n^*)/R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) < 1$.

Example 5.2 (Fast decaying $\theta_{n,m}$). The true coefficients are set to $\beta_m = \exp(-cm)$ with $c > 0$. Then, $\theta_{n,m} = \exp(-2cm)/\sigma^2$ and Condition A2 is satisfied. The global optimal model should include the first $m_n^{**} \sim \frac{1}{2c}\log\left(\frac{n}{\sigma^2}\right)$ terms. We consider the following three situations regarding the number of candidate models M_n.

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

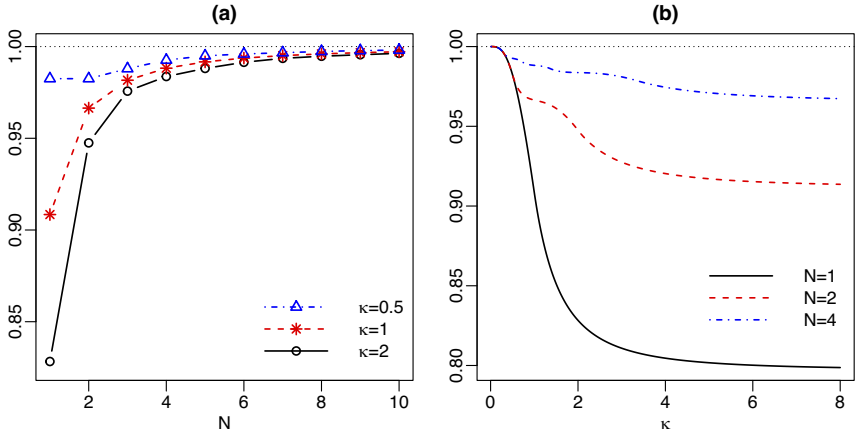

FIGURE 2. Numerical illustration for Example 5.1 with $\alpha = 0.8$. (a): Plots of $\lim_{n\to\infty} R_n^{\mathrm{MA}}(\mathbf{w}_n^*)/R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*)$ against $N \in \{1,\ldots,10\}$ for $\kappa = 0.5, 1, 2$, respectively. (b): Plots of $\lim_{n\to\infty} R_n^{\mathrm{MA}}(\mathbf{w}_n^*)/R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*)$ against $\kappa \in (0, 8)$ for $N = 1, 2, 4$, respectively.

(i) When $\limsup_{n\to\infty} M_n/m_n^{**} < 1$, we have

$$\frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \sim \frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \sim \frac{\exp(-2cM_n)}{\exp(2c)-1},$$

which verifies that $R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \sim R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ under Condition M1 and accords with Theorem 3 and the third conclusion of Theorem 6.

(ii) When $M_n < m_n^{**}$ for any sufficiently large n but $\lim_{n\to\infty} M_n/m_n^{**} = 1$, we have

$$\frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \sim \frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \sim \frac{1}{2c}\frac{\sigma^2}{n}\log\left(\frac{n}{\sigma^2}\right) + \frac{\exp(-2cM_n)-\exp(-2cp_n)}{\exp(2c)-1},$$

which verifies that $R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \sim R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ under Conditions M2 and A2 and accords with the second conclusion of Theorem 4 and the third conclusion of Theorem 6.

(iii) When $M_n \geq m_n^{**}$ for any sufficiently large n, we have

$$\frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \sim \frac{1}{n}R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \sim \frac{1}{2c}\frac{\sigma^2}{n}\log\left(\frac{n}{\sigma^2}\right),$$

which also verifies that $R_n^{\mathrm{MA}}(\mathbf{w}_{n,N}^*) \sim R_n^{\mathrm{MA}}(\mathbf{w}_n^*)$ under Conditions M2 and A2 and accords with the second conclusion of Theorem 4 and the third conclusion of Theorem 6.

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

6. SIMULATION STUDIES

In this section, we conduct several simulation studies to illustrate the theoretical results presented in Corollaries 1 and 3, where specific MS and MA methods are compared. We choose AIC, BIC, and LOO-CV as MS methods and MMA and JMA as MA methods. Additionally, we conduct simulation studies to compare the oracle optimal MAs with different weight sets, and to illustrate the theoretical results in Theorems 5 and 6. Specifically, we use the following three examples:

- **Example 1**: **General nested framework** (i.e., $v_m \neq m$), homoscedastic and uncorrelated errors, and (approximately) orthonormal design.
- **Example 2**: Typical nested framework (i.e., $v_m = m$), **heteroscedastic and autocorrelated errors**, and (approximately) orthonormal design.
- **Example 3**: Typical nested framework (i.e., $v_m = m$), homoscedastic and uncorrelated errors, and **non-orthonormal design**.

To evaluate the estimators, we compute the risks of the competing methods by computing averages across 1000 replications.

Example 1 (General nested framework). We use the same set-up as that of Peng and Yang (2022) except for the coefficients β_m's. Specifically, suppose the data come from the model (1), where $p_n = \lfloor 5n^{2/3} \rfloor$, $x_{i1} = 1$, the remaining x_{ij} are independent and identically distributed (i.i.d.) from $N(0,1)$, and the random errors ε_i are i.i.d. from $N(0,\sigma^2)$ and are independent of x_{ij}'s. The population R^2 is denoted by $R^2 = \mathrm{Var}(\mu_i)/\mathrm{Var}(y_i)$, which is controlled to be 0.05, 0.25, 0.5, or 0.75 via the parameter σ^2. We consider a more general nested model setting than that of Peng and Yang (2022) by setting

$$v_m = \begin{cases} 5\lfloor m/2 \rfloor + 2, & \text{if } m \text{ is odd} \\ 5\lfloor m/2 \rfloor, & \text{if } m \text{ is even} \end{cases}, \quad m = 1, \ldots, q_n - 1$$

and $v_{q_n} = p_n$. Thus, the size of the mth group of predictors is 2 when m is odd and 3 when m is even, $m = 1, \ldots, q_n - 1$. We consider two cases with different coefficient decaying orders:

- Case 1. $\beta_j = m^{-\alpha_1}$ when x_{ij} is in the mth group, and α_1 is set to be 1, 1.5, or 2.
- Case 2. $\beta_j = \exp(-\alpha_2 m)$ when x_{ij} is in the mth group, and α_2 is set to be 1, 1.5, or 2.

For Case 1, we know that $\theta_{n,m}$ converges to $\theta_m^* = m^{-2\alpha_1}/\sigma^2$ a.s., and then Condition B1 is satisfied. For Case 2, $\theta_{n,m}$ converges to $\theta_m^* = \exp(-2\alpha_2 m)/\sigma^2$ a.s., and then Condition B2 is satisfied. The sample size n varies at 50, 500, 1000, 2000, 3000, and 4000. The number of candidate models is determined by $M_n = \mathrm{INT}(3n^{1/3})$, where the function $\mathrm{INT}(a)$ returns the nearest integer from a. In each simulation setting of the combination of n, R^2, and α_1 (or α_2), we normalize the risks of the MS methods by dividing by the risk of MMA.

Figure 3 presents the simulation results for Case 1 with $\alpha_1 = 1.5$ and Case 2 with $\alpha_2 = 1.5$. Due to limited space, the other simulation results are summarized in

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

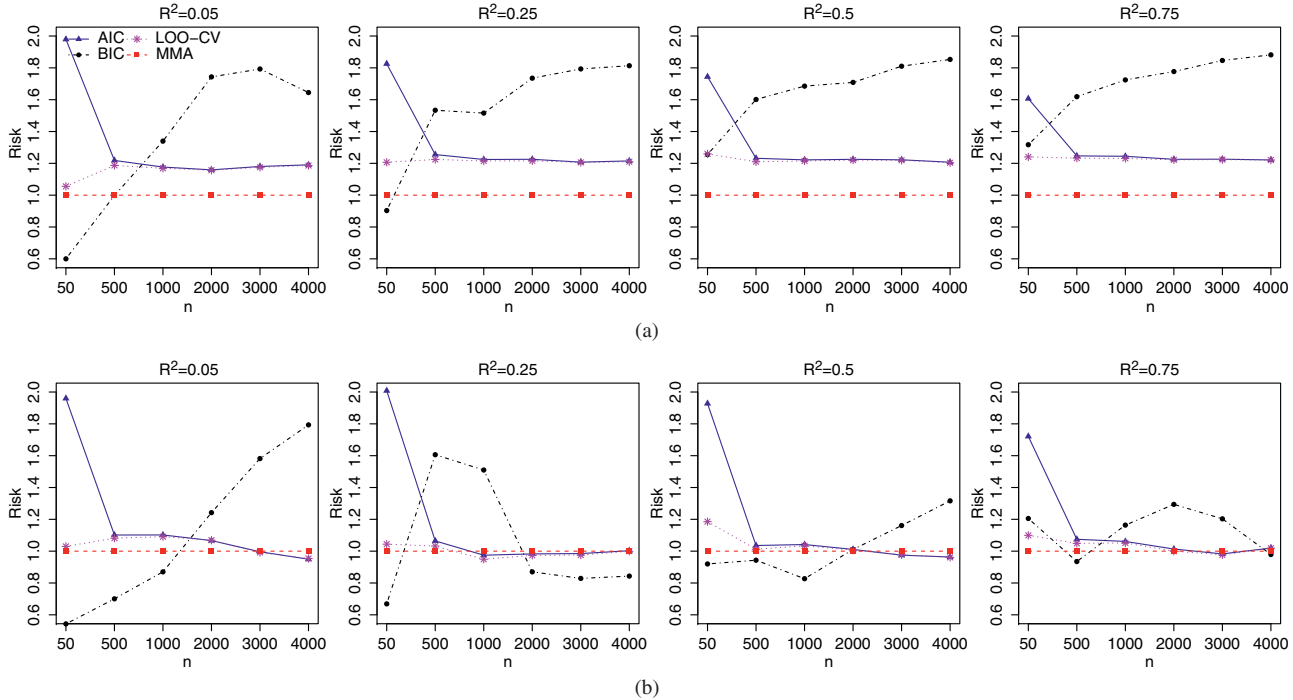

FIGURE 3. Normalized risk functions for AIC, BIC, LOO-CV, and MMA in Example 1. Row (a): $\theta_m^* = m^{-2\alpha_1}/\sigma^2$ with $\alpha_1 = 1.5$, corresponding to the case of slowly decaying θ_m^*. Row (b): $\theta_m^* = \exp(-2\alpha_2 m)/\sigma^2$ with $\alpha_2 = 1.5$, corresponding to the case of fast decaying θ_m^*.

Figures S.2 and S.3 of the Supplementary Material. Note that Figure 3(a) refers to Case 1 (slowly decaying coefficients) and (b) to Case 2 (fast decaying coefficients). Since both AIC and LOO-CV are asymptotically optimal for Example 1, as expected, their performances are very close for large sample sizes. In the slowly decaying θ_m^* case, the performance gap between AIC (or LOO-CV) and MMA does not vanish when n increases, while in the fast decaying θ_m^* case, it becomes very small when n is large. These results are consistent with the results of Corollary 1.

Following Peng and Yang (2022), we also include BIC in our simulation, although often, BIC is not asymptotically optimal. In Case 1, the advantage of AIC over BIC becomes increasingly larger as n increases from 50 to 4000, while in Case 2 with fast decaying θ_m^*, BIC is competitive with AIC in some scenarios. This phenomenon was also observed by Peng and Yang (2022).

Example 2 (Heteroskedastic and autocorrelated errors). The setting of this example is the same as that of Example 1, except that the typical nested framework with $\nu_m = m$ and heteroscedastic and autocorrelated errors are considered. We utilize the same error process of Zhang et al. (2013), one that is both heteroscedastic and autocorrelated. Specifically, the error process is given by $\varepsilon_i = \varepsilon_{i1} + \varepsilon_{i2}$, where the ε_{i1}'s are independent observations from the $N(0, x_{i2}^2)$ distribution, and ε_{i2} follows an AR(1) process with an autocorrelation coefficient $\rho_1 = 0.5$, where $\varepsilon_{i2} = \rho_1 \varepsilon_{i-1,2} + e_i$, $\varepsilon_{12} \sim N(0,1)$, and the e_i's are i.i.d. from $N(0, 1-\rho_1^2)$ and are independent of ε_{i2}'s. Then, the conditional covariance matrix of $\boldsymbol{\varepsilon}$ given the x_{ij}'s is $\boldsymbol{\Omega} = \boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2$, where $\boldsymbol{\Omega}_1 = \mathrm{diag}\{x_{12}^2, \ldots, x_{n2}^2\}$ and $\boldsymbol{\Omega}_2 = (\rho_1^{|k-l|})_{k,l=1,\ldots,n}$. By a simple calculation, we have

$$\mathrm{tr}(\mathbf{P}_m \boldsymbol{\Omega}_1) \to_{a.s.} \begin{cases} 1, & \text{if } \nu_m = 1 \\ \nu_m + 2, & \text{if } \nu_m \geq 2 \end{cases} \quad \text{and} \quad \mathrm{tr}(\mathbf{P}_m \boldsymbol{\Omega}_2) \to_{a.s.} \frac{2\rho_1}{1-\rho_1} + \nu_m.$$

Therefore, for any fixed m, $\theta_{n,m} \to_{a.s.} \theta_m^* = \beta_m^2/\zeta_m$, where

$$\zeta_m = \begin{cases} \frac{2}{1-\rho_1}, & \text{if } m = 1, \\ 4, & \text{if } m = 2, \\ 2, & \text{if } m \geq 3. \end{cases}$$

We consider two cases with different decaying orders of θ_m^*:

- Case 1 (With θ_m^* satisfying Condition B1). Here, $\beta_m = c\sqrt{\zeta_m} m^{-\alpha_1}$, and α_1 is set to be 1, 1.5, or 2.
- Case 2 (With θ_m^* satisfying Condition B2). Here, $\beta_m = c\sqrt{\zeta_m} \exp(-\alpha_2 m)$, and α_2 is set to be 1, 1.5, or 2.

As in Hansen and Racine (2012), the parameter c is selected to control the approximate population $\tilde{R}^2 = c^2/(1+c^2)$ to vary on 0.05, 0.25, 0.5, and 0.75. The sample size is varied among $n = 500$, 1000, 2000, 3000, 4000, and 5000. To illustrate the results in Corollary 3, we include JMA without the restriction $\sum_{m=1}^{M_n} w_m = 1$, denoted by JMA2, as a competing method. In each simulation

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

setting of combination of n, $\tilde{R}^2$, and α_1 (or α_2), we normalize the risks of the MS methods and JMA2 by dividing by the risk of JMA.

Figure 4 only shows the simulation results for Case 1 with $\alpha_1 = 1.5$ and Case 2 with $\alpha_2 = 1.5$. Due to limited space, the other simulation results are summarized in Figures S.4 and S.5 of the Supplementary Material. From Figure 4(a), we see that in the slowly decaying θ_m^* case, the performance gap between LOO-CV and JMA does not vanish when the sample size increases. In contrast, in the fast decaying θ_m^* case (Figure 4(b)), it becomes very small when the sample size is large. These results are consistent with Corollary 1. Note that from Figure 4(b), the performances of AIC and JMA are not consistently close since AIC may not be asymptotically optimal in Example 2 because of heteroscedasticity. Another observation is that the performances of JMA2 and JMA are very close when n is sufficiently large, which illustrates the results in Corollary 3. Moreover, we can observe the same phenomena seen in the Example 1 for the comparison of AIC and BIC.

Example 3 (Non-orthonormal design). The setting of this example is the same as that of Example 1 except that the typical nested framework with $\nu_m = m$ and predictors are non-orthonormal. Specifically, the predictors $(x_{i1}, \ldots, x_{ip_n})^\top$, $i = 1, \ldots, n$ are i.i.d. normal random vectors with zero mean and the covariance matrix between the kth and lth elements being $\rho_2^{|k-l|}$, with the random errors ε_i being i.i.d. from $N(0, \sigma^2)$ and independent of x_{ij}'s. Here, ρ_2 is set to be 0.5. It is easy to prove that for any fixed m,

$$\frac{1}{n}\boldsymbol{\mu}^\top \mathbf{P}_m \boldsymbol{\mu} \to_{a.s.} \lim_{n\to\infty} \boldsymbol{\beta}_{p_n}^\top \boldsymbol{\Sigma}_{m\times p_n}^\top \boldsymbol{\Sigma}_{m\times m}^{-1} \boldsymbol{\Sigma}_{m\times p_n} \boldsymbol{\beta}_{p_n},$$

where $\boldsymbol{\Sigma}_{d_1\times d_2}$ is a $d_1 \times d_2$ matrix with (k,l)th element being $\rho_2^{|k-l|}$ and $\boldsymbol{\beta}_{p_n} = (\beta_1, \ldots, \beta_{p_n})^\top$. It follows that $\theta_{n,m} = \boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}/(n\sigma^2) \to \theta_m^* = \xi_m/\sigma^2$ a.s., where $\xi_1 = \lim_{n\to\infty}(\boldsymbol{\Sigma}_{1\times p_n}\boldsymbol{\beta}_{p_n})^2$ and

$$\xi_m = \lim_{n\to\infty} \boldsymbol{\beta}_{p_n}^\top \left\{ \boldsymbol{\Sigma}_{m\times p_n}^\top \boldsymbol{\Sigma}_{m\times m}^{-1} \boldsymbol{\Sigma}_{m\times p_n} - \boldsymbol{\Sigma}_{(m-1)\times p_n}^\top \boldsymbol{\Sigma}_{(m-1)\times(m-1)}^{-1} \boldsymbol{\Sigma}_{(m-1)\times p_n} \right\} \boldsymbol{\beta}_{p_n}$$

for $m = 2, \ldots, p_n$. By calculations, we can obtain simple forms of ξ_m as follows

$$\xi_1 = \lim_{n\to\infty} \left(\sum_{j=1}^{p_n} \beta_j \rho_2^{j-1} \right)^2 \quad \text{and} \quad \xi_m = \lim_{n\to\infty} (1-\rho_2^2) \left(\sum_{j=m}^{p_n} \beta_j \rho_2^{j-m} \right)^2, \ m \geq 2. \tag{12}$$

We consider two cases with different decaying orders of θ_m^*:

- Case 1 (With θ_m^* satisfying Condition B1). Here, $\xi_m = m^{-2\alpha_1}$ and α_1 is set to be 1, 1.5, or 2.
- Case 2 (With θ_m^* satisfying Condition B2). Here, $\xi_m = \exp(-2\alpha_2 m)$ and α_2 is set to be 1, 1.5, or 2.

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

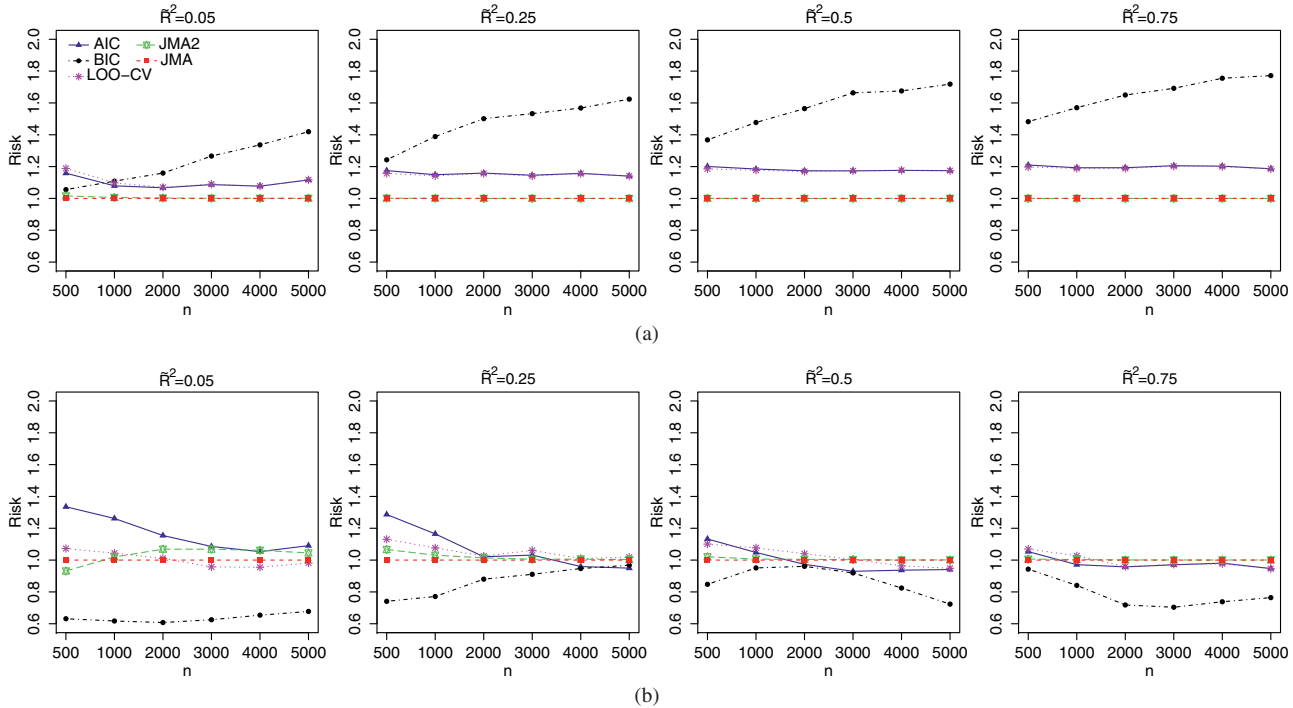

FIGURE 4. Normalized risk functions for AIC, BIC, LOO-CV, JMA2, and JMA in Example 2. Row (a): $\theta_m^* = c^2 m^{-2\alpha_1}$ with $\alpha_1 = 1.5$, corresponding to the case of slowly decaying θ_m^*. Row (b): $\theta_m^* = c^2 \exp(-2\alpha_2 m)$ with $\alpha_2 = 1.5$, corresponding to the case of fast decaying θ_m^*.

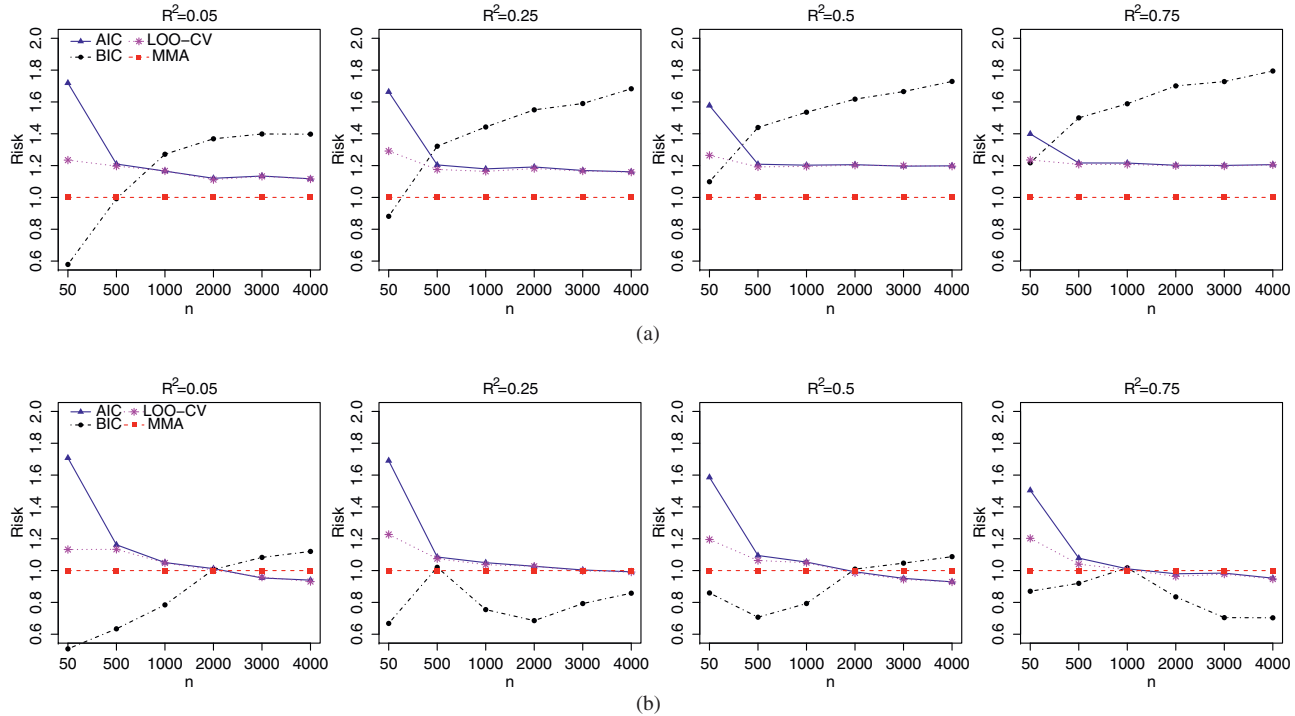

FIGURE 5. Normalized risk functions for AIC, BIC, LOO-CV, and MMA in Example 3. Row (a): $\theta_m^* = m^{-2\alpha_1}/\sigma^2$ with $\alpha_1 = 1.5$, corresponding to the case of slowly decaying θ_m^*. Row (b): $\theta_m^* = \exp(-2\alpha_2 m)/\sigma^2$ with $\alpha_2 = 1.5$, corresponding to the case of fast decaying θ_m^*.

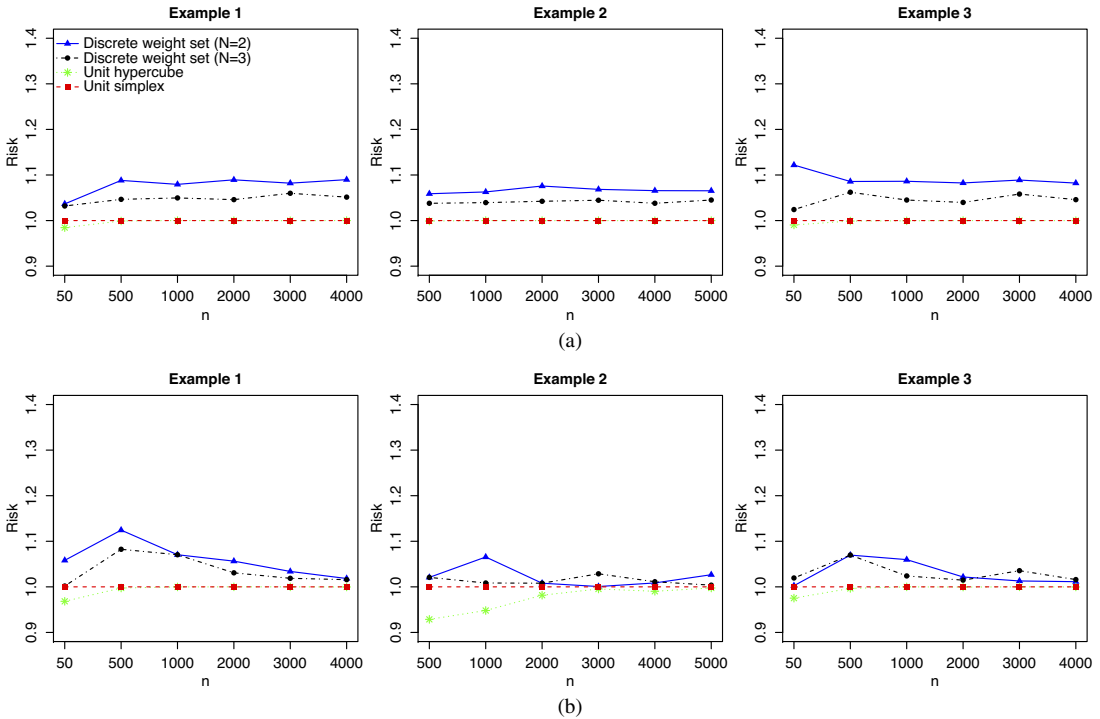

FIGURE 6. Normalized risk functions of the oracle optimal MAs with different weight sets for Examples 1–3, where $R^2 = 0.5$ (or $\tilde{R}^2 = 0.5$). Row (a): Case 1 with $\alpha_1 = 2$, corresponding to the case of slowly decaying GVI. Row (b): Case 2 with $\alpha_2 = 2$, corresponding to the case of fast decaying GVI.

We can set different coefficient β_j via (12) such that Case 1 and Case 2 hold, respectively. Without loss of generality, we assume $\beta_j \geq 0$ for all j. Then from (12), we have

$$\beta_1 = \sqrt{\xi_1} - \rho_2 \sqrt{\frac{\xi_2}{1-\rho_2^2}} \quad \text{and} \quad \beta_j = \frac{\sqrt{\xi_j} - \rho_2 \sqrt{\xi_{j+1}}}{\sqrt{1-\rho_2^2}}, \quad j \geq 2.$$

The sample size n varies at 50, 500, 1000, 2000, 3000, and 4000. In each simulation setting of the combination of n, R^2, and α_1 (or α_2), we normalize the risks of the MS methods by dividing by the risk of MMA. Figure 5 displays the simulation results for Case 1 with $\alpha_1 = 1.5$ and Case 2 with $\alpha_2 = 1.5$. Due to limited space, the other simulation results are summarized in Figures S.6 and S.7 of the Supplementary Material. From these results, we can see the same observations as in Example 1, again agreeing with the previous theoretical findings.

Finally, we illustrate the theoretical results presented in Theorems 5 and 6. We consider the same settings as in Examples 1–3 and fix $R^2 = 0.5$ (or $\tilde{R}^2 = 0.5$) and $\alpha_1 = 2$ (or $\alpha_2 = 2$). For each model setting, we compute the risks of the oracle optimal MAs with three different weight sets $\mathcal{W}_n(N)$, $\mathcal{W}_n$, and $\mathcal{Q}_n$, where $N = 2, 3$. Then, we normalize all risks by dividing by the risk of the oracle optimal MA with weights belonging to $\mathcal{W}_n$. Figure 6 shows the simulation results for Examples 1–3, where (a) refers to Case 1 (slowly decaying GVI) and (b) to Case 2 (fast decaying GVI). From Figure 6(a), the performance gap between MAs with weight sets $\mathcal{W}_n(N)$ and $\mathcal{W}_n$ does not vanish as n increases, while it becomes very small as n increases in Figure 6(b). This is consistent with the results in Theorem 6. Additionally, in both (a) and (b), the performance of MAs with weight sets $\mathcal{W}_n$ and $\mathcal{Q}_n$ are very close when n is large. This is consistent with the results in Theorem 5.

In the examples above, we utilize the decaying orders of GVI to determine the performance of MAs and MSs and set M_n to be sufficiently large (i.e., Condition M2 is satisfied). Additionally, in Section S.5 of the Supplementary Material, we design Example 4 to illustrate Corollary 1 when M_n is too small (Condition M1 is satisfied).

7. CONCLUSION

This paper compares MS and MA in a general model setting, allowing the predictors to be non-orthonormal, the error terms to be heteroscedastic and autocorrelated, and some predictors to be totally unimportant. We obtain the results that the number of candidate models M_n and the decaying order of $\{\theta_{n,m}\}_{m=1}^{d_n}$ determine when MA is better than MS. Specifically, when M_n is large enough and $\theta_{n,m}$ decays slowly in m, the benefit of MA over MS is real. However, when either M_n is too small or M_n is large enough and $\theta_{n,m}$ decays fast in m, the risks of MA and MS are asymptotically equivalent. Furthermore, the obtained results are extended to compare MAs with weights belonging to three different weight sets.

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

The results of this paper provide practical insights. First, MA has the potential to outperform MS significantly in risk. The comparison between them depends on both the number of candidate models and the decaying order of a sequence of indices related to the model coefficients. Second, to achieve optimal performance for MS and MA, the number of candidate models is required to diverge to infinity with the sample size. Last, a large weight set may not improve the performance of MA. Under certain reasonable conditions, MA with the large weight set $\mathcal{Q}_n$ has the same performance as MA with $\mathcal{W}_n$ in a large sample sense.

The results of this paper suggest several open questions. First, an interesting issue is how to order the predictors and prepare nested candidate models such that the risk gain of MA is optimal. Although various procedures are proposed to order the predictors in the implementation of MA, such as the forward selection approach (Claeskens et al., 2006), marginal correlation (Ando and Li, 2014, 2017; Zhang et al., 2016b), and solution path algorithm of penalized regression (Feng and Liu, 2020; Zhang et al., 2020), the literature still lacks theoretical study on the optimal way of ordering the predictors. Second, it would be interesting to develop a data-driven way to choose the number of the candidate models. Third, it would be valuable to extend the current work to dynamic models that include lagged dependent variables as predictor variables.

Finally, we discuss the challenge of comparing MS and MA in the non-nested model setting. First, the monotonicity of the squared bias and variance terms of $\hat{\boldsymbol{\mu}}_m$, as illustrated in Remark 2, generally does not hold in the non-nested setting. Second, the property that $\{\mathbf{P}_m - \mathbf{P}_{m-1}\}_{m=1}^{M_n}$ are mutually orthonormal projection matrices, as detailed in Appendix, does not apply in the non-nested setting. These observations make it difficult to characterize the unknown optimal model index m_n^* and weights $\mathbf{w}_n^*$. Consequently, deriving explicit expressions for the optimal risks of MS and MA becomes challenging, introducing complexity to the comparison of MS and MA. A deeper and more detailed investigation of these issues is warranted.

APPENDIX

A. Proof of Theorem 1

Given that Assumptions 1–3 and 5–6, Conditions M1–M2, and Conditions A1–A2 are imposed almost surely, there exists an event $\mathcal{F}$ with $\Pr(\mathcal{F}) = 1$ such that these assumptions and conditions can be imposed surely on $\mathcal{F}$. In the subsequent proofs, all results will be derived on $\mathcal{F}$ when using these assumptions and conditions, hence they hold almost surely. The risk of the mth candidate model is

$$\begin{aligned} R_n^{\mathrm{MS}}(m) &= E\left\{\|\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 | \mathbf{X}\right\} \\ &= \mathrm{tr}\left[E\left\{(\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu})^\top | \mathbf{X}\right\}\right] \\ &= \mathrm{tr}\left[\{E(\hat{\boldsymbol{\mu}}_m|\mathbf{X}) - \boldsymbol{\mu}\}\{E(\hat{\boldsymbol{\mu}}_m|\mathbf{X}) - \boldsymbol{\mu}\}^\top\right] + \mathrm{tr}\left\{\mathrm{Var}(\hat{\boldsymbol{\mu}}_m|\mathbf{X})\right\} \\ &= \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_m)\boldsymbol{\mu} + \mathrm{tr}(\mathbf{P}_m\boldsymbol{\Omega}). \end{aligned}$$

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

Observe that

$$R_n^{\mathrm{MS}}(m) - R_n^{\mathrm{MS}}(m-1) = \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\} - \boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}$$
$$= n\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}\left(\frac{1}{n} - \theta_{n,m}\right), \tag{A.1}$$

where $\theta_{n,m}$ is defined in (6). Under Assumptions 2–3 we have $\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\} \geq c_1$ and $\{\theta_{n,m}\}_{m=1}^{q_n}$ is nonincreasing. Combining these results with Assumption 6, it is easy to see that the global optimal model m_n^{**} that minimizes $R_n^{\mathrm{MS}}(m)$ on $\{1, \ldots, d_n\}$ satisfies

$$\theta_{n,m_n^{**}} > \frac{1}{n} \geq \theta_{n,m_n^{**}+1}. \tag{A.2}$$

Hence, the risk of the optimal model m_n^* is

$$R_n^{\mathrm{MS}}(m_n^*) = \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_{m_n^*})\boldsymbol{\mu} + \mathrm{tr}(\mathbf{P}_{m_n^*}\boldsymbol{\Omega}), \tag{A.3}$$

where when $M_n < m_n^{**}$, we have $m_n^* = M_n$ and $\theta_{n,m_n^*} > 1/n$; when $M_n \geq m_n^{**}$, we have $m_n^* = m_n^{**}$ and

$$\theta_{n,m_n^*} > \frac{1}{n} \geq \theta_{n,m_n^*+1}. \tag{A.4}$$

The risk of the MA estimator with weights $\mathbf{w}$ is

$$R_n^{\mathrm{MA}}(\mathbf{w}) = E\left\{\|\hat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}\|^2|\mathbf{X}\right\} = E\left\{\|\mathbf{P}(\mathbf{w})\mathbf{y} - \boldsymbol{\mu}\|^2|\mathbf{X}\right\}$$
$$= \boldsymbol{\mu}^\top\{\mathbf{P}(\mathbf{w}) - \mathbf{I}_n\}^2\boldsymbol{\mu} + \mathrm{tr}\{\mathbf{P}^2(\mathbf{w})\boldsymbol{\Omega}\}.$$

Rewrite $\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M_n}\gamma_m(\mathbf{P}_m - \mathbf{P}_{m-1})$, where $\gamma_m = \sum_{j=m}^{M_n} w_j$ and $\mathbf{P}_0 = \mathbf{0}$. Since $\mathbf{P}_m\mathbf{P}_l = \mathbf{P}_{\min(m,l)}$ for the nested candidate models, it is easy to verify that $\{\mathbf{P}_m - \mathbf{P}_{m-1}\}_{m=1}^{M_n}$ are mutually orthonormal projection matrices, i.e.,

$$(\mathbf{P}_{m_1} - \mathbf{P}_{m_1-1})(\mathbf{P}_{m_2} - \mathbf{P}_{m_2-1}) = \begin{cases} \mathbf{P}_{m_1} - \mathbf{P}_{m_1-1}, & \text{if } m_1 = m_2, \\ \mathbf{0}, & \text{if } m_1 \neq m_2. \end{cases}$$

Using the above fact, $R_m(\mathbf{w})$ is further expanded as

$$R_n^{\mathrm{MA}}(\mathbf{w}) = \sum_{m=1}^{M_n}\Big(\gamma_m^2\Big[\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} + \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}\Big]$$
$$- 2\gamma_m\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} + \boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}\Big) + \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_{M_n})\boldsymbol{\mu}. \tag{A.5}$$

Under Assumption 3, it is straightforward to show that the infeasible optimal weights $\mathbf{w}_n^* = (w_{n,1}^*, \ldots, w_{n,M_n}^*)^\top$ can be obtained by setting $w_{n,m}^* = \gamma_{n,m}^* - \gamma_{n,m+1}^*$ for $m = 1, \ldots, M_n - 1$ and $w_{n,M_n}^* = \gamma_{n,M_n}^*$, where $\gamma_{n,1}^* = 1$ and

$$\gamma_{n,m}^* = \frac{\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}}{\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} + \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}} = \frac{\theta_{n,m}}{\theta_{n,m} + 1/n}, \quad m = 2, \ldots, M_n. \tag{A.6}$$

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

Hence, the risk of the optimal MA estimator is

$$R_n^{\mathrm{MA}}(\mathbf{w}_n^*) = \mathrm{tr}(\mathbf{P}_1\boldsymbol{\Omega}) + \sum_{m=2}^{M_n} \frac{\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}}{\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} + \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}} + \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_{M_n})\boldsymbol{\mu}. \tag{A.7}$$

Combining (A.3) and (A.7), the potential advantage of MA over MS is

$$\begin{aligned}\Delta_n &= R_n^{\mathrm{MS}}(m_n^*) - R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \\ &= \sum_{m=2}^{m_n^*}\left[\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\} - \frac{\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}}{\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} + \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}}\right] \\ &\quad + \sum_{m=m_n^*+1}^{M_n} \frac{\{\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}\}^2}{\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} + \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}},\end{aligned} \tag{A.8}$$

which implies that, as is expected, the optimal risk of MA is not larger than that of MS, i.e., $R_n^{\mathrm{MA}}(\mathbf{w}_n^*) \le R_n^{\mathrm{MS}}(m_n^*)$. We consider two scenarios: $M_n < m_n^{**}$ and $M_n \ge m_n^{**}$.

When $M_n < m_n^{**}$, we have $m_n^* = M_n$. It follows from Assumption 3, $\theta_{n,m_n^*} > 1/n$, and (A.6) that $\{\gamma_{n,m}^*\}_{m=1}^{M_n}$ is nonincreasing and $\gamma_{n,m_n^*}^* > 1/2$. Then, for a sufficiently large n,

$$\begin{aligned}R_n^{\mathrm{MA}}(\mathbf{w}_n^*) &\ge \sum_{m=1}^{m_n^*} \frac{\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}}{\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} + \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}} + \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_{m_n^*})\boldsymbol{\mu} \\ &= \sum_{m=1}^{m_n^*} \gamma_{n,m}^* \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\} + \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_{m_n^*})\boldsymbol{\mu} \\ &\ge \gamma_{n,m_n^*}^* \sum_{m=1}^{m_n^*} \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\} + \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_{m_n^*})\boldsymbol{\mu} \\ &\ge \frac{1}{2}\mathrm{tr}(\mathbf{P}_{m_n^*}\boldsymbol{\Omega}) + \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_{m_n^*})\boldsymbol{\mu} \ge \frac{1}{2}R_n^{\mathrm{MS}}(m_n^*).\end{aligned}$$

When $M_n \ge m_n^{**}$, it follows from Assumption 3 and (A.4) that $\gamma_{n,m_n^*}^* > 1/2 \ge \gamma_{n,m_n^*+1}^*$. Then, for a sufficiently large n,

$$\begin{aligned}R_n^{\mathrm{MA}}(\mathbf{w}_n^*) &\ge \sum_{m=1}^{M_n} \frac{\mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu}}{\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} + \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}} + \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_{M_n})\boldsymbol{\mu} \\ &= \sum_{m=1}^{m_n^*} \gamma_{n,m}^* \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\} \\ &\quad + \sum_{m=m_n^*+1}^{M_n} (1-\gamma_{n,m}^*)\boldsymbol{\mu}^\top(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\mu} + \boldsymbol{\mu}^\top(\mathbf{I}_n - \mathbf{P}_{M_n})\boldsymbol{\mu} \\ &\ge \gamma_{n,m_n^*}^* \sum_{m=1}^{m_n^*} \mathrm{tr}\{(\mathbf{P}_m - \mathbf{P}_{m-1})\boldsymbol{\Omega}\}\end{aligned}$$

https://doi.org/10.1017/S0266466624000355 Published online by Cambridge University Press

$$
\begin{aligned}
&+(1-\gamma^*_{n,m^*_n+1})\sum_{m=m^*_n+1}^{M_n}\boldsymbol{\mu}^\top(\mathbf{P}_m-\mathbf{P}_{m-1})\boldsymbol{\mu}+\boldsymbol{\mu}^\top(\mathbf{I}_n-\mathbf{P}_{M_n})\boldsymbol{\mu}\\
&\geq \frac{1}{2}\mathrm{tr}(\mathbf{P}_{m^*_n}\boldsymbol{\Omega})+\frac{1}{2}\boldsymbol{\mu}^\top(\mathbf{P}_{M_n}-\mathbf{P}_{m^*_n})\boldsymbol{\mu}+\boldsymbol{\mu}^\top(\mathbf{I}_n-\mathbf{P}_{M_n})\boldsymbol{\mu}\\
&\geq \frac{1}{2}R_n^{\mathrm{MS}}(m^*_n).
\end{aligned}
$$

Therefore, we have $R_n^{\mathrm{MS}}(m^*_n)\geq R_n^{\mathrm{MA}}(\mathbf{w}^*_n)\geq R_n^{\mathrm{MS}}(m^*_n)/2$ for any sufficiently large n, which yields that $R_n^{\mathrm{MS}}(m^*_n)$ and $R_n^{\mathrm{MA}}(\mathbf{w}^*_n)$ have the same order. This completes the proof of Theorem 1.

Supplementary Material

The supplementary material for this article can be found at https://doi.org/10.1017/S0266466624000355

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petroc, & F. Csake (Eds.), *Second international symposium on information theory* (pp. 268–281). Akademiai Kiado.

Alhorn, K., Dette, H., & Schorning, K. (2021). Optimal designs for model averaging in non-nested models. *Sankhya A*, 83, 745–778.

Alhorn, K., Schorning, K., & Dette, H. (2019). Optimal designs for frequentist model averaging. *Biometrika*, 106, 665–682.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16, 125–127.

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

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

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

Claeskens, G., Croux, C., & Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*, 62, 972–979.

Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35, 16–34.

Fang, F., Lan, W., Tong, J., & Shao, J. (2019). Model averaging for prediction with fragmentary data. *Journal of Business & Economic Statistics*, 37, 517–527.

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

Feng, Y., & Liu, Q. (2020). Nested model averaging on solution path for high-dimensional linear regression. *Stat*, 9, e317.

Feng, Y., Liu, Q., Yao, Q., & Zhao, G. (2022). Model averaging for nonlinear regression models. *Journal of Business & Economic Statistics*, 40, 785–798.

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

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

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

He, B., Liu, Y., Wu, Y., Yin, G., & Zhao, X. (2020). Functional martingale residual process for high-dimensional Cox regression with model averaging. *Journal of Machine Learning Research*, 21, 1–37.

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

Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *The Annals of Statistics*, 35, 1238–1277.

Ing, C.-K., & Wei, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *The Annals of Statistics*, 33, 2423–2474.

Lehrer, S., & Xie, T. (2017). Box office buzz: Does social media data steal the show from model uncertainty when forecasting for Hollywood? *Review of Economics and Statistics*, 99, 749–755.

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.

Liao, J., Zou, G., Gao, Y., & Zhang, X. (2021). Model averaging prediction for time series models with a diverging number of parameters. *Journal of Econometrics*, 223, 190–221.

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

Liu, Q., & Okui, R. (2013). Heteroscedasticity-robust C_p model averaging. *The Econometrics Journal*, 16, 463–472.

Liu, Q., Okui, R., & Yoshimura, A. (2016). Generalized least squares model averaging. *Econometric Reviews*, 35, 1692–1752.

Magnus, J. R., & Luca, G. D. (2016). Weighted-average least squares (WALS): A survey. *Journal of Economic Surveys*, 30, 117–148.

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

Mallows, C. (1973). Some comments on C_p. *Technometrics*, 15, 661–675.

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

Peng, J., & Yang, Y. (2022). On improvability of model selection by model averaging. *Journal of Econometrics*, 229, 246–262.

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. *Statistica Sinica*, 7, 221–242.

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics*, 35, 415–423.

Steel, M. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, 58, 644–719.

Stone, M. (1974). Cross-validation choice and assessment of statistical procedures. *Journal of the Royal Statistical Society: Series B*, 36, 111–147.

Trench, W. F. (1999). Asymptotic distribution of the spectra of a class of generalized Kac-Murdock-Szegö matrices. *Linear Algebra and its Applications*, 294, 181–192.

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

Wang, M., Zhang, X., Wan, A. T., You, K., & Zou, G. (2023). Jackknife model averaging for high-dimensional quantile regression. *Biometrics*, 79, 178–189.

Yan, X., Wang, H., Wang, W., Xie, J., Ren, Y., & Wang, X. (2021). Optimal model averaging forecasting in high-dimensional survival analysis. *International Journal of Forecasting*, 37, 1147–1155.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica*, 9, 475–499.

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

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

Zhang, X., Ullah, A., & Zhao, S. (2016a). On the dominance of Mallows model averaging estimator over ordinary least squares estimator. *Economics Letters*, 142, 69–73.

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

Zhang, X., & Wang, W. (2019). Optimal model averaging estimation for partially linear models. *Statistica Sinica*, 29, 693–718.

Zhang, X., Yu, D., Zou, G., & Liang, H. (2016b). 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., Zou, G., Liang, H., & Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, 115, 972–984.

Zhao, S., Zhou, J., & Li, H. (2016). Model averaging with high-dimensional dependent data. *Economics Letters*, 148, 68–71.