FEDERAL RESERVE BANK
PHILADELPHIA

Working Papers
RESEARCH DEPARTMENT

Model Averaging Prediction for Possibly Nonstationary Autoregressions

Tzu-Chi (Simon) Lin
Federal Reserve Bank of Philadelphia Supervision, Regulation, and Credit Department

WP 23-08

PUBLISHED
May 2023

ISSN: 1962-5361

DOI: https://doi.org/10.21799/frbp.wp.2023.08

Model Averaging Prediction for Possibly Nonstationary Autoregressions

Tzu-Chi (Simon) Lin[§]

Federal Reserve Bank of Philadelphia

Abstract

This paper considers the problem of model averaging (MA) predictions for the integrated autoregressive processes of infinite order (AR(∞)), which accommodates many stationary and nonstationary models in practice. We adopt the MA approach to forecast future observations and obtain the uniformly asymptotic expression for the mean squared prediction error (MSPE) of the averaging predictor. The MSPE can be decomposed into three components: non-stationary integration order, model complexity, and goodness-of-fit. The decomposition justifies that the advantage of MA comes from the diverse model intersections and provides the separation conditions under which the MA can attain strictly lower MSPE over model selection (MS). Regarding the predictive risk reduction by MA, it can be shown that the magnitude of MA improvement has the same order as the oracle minimum risk of MS under algebraic-decay case, while the magnitude is negligible under exponential-decay case. To pick the best choice of weights, we propose Shibata model averaging (SMA) criterion and show that, even without the integration order information, the selected weights by minimizing SMA and its variants including AIC-type and Mallow's MA criteria are asymptotically optimal in the sense that: (i) The probability of a criteria minimizer with positive weights on models of dimension less than the integration order is negligible almost surely; (ii) The averaging predictor formed by the selected weights will ultimately achieve the lowest possible MSPE.

Keywords: Time Series, Model Averaging, Model Selection, Nonstationary Autoregressions, Asymptotic Optimality

JEL classification: C22, C52, C53, E17

[§]10 N. Independence Mall W., Philadelphia, PA 19106; email: Simon.X.Lin@phil.frb.org

1 Introduction

To single out a good model for prediction is an important but challenging topic in statistics and econometrics literature, especially when enormous candidate models are considered. For a practitioner, model selection methods such as Akaike information criterion (Akaike, 1974), Bayesian information criterion (Schwarz, 1978), Mallows' C_p (Mallows, 1973), cross validation (Stone, 1974) are standard procedures to pin down the best predictive model. However, the issue of model uncertainty is raised by other researchers such as Yang (2001, 2007), Yuan and Yang (2005), and Hjort and Claeskens (2003). For any small perturbation of a data set, the model selection method may suggest diverse models and create controversial inference. Besides, under the single-best model approach, it may neglect the information embedded in other models and produce high prediction bias. The other approach for avoiding such shortcomings is model averaging. As a smoothed extension of model selection, model averaging combines information and predictive power by assigning different weights on all the candidate models and has better prediction performance than model selection under general circumstances.

The research core of model averaging is the weighting strategy, and the development based on frequentist-type methods has been burgeoning in the past two decades. Those methods are shown to be asymptotically optimal in the sense of achieving the lowest possible prediction losses. Under linear regression models, Hansen (2007) proposed a Mallow criterion to choose weights, and Hansen and Racine (2012) and Liu and Okui (2013) developed Jackknife model averaging and robust C_p, respectively, for heteroscedastic cases. Ando and Li (2014, 2017) suggested leave-one-out cross-validation and demonstrate the asymptotic optimality in the context of high-dimensional regressions. In recent years, model averaging methods and applications on time series models or dependent data have attracted growing interest. Those studies include lagged dependent variables (Zhang et al., 2013), time series errors (Cheng et al., 2015), factor-augmented regression (Cheng and Hansen, 2015), longitudinal data (Gao et al., 2016), vector autoregression (VAR) (Liao et al., 2019; Liao and Tsay, 2020), stationary AR(∞) process (Liao et al., 2021), time-varying model (Sun et al., 2021), and panel data VAR (Greenaway-McGrevy, 2022).

However, existing model averaging methods on dependent data in the literature are designed to stationary or local stationary time series. It precludes many economic, financial, or climate change data with non-stationary patterns. Second, as Liao et al. (2021) and Zhang and Liu (2022) pointed out, model averaging methods for predicting future observations are not explored well. Many asymptotic optimality results are established based on in-sample empirical losses, which may hinder practical implications since most of the forecasting problems concern the out-of-sample performances. In addition, it is even common in MS studies to examine the bias-variance trade-off under time series data (Shibata, 1980; Ing and Wei, 2003, 2005; Ing, 2007, 2020; Ing et al., 2010, 2012; Greenaway-McGrevy, 2015, 2019), but the similar trade-off analysis for MA methods are not well explored. Better investigation of the MA bias-variance trade-off will help to understand the mechanism, improve the prediction, and bring about the asymptotic optimality. The other fundamental issue of MA is only limited studies addressing the MA-MS comparisons. Theoretically, MA is the smoothed generalization of MS and can offer better performances than MS. Despite the rich literature focusing on MA prediction efficiency, what is the key aspect that attributes to the advantage of MA and whether the MA is favorable over MS under general circumstances are ambiguous. Last but not least, for frequentist-type MA methods aforementioned, there are two main approaches: Mallows MA (Hansen, 2007; Cheng et al., 2015; Liao and Tsay, 2020; Liao et al., 2021) and cross validation (Hansen and Racine, 2012; Zhang et al., 2013; Ando and Li, 2014; Gao et al., 2016; Liao et al., 2019; Zhang and Liu, 2022); however, few studies consider other MA criteria and analyze their relationships with existing methods.

To fill these gaps, we assume the data are generated from integrated AR(∞) processes and give an asymptotic expression for the MSPE of the model averaging predictor, under one-step ahead out-of-sample framework. The expression decomposes the MSPE into three components: non-stationary integration order, model complexity, and goodness-of-fit. It is not only a nontrivial extension of Ing and Wei (2003), Ing et al. (2010), but also provides the theoretical foundation for bias-variance trade-off analysis and MA-MS comparisons. According to the MSPE decomposition, the advantage of MA over MS comes from the cross intersection between the parsimonious model and the aggressive model. The cross product of diverse models only increases the model complexity term by the dimensionality of the parsimonious model, thus reducing the estimation variability. In contrast, the cross product on the goodness-of-fit term improves the fitting from the aggressive model, which has less-biased approximation of the AR(∞) process and further decreases the total MSPE. Compared with pure model selection, MA takes the advantage of intersections among diverse models for MSPE reduction via bias-variance trade-off.

Also, the MSPE decomposition offers another direction to gauge the MA-MS comparison. We identify the conditions under which the MA can reach strictly lower MSPE over MS. In a nutshell, if there is one candidate model whose model misspecification term is separable from other models, then there exists a weight vector, and the MSPE of the corresponding MA prediction is lower than the MSPEs of all models. From the separation condition, the potential MSPE reduction from the best MA approach is strictly greater than the best MS model. We would like to investigate whether the improvement is substantial or vanishing relative to the best MS model. Inspired by Peng and Yang (2022), we compare the magnitude of potential MSPE reduction from MA over MS with the oracle minimum predictive risk of MS. We demonstrate that, under the integrated AR(∞) model, if the goodness-of-fit term is algebraic decay as the dimension of fitted model increases, the magnitude of potential MSPE reduction has the same order as the oracle minimum predictive risk by MS. While under the exponential-decay scenario, the magnitude is asymptotic negligible. These results are consistent with current findings such as Peng and Yang (2022) and Xu and Zhang (2022), in which they consider non-stochastic regression design, while our extensions are under general autoregressive models with broader applicability.

We also define the asymptotical optimality conditions of MA criterion under the integrated AR(∞) model and propose three MA criteria: Akaike model averaging (AMA), Mallow's model averaging (MMA), and Shibata model averaging (SMA) criteria that are prediction efficient and asymptotic equivalent. The MA criterion is said to be asymptotically optimal under dth-order integrated (I(d)) AR(∞) process where d is unknown, if: (i) the unrestrictive weights selected by minimizing the criterion converges almost surely to the weight estimator constrained on the restrictive set, in which the first $d-1$ elements of any weight vectors are equal to zero; (ii) the restrictive weight estimator of the criterion will achieve the lowest MSPE asymptotically. Since any finite AR model with less than dth-order is severely underfitting and causes explosive bias, the first condition describes that, even without the integration order information, the selected weight vectors will eliminate any AR models with less than dth-order for prediction. This oracle property makes sure the MA prediction will only consider large enough models to control the forecast errors. The second condition is typical in MA literature: The selected weights will attain the lowest possible MSPE as do the infeasible optimal weights.

In summary, the major contributions of the current work can be listed as follows:

1. We adopt the MA approach to forecast integrated AR(∞) processes and obtain the uniformly asymptotic expression for the mean squared prediction error (MSPE) of the averaging predictor.

2. By virtue of the expression, the MSPE can be decomposed into three components: non-stationary integration order, model complexity, and goodness-of-fit. The bias-variance representation justifies that the diverse model intersections from the MA approach reduces the prediction errors and provides the separation conditions under which the MA can attain strictly lower MSPE over MS.

3. We demonstrate that, if the goodness-of-fit term is algebraic decay with the model

dimension, the magnitude of potential MSPE reduction has the same order as the oracle minimum predictive risk by MS. While under the exponential-decay scenario, the magnitude is asymptotic negligible. These results are consistent with existing literature but have broader applicability under the time series model.

4. We define the asymptotic optimality conditions under the integrated AR(∞) process, propose three MA criteria, and show they are asymptotic equivalent and all satisfy the asymptotic optimality conditions. It extends non-trivially the time series MA literature to nonstationarity and that of Ing and Wei (2005) and Ing et al. (2012) to MA. The convergence rate of unrestrictive weight estimators to the restrictive set for three MA criteria are also provided.

The rest of this paper is organized as follows. Section 2 demonstrates the integrated AR(∞) model, MA prediction, and assumptions and notations for this paper. Section 3 presents the asymptotic expression of the MSPE for the MA predictor and the details of MA-MS comparisons. Section 4 introduces the proposed SMA and shows the asymptotically optimal properties of the weight estimator based on SMA. Asymptotic equivalences of SMA and its relationship with other variants are also discussed. Section 5 concludes, and the technical proofs are relegated to the Appendix.

2 Model framework and assumptions

In this working paper, we follow the model setup as in Ing et al. (2010, 2012). Assume the data $\{y_1, ..., y_n\}$ generated from a dth-order integrated (I(d)) AR(∞) process are as below:

$$\big(1 + \sum_{j=1}^{\infty} a_j L^j\big)(1-L)^d y_t = \epsilon_t, \tag{2.1}$$

where $A(z) = 1 + \sum_{j=1}^{\infty} a_j z^j$ is the stationary component of the process satisfying

$$A(z) \neq 0 \text{ for all } |z| \leq 1 \text{ and } \sum_{j=1}^{\infty} |j a_j| < \infty, \tag{2.2}$$

L is the backshift operator, $0 \leq d < \infty$ is an unknown integer, and $\{\epsilon_t, t = 0, \pm 1, \pm 2, ...\}$ are independent variables having zero mean and same variance σ^2. Note that ϵ_t does not necessarily come from the same distribution. Besides, by Theorem 3.8.4 of Brillinger (2001), (2.2) yields

$$A^{-1}(z) = B(z) = 1 + \sum_{j=1}^{\infty} b_j z^j \neq 0 \text{ for all } |z| \leq 1 \text{ and } \sum_{j=1}^{\infty} |j b_j| < \infty, \tag{2.3}$$

where $b_0 = 1$. We also pose the initial condition $y_t = 0$ for $t \leq 0$. This dth-order integrated AR(∞) process includes many other models like the basic ARIMA(p, d, q) process, which is frequently used in nonstationary time series analysis.

To make a one-step-ahead prediction $\hat{y}_{n+1}$, a pool of finite-order models, AR(1), ..., AR(K_n) is considered, where K_n can increase to infinity as sample size n does with a slower rate. Note that since the true data generating process is AR(∞), every single AR(k) model is misspecified and only an approximation. For each AR(k) model, the least squares type estimator is defined as:

$$-\hat{\mathbf{a}}(k) = [\sum_{j=K_n}^{n-1} \mathbf{y}_j(k)\mathbf{y}_j^{'}(k)]^{-1} \sum_{j=K_n}^{n-1} \mathbf{y}_j(k) y_{j+1}, \quad 1 \leq k \leq K_n,$$

where $-\hat{\mathbf{a}}(k)$ and $\mathbf{y}_j(k) = (y_j, ..., y_{j-k+1})'$ are both $k \times 1$ vectors, and $\sum_{j=K_n}^{n-1} \mathbf{y}_j(k)\mathbf{y}_j^{'}(k)$ is a $k \times k$ matrix, $N = n - K_n$. We assume that for all $1 \leq k \leq K_n$, the inverse of $\sum_{j=K_n}^{n-1} \mathbf{y}_j(k)\mathbf{y}_j^{'}(k)$ exists. Thus, for each AR (k), the one-step-head prediction is $\hat{y}_{n+1}(k) = -\mathbf{y}_n^{'}(k)\hat{\mathbf{a}}(k)$.

Let $z_t = (1 - L)^d y_t$ be the dth differenced term. Then, $z_t = \sum_{j=1}^{t-1} b_j \epsilon_{t-j}$. Define $z_{t,\infty} = \sum_{i=0}^{\infty} b_i \epsilon_{t-i}$, $\mathbf{z}_t(v) = (z_t, ..., z_{t-v+1})'$, $\mathbf{z}_{t,\infty}(v) = (z_{t,\infty}, ..., z_{t-v+1,\infty})'$, and define $\mathbf{a}(v) = (a_1(v), ..., a_v(v))' = \text{argmin}_{\mathbf{c} \in R^v} \text{E}(z_{t,\infty} + \mathbf{z}_t^{'}(v)\mathbf{c})^2$. For $1 \leq k \leq K_n$ and $N = n - K_n$, the $y_{n+1} - \hat{y}_{n+1}(k)$ can be rewritten as :

$$\begin{aligned}
y_{n+1} - \hat{y}_{n+1}(k) &= y_{n+1} - \mathbf{y}_n^{'}(k) \Big[\sum_{j=K_n}^{n-1} \mathbf{y}_j(k)\mathbf{y}_j^{'}(k) \Big]^{-1} \sum_{j=K_n}^{n-1} \mathbf{y}_j(k) y_{j+1} \\
&= y_{n+1} \\
&- \mathbf{y}_n^{'}(k) Q^{'}(k) \Big[\sum_{j=K_n}^{n-1} Q(k)\mathbf{y}_j(k)\mathbf{y}_j^{'}(k)Q^{'}(k) \Big]^{-1} \\
&\times \sum_{j=K_n}^{n-1} Q(k)\mathbf{y}_j(k)[\epsilon_{j+1,k-d} - \mathbf{y}_j^{\prime}(k)Q^{'}(k)b(k)] \\
&= \epsilon_{n+1,k-d} - \Big\{ N^{-1}\mathbf{s}_{n,n}^{'}(k) \Big[N^{-1} \sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(k)\mathbf{s}_{j,n}^{'}(k) \Big]^{-1} \sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(k)\epsilon_{j+1,k-d} \Big\} \\
&= \epsilon_{n+1,k-d} - \Big\{ N^{-1}\mathbf{s}_{n,n}^{'}(k)\hat{\Omega}_n^{-1}(k) \sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(k)\epsilon_{j+1,k-d} \Big\}, \qquad (2.4)
\end{aligned}$$

where

$$\epsilon_{n+1,k-d} = \begin{cases} z_{j+1}, & k = d, \\ z_{j+1} + \mathbf{a}^{'}(k-d)\mathbf{z}_j(k-d), & k > d; \end{cases}$$

$$b(k) = \begin{cases} \left(\mathbf{a}'(k-d), -\mathbf{1}_d'\right)', & k > d \geq 1, \\ -\mathbf{1}_k, & d \geq k \geq 1, \\ \mathbf{a}(k), & d = 0, \end{cases}$$

with $\mathbf{1}_k$ is the k-dimensional vector of 1's; and $\mathbf{s}_{j,n}(k) = G_n(k)Q(k)\mathbf{y}_j(k)$.

The $G_n(k)$, $Q(k)$ are $k \times k$ matrix defined by

$$G_n(k) = \begin{cases} \mathrm{diag}(1, ..., 1, N^{-d+1/2}, ..., N^{-1/2}), & k > d \geq 1, \\ \mathrm{diag}(N^{-d+1/2}, ..., N^{-d+k-1/2}), & d \geq k \geq 1, \\ \mathrm{diag}(1, ..., 1), & d = 0, \end{cases}$$

$$Q_n(k)y_j(k) = \begin{cases} \left(\mathbf{z}_j'(k-d)1, y_j(d), ..., y_j(1)\right)', & k > d \geq 1, \\ (y_j(d), ..., y_j(d-k+1))', & d \geq k \geq 1, \\ \mathbf{z}_j(k), & d = 0, \end{cases}$$

with $y_j(v) = (1-L)^{d-v}y_j$, and diag(.) represents diagonal matrix. Let $a_i(v) = 0$ if $i > v \geq 0$ or $v \leq 0$. In this working paper, $\mathbf{a}(v)$ is sometimes viewed as an infinite dimensional vector with the ith element equal to $a_i(v), i = 1, 2, ...$ Define $\|\mathbf{d}\|_z^2 = \sum_{1 \leq i,j \leq \infty} d_i d_j \chi_{i-j}$, where $\chi_{i-j} = \mathrm{E}(z_{i,\infty} z_{j,\infty})$, and $\mathbf{d} = (d_1, d_2, ...)'$ is an infinite dimensional vector that belongs to $l^2(\mathcal{Z}^+)$, i.e., $\sum_{i \in \mathcal{Z}^+} d_i^2 < \infty$. Since

$$z_{t,\infty} + \sum_{i=1}^{\infty} a_i z_{t-i,\infty} = \epsilon_t,$$

then for all $v \geq 0$,

$$\|\mathbf{a} - \mathbf{a}(v)\|_z^2 = \mathrm{E}\Big[\sum_{i=1}^{\infty} (a_i - a_i(v)) z_{t-i,\infty}\Big]^2 = \mathrm{E}\Big[z_{t,\infty} + \sum_{i=1}^{v} a_i(v) z_{t-i,\infty}\Big]^2 - \sigma^2. \tag{2.5}$$

Let $\mathbf{w} = (w_1, ..., w_{K_n})'$ be a weight vector such that

$$\sum_{k=1}^{K_n} w_k = 1, \quad w_k \geq 0, \quad \forall\, k \in \{1, ..., K_n\},$$

then, the weight vector $\mathbf{w}$ belongs to the set $\mathcal{H}_n := \{\mathbf{w} \in [0,1]^{K_n} : \sum_{k=1}^{K_n} w_k = 1\}$. Combining all possible one-step-head prediction $\hat{y}_{n+1}(k)$, we construct an averaging predictor as

$$y_{n+1}(\mathbf{w}) = \sum_{k=1}^{K_n} w_k \hat{y}_{n+1}(k),$$

and the MSPE of averaging predictor is $E(y_{n+1} - y_{n+1}(\mathbf{w}))$.

Before the section ends, we list and discuss the assumptions for this working paper:

Assumption 1. d is a fixed nonnegative integer and bounded by some $\bar{d} < \infty$.
Assumption 2. Let $F_{t,m,\mathbf{v}_m}(.)$ be the distribution function of the linear combination of innovations: $\{\mathbf{v}'_m(\epsilon_t, ..., \epsilon_{t-m+1})\}$, where $\{\mathbf{v}_m = (v_1, ..., v_m)' \in R^m, \ \sum_{j=1}^m v_j^2 = 1\}$. For all $m \geq 1$, $m \leq t < \infty$, there exists some real positive numbers α, C, and δ such that the distribution function $F_{t,m,\mathbf{v}_m}(.)$ satisfies the locally Hölder condition of order α:

$$|F_{t,m,\mathbf{v}_m}(x) - F_{t,m,\mathbf{v}_m}(y)| \leq C|x-y|^{\alpha}, \quad |x-y| \leq \delta.$$

Assumption 3. $\sup_{-\infty<t<\infty} \mathrm{E}|\epsilon_t|^q < \infty$, and $q \in \mathcal{N}$.
Assumption 4. $K_n^{\max\{4d-1,3\}} = o(n)$.

Assumption 1 implies that the y_t is generated from an integrated autoregressive process with a finite integration order d, $0 \leq d < \bar{d}$. Assumption 2 is the non-singularity assumption of Ing et al. (2010, 2012) used to establish the negative moment bounds of the minimum eigenvalue of the Fisher information matrix in (2.4). Assumption 2 is fulfilled by most continuous-type distributions, such as normal distribution, see Ing and Sin (2006) for more details. Assumption 3 is the moment conditions of ϵ_t. Assumption 4 imposes the limitations on the number of models relative to the sample size, and reflects the facts that, the time series y_t has higher correlations with higher integration order and yields smaller minimum eigenvalue of the information matrix. There is also a trade-off between the moment condition on ϵ_t in Assumption 3 and divergence rate of K_n in Assumption 4. With weaker moment condition in Assumption 3, the K_n will be more restrictive than Assumption 4. When $d = 0$, the bound of K_n is slightly restrictive than the MS cases in Ing et al. (2010, 2012), where $K_n^{2+\delta} = o(n)$ for some $\delta > 0$ and ours is $K_n^3 = o(n)$. This is the price paid for MA approach because MA generalizes MS from countable model comparisons to finding an extreme weight estimator on an uncountable set. When $d \geq 1$, there is no difference with Ing et al. (2010, 2012).

In the following sections, we use C to denote some positive constant, which is independent from sample size n. And C may represent different values in different equations. $\xrightarrow{p}$ and $\xrightarrow{a.s.}$ represent convergence in probability and almost surely respectively. $\|\mathbf{v}\|_2$ is the Euclidean norm for vector $\mathbf{v}$ and $\|A\|^2 = \lambda_{\max}(A'A)$ is the maximum eigenvalue of matrix $A'A$. $a_n = \Theta(b_n)$ means a_n has the same order as b_n if $c_1 \, b_n \leq a_n \leq c_2 \, b_n$ for some positive numbers but c_1, c_2 do not depends on n. The $K_n \times K_n$ matrices defined below are also frequently used in this article.

$$\Pi_{\min}(K_n) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \dots & K_n \end{pmatrix},$$

$$\Pi_{\min}(K_n)_{ij} = \min(i, j), \quad 1 \le i,\, j \le K_n,$$

and

$$\Pi_{\max}(K_n) = \begin{pmatrix} 1 & 2 & \dots & K_n \\ 2 & 2 & \dots & K_n \\ \vdots & \vdots & \ddots & \vdots \\ K_n & K_n & \dots & K_n \end{pmatrix},$$

$$\Pi_{\max}(K_n)_{ij} = \max(i, j), \quad 1 \le i,\, j \le K_n. \tag{2.6}$$

3 Asymptotic expression of the MSPE

In this section, assume the integration order d is known, and we consider combining a class of finite-order models, AR($\max(1, d)$), ..., AR(K_n), to make forecast and derive the asymptotic expression of the MSPE of the averaging predictors. The most parsimonious candidate model in the class has dimension of order d if $d \ge 1$. In practice, the class of models starts from AR(1) since d is unknown to the researcher and may equal to zero (i.e., the data generating process is stationary). In the later section, we will show that the weights of any models with dimensions less than d converge to zero almost surely if the MA weights are selected based on SMA or its variants.

Given the family of AR(k) models, $1 \le k \le K_n$, let $\mathbf{w} = (w_1, ..., w_{K_n})$ be a weight vector such that

$$\sum_{k=1}^{K_n} w_k = 1, \quad w_k \ge 0, \quad w_k = 0,\ 1 \le k < d.$$

The restrictive weight vector $\mathbf{w}$ belongs to the set $\mathcal{H}_n^d := \{\mathbf{w} \in [0,1]^{K_n} : \sum_{k=1}^{K_n} w_k = 1, w_k = 0,\ 1 \le k < d\}$, where $\mathcal{H}_n^d$ is a subset of $\mathcal{H}_n$.

With fix $\mathbf{w} \in \mathcal{H}_n^d$, and follow the decomposition of Ing et al. (2010), we can rewrite the

MSPE of averaging prediction as

$$
\begin{aligned}
\mathrm{E}(y_{n+1}-\hat{y}_{n+1}(\mathbf{w}))^2-\sigma^2 &= \mathrm{E}(y_{n+1}-\sum_{k=\max(1,d)}^{K_n} w_k\hat{y}_{n+1}(k))^2-\sigma^2 \\
&= \mathrm{E}(\sum_{k=\max(1,d)}^{K_n} w_k f_n(k) \; + \sum_{k=\max(1,d)}^{K_n} w_k S_n(k-d))^2, \qquad (3.1)
\end{aligned}
$$

where

$$
\begin{aligned}
f_n(k) &= \frac{\mathbf{s}_{n,n}^{'}(k)}{\sqrt{N}}\hat{\Omega}_n^{-1}(k)\left(\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(k)\epsilon_{j+1,k-d}\right), \\
S_n(k-d) &= \epsilon_{n+1,k-d}-\epsilon_{n+1} = \sum_{i=1}^{n}(a_i-a_i(k-d))z_{n+1-i},
\end{aligned}
$$

and $\epsilon_{n+1,k-d}$, $\mathbf{s}_{j,n}(k)$, $\hat{\Omega}_n^{-1}(k)$ are defined after (2.4). In the following theorem, we obtain an uniformly asymptotic expression of the MSPE of the averaging prediction.

Theorem 1. Assume Assumption 1-4 and (2.3), then

$$
\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}(y_{n+1}-\hat{y}_{n+1}(\mathbf{w}))^2-\sigma^2}{L_n^d(\mathbf{w})}-1\right| = 0,
$$

where

$$
\begin{aligned}
L_n^d(\mathbf{w}) &= \sigma^2\frac{d^2+d}{N}+\sigma^2\frac{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}{N}+\|\mathbf{a}-\mathbf{a}(\mathbf{w},d)\|_z^2 \\
&= \sigma^2\frac{d^2}{N}+\sigma^2\frac{\sum_{\max(1,d)\le i,j\le K_n} w_iw_j\min(i,j)}{N}+\sum_{\max(1,d)\le i,j\le K_n} w_iw_j\|\mathbf{a}-\mathbf{a}(\max(i,j)-d)\|_z^2.
\end{aligned}
\qquad (3.2)
$$

From (3.2), $L_n^d(\mathbf{w})$ can be decomposed as three parts: First is the integradness term, $\sigma^2\frac{d^2+d}{N}$, which is an estimation from a nonstationary component. The second one is the model complexity under model averaging, $\sigma^2\frac{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}{N}$. And the third one, $\|\mathbf{a}-$

$\mathbf{a}(\mathbf{w}, d)\|_z^2$ represents as model misspecification under averaging. Note that for any $\mathbf{w} \in \mathcal{H}_n^d$,

$$\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w} = \sum_{\max(1,d)\le i,j\le K_n} w_i w_j \min(i,j)$$

$$\|\mathbf{a} - \mathbf{a}(\mathbf{w}, d)\|_z^2 = \mathrm{E}\big[\sum_{v=1}^{K_n} w_v \,(\sum_{i=1}^{\infty}(a_i - a_i(v-d))z_{t-i,\infty})\big]^2$$
$$= \sum_{\max(1,d)\le i,j\le K_n} w_i w_j \|\mathbf{a} - \mathbf{a}(\max(i,j)-d)\|_z^2,$$

$\Pi_{\min}(K_n)$ are defined after (2.6).

Besides, if only one model, say AR(k) is used in the prediction, the weight vector $\mathbf{w}$ deduces to $\mathbf{w}_k$, where $\mathbf{w}_k \in \mathcal{H}_n^d$ is the weight vector where the k-th element equals 1 and other elements are zeros. Then, (3.2) can be rewritten as

$$L_n^d(\mathbf{w}_k) = \sigma^2\frac{d^2+d}{N} + \sigma^2\frac{\mathbf{w}_k'\Pi_{\min}(K_n)\mathbf{w}_k - d}{N} + \|\mathbf{a} - \mathbf{a}(\mathbf{w}_k, d)\|_z^2,$$
$$= \sigma^2\frac{d^2+d}{N} + \sigma^2\frac{k-d}{N} + \|\mathbf{a} - \mathbf{a}(k-d)\|_z^2, \tag{3.3}$$

which is the MSPE representation derived in Theorem 2 of Ing et al. (2010), and when $d=0$, the autoregressive process is stationary, (3.3) becomes exactly the AR(k) expression in Ing and Wei (2003, 2005). Thus, (3.2) is a generalization of asymptotic expressions in Ing et al. (2010); Ing and Wei (2003, 2005). Theorem 1 also provides the explicit bias-variance trade-off for the MA approach. In (3.2), the cross product of diverse models increases the model complexity term by the dimensionality of the parsimonious one, thus reducing the estimation variability. In contrast, the cross product on the goodness-of-fit term improves the fitting from the aggressive model, which has a less-biased approximation of the AR(∞) process, and further decreases the total MSPE. Compared with pure model selection, MA takes the advantage of intersections among diverse models for MSPE reduction.

As a byproduct from Theorem 1, we can identify the conditions under which there exists at least one MA prediction that can reach strictly lower MSPE than MS. Observe that

$$\mathcal{H}_n^d = \{\mathbf{w} \in [0,1]^{K_n} : \sum_{k=1}^{K_n} w_k = 1, w_k = 0,\ 1 \le k < d\}$$
$$= \{\mathbf{w} \in [0,1]^{K_n - \max(1,d)+1} : \sum_{k=\max(1,d)}^{K_n} w_k = 1, w_k \ge 0,\ \max(1,d) \le k \le K_n\},$$

and the model selection of AR($\max(1,d)$), ..., AR(K_n), is one-to-one corresponding to vertices of $\mathcal{H}_n^d$. For instance, if model AR(K_n) is used, then the corresponding $\mathbf{w}$ is

$\mathbf{w}_{K_n} := (0, 0, ..., 1) \in \mathcal{H}_n^d$, where the last element is 1 and others are equal to zero. Define $\mathcal{V}(\mathcal{H}_n^d)$ is the set of all the vertices in $\mathcal{H}_n^d$.

Corollary 1. Assume Assumption 1-4 and (2.3) as in Theorem 1. If there is a $k \in \{\max(1, d), ..., K_n\}$ such that

$$\|\mathbf{a} - \mathbf{a}(k-d)\|_z^2 \neq \|\mathbf{a} - \mathbf{a}(l-d)\|_z^2, \ \ \forall\, l \in \{\max(1, d), ..., K_n\}, \ \ l \neq k, \qquad (3.4)$$

then there exists at least one weight vector $\mathbf{w}_n^\diamond$ in $\mathcal{H}_n^d/\mathcal{V}(\mathcal{H}_n^d)$ such that

$$\inf_{\mathbf{w}\in\mathcal{H}_n^d/\mathcal{V}(\mathcal{H}_n^d)} L_n^d(\mathbf{w}) \leq L_n^d(\mathbf{w}^\diamond) < \min_{\mathbf{w}\in\mathcal{V}(\mathcal{H}_n^d)} L_n^d(\mathbf{w}),$$

where $\|\mathbf{a} - \mathbf{a}(v)\|_z^2$ is defined in (2.5).

Corollary 1 and sufficient condition (3.4) show that, if the prediction bias of one model is separable from others, model averaging can further reduce the forecast risk from the model selection. And there exists at least one model averaging predictor such that the corresponding MSPE is strictly less than the MSPEs of all the model selection predictors.

Based on the asymptotic expression, we can also compare the predictive risk of model averaging and model selection. Let $\mathcal{V}(\mathcal{H}_n^d)$ be the set of all the vertices in $\mathcal{H}_n^d$ defined after Corollary 1. Denote $\mathbf{w}_{k_n^*} := \arg\min_{\mathbf{w}\in\mathcal{V}(\mathcal{H}_n^d)} L_n^d(\mathbf{w})$, then $L_n^d(\mathbf{w}_{k_n^*})$ is the minimum predictive risk of model selection methods.

As Peng and Yang (2022), the appropriate comparison between model averaging and model selection is to analyze the potential risk reduction of MA from MS defined as:

$$\Delta_n = L_n^d(\mathbf{w}_{k_n^*}) - L_n^d(\mathbf{w}_n^*), \qquad (3.5)$$

where $L_n^d(\mathbf{w}_n^*) := \inf_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w})$. And we investigate the magnitude of Δ_n relative to $L_n^d(\mathbf{w}_{k_n^*})$ under the exponential-decay case and algebraic-decay case, which are frequently used in time series research.

(i) Exponential-decay case:

$$\|\mathbf{a} - \mathbf{a}(v)\|_z^2 = C\exp(-\alpha(v)),$$

where α is a positive constant.

(ii) Algebraic-decay case:

$$\|\mathbf{a} - \mathbf{a}(v)\|_z^2 = Cv^{-\alpha},$$

where α is a positive constant. The exponential-decay and algebraic-decay scenarios used are simplified but have the same orders of k_n^* as the example 1 and 2 in Ing and Wei (2005).

Theorem 2. Assume Assumption 1-4 and (2.3).
(i) Let $A_j = \|\mathbf{a} - \mathbf{a}(j-d)\|_z^2$, then

$$L_n^d(\mathbf{w}_n^*) = \frac{\sigma^2 d^2}{N} + \frac{\sigma^2 \max(1,d)}{N} + A_{K_n} + \sum_{j=\max(1,d)+1}^{K_n} \frac{\frac{\sigma^2}{N}(A_{j-1} - A_j)}{\frac{\sigma^2}{N} + A_{j-1} - A_j}. \tag{3.6}$$

(ii) under exponential-decay case:

$$\Delta_n = o(L_n^d(\mathbf{w}_{k_n^*})),$$

(iii) under algebraic-decay case:

$$\Delta_n = \Theta(L_n^d(\mathbf{w}_{k_n^*})),$$

where Δ_n defined after (3.5).

From Corollary 1, we know that if there exists a model whose prediction bias is separable from the other models, Δ_n is greater than zero. Theorem 2 further provides a measurement of the potential improvability by MA, relative to the minimum predictive risk attained by MS. Under the integrated AR(∞) model, if the goodness-of-fit term is algebraic decay as the model dimension increases, the magnitude of potential MSPE reduction has the same order as the oracle minimum predictive risk by MS. While under the exponential-decay scenario, the magnitude is asymptotic negligible. These results are consistent with current findings such as Peng and Yang (2022) and Xu and Zhang (2022), in which they consider non-stochastic regression design while our extensions are under general autoregressive models with broader applicability.

For any MS and MA criterion, $MS_n(\mathbf{w})$ and $MAC_n(\mathbf{w})$, respectively, define

$$\hat{\mathbf{w}}_{MS_n}^d := \arg\min_{\mathbf{w}\in\mathcal{V}(\mathcal{H}_n^d)} MAC_n(\mathbf{w}),$$
$$\hat{\mathbf{w}}_{MAC_n}^d := \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} MAC_n(\mathbf{w}),$$

and define their difference as:

$$\hat{\Delta}_n := L_n^d(\hat{\mathbf{w}}_{MS_n}^d) - L_n^d(\hat{\mathbf{w}}_{MAC_n}^d).$$

Then, we can get the following consequences:

Corollary 2. Assume Assumption 1-4 and (2.3). If

$$\frac{L_n^d(\hat{\mathbf{w}}_{MS_n}^d)}{L_n^d(\mathbf{w}_{k_n^*})} \xrightarrow{p} 1, \quad \frac{L_n^d(\hat{\mathbf{w}}_{MAC_n}^d)}{L_n^d(\mathbf{w}_n^*)} \xrightarrow{p} 1, \tag{3.7}$$

then, under the exponential-decay case:

$$\hat{\Delta}_n = o(L_n^d(\hat{\mathbf{w}}_{MS_n}^d)).$$

Under the algebraic-decay case:

$$\hat{\Delta}_n = \Theta(L_n^d(\hat{\mathbf{w}}_{MS_n}^d)).$$

And

$$L_n^d(\hat{\mathbf{w}}_{MAC_n}^d) = \Theta(L_n^d(\hat{\mathbf{w}}_{MS_n}^d)),$$

under either conditions.

Corollary 2 conducts the MSPE comparison based on the selected model and weight estimator from the data. If the MS and MA are asymptotic optimal, the results of Theorem 2 hold for the selected model and weight estimator. Under Assumption 1-4, Theorem 3.1 of Ing et al. (2012) shows that the selected model from AIC or its equivalent MS criteria satisfies the asymptotic condition of MS part in (3.7). In the next section, we will show that the weights selected by SMA and its equivalent MA criteria satisfy the asymptotic condition of the MA part in (3.7).

4 The MA criteria

In practice, the integration order d is unknown and so is the restrictive set $\mathcal{H}_n^d$. To form a MA prediction under $0 \leq d < \bar{d}$, we consider combining all finite-order AR models, from AR(1) to AR(K_n) to forecast future observation. In this section, we will propose three MA criteria and show that the weight estimator by minimizing those criteria are asymptotically optimal even if d is unknown to us. The asymptotic optimality conditions under the integrated AR(∞) process are defined as below:

For any MA criterion, $MAC_n(\mathbf{w})$, define

$$\hat{\mathbf{w}}_{MAC_n} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} MAC_n(\mathbf{w}),$$
$$\hat{\mathbf{w}}_{MAC_n}^d := \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} MAC_n(\mathbf{w}).$$

The MA criterion is said to be asymptotically optimal without the integration order information if

$$\|\hat{\mathbf{w}}_{MAC_n} - \hat{\mathbf{w}}_{MAC_n}^d\|_2 \xrightarrow{a.s.} 0, \tag{4.1}$$

and

$$\lim_{n\to\infty} \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{MAC_n}^d)} \xrightarrow{p} 1, \tag{4.2}$$

where $L_n^d(\mathbf{w}_n^*) := \inf_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w})$. The weight minimizer of the criterion on the unrestrictive H_n satisfying condition (4.1) will converge to the weight minimizer on the restrictive H_n^d almost surely even d is unknown. By Theorem 1, $L_n^d(\mathbf{w})$ is the uniformly asymptotic expression of the averaging prediction MSPE, the selected weights on the restrictive H_n^d satisfying (4.2) can achieve the lowest possible MSPE asymptotically.

Inspired by Shibata (1980), we propose the Shibata model averaging (SMA) criterion:

$$S_n(\mathbf{w}) = (N + \mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})\hat{\sigma}_w^2, \tag{4.3}$$

where $\Pi_{\min}(K_n)$ and $\Pi_{\max}(K_n)$ are defined after (2.6), and $\hat{\sigma}_w^2 = \frac{1}{N}\sum_{t=K_n}^{n-1}(y_{t+1}-\hat{y}_{t+1}(\mathbf{w}))^2$. SMA generalizes the Shibata (1980) model selection criterion to model averaging, and it has a close relationship with AIC-type model averaging criterion criterion and Mallow's model averaging criterion. While $\mathbf{w}$ shrinks to $\mathbf{w}_k$, where $\mathbf{w}_k \in \mathcal{H}_n$ is the weight vector where the k-th element equals 1 and other elements are zeros. Then, $S_n(\mathbf{w})$ will deduce to the Shibata's model selection criterion $S_n(k)$, where $S_n(k) := (N + 2k)\hat{\sigma}^2(k)$, with the empirical MSPE of AR(k), $\hat{\sigma}^2(k) = \frac{1}{N}\sum_{t=K_n}^{n-1}(y_{t+1} - \hat{y}_{t+1}(k))^2$. Intuitively, the usage of matrix $\Pi_{\min}(K_n) + \Pi_{\max}(K_n)$ rather than $2\Pi_{\min}(K_n)$ or $2\Pi_{\max}(K_n)$ is to balance the bias-variance trade-off represented in (3.2) of Theorem 1. To prove the asymptotic optimality of SMA, we need an additional assumption:

Assumption 5. $K_n^{1/2}/N\eta_n^d \longrightarrow 0$, where $\eta_n^d := L_n^d(\mathbf{w}_n^*) = \inf_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w})$.

As pointed out in Cheng et al. (2015) or Liao et al. (2021), many MA approaches require the strong assumption on K_n and may result in inferior prediction due to preclude the optimal model. Assumption 5 is quite mild and consists of the best model. For example, under either the exponential decay or algebraic decay as in Theorem 2, $\|\mathbf{a} - \mathbf{a}(v)\|_z^2 = C\exp(-\alpha(v))$ or $\|\mathbf{a} - \mathbf{a}(v)\|_z^2 = Cv^{-\alpha}$, by (3.6), it can be shown that

$$\begin{aligned} L_n^d(\mathbf{w}_n^*) &\geq \sum_{j=\max(1,d)+1}^{k_n^*} \frac{\frac{\sigma^2}{N}(A_{j-1} - A_j)}{\frac{\sigma^2}{N} + A_{j-1} - A_j} \\ &\geq C\frac{k_n^*}{N}, \end{aligned}$$

where the second inequality is guaranteed by the fact $\sigma^2/N < C\big(A_{j-1} - A_j\big)$ for some large C and for all $j = \max(1, d), ..., k_n^*$ under exponential decay or algebraic decay. In either case, K_n can have order up to $o\big((k_n^*)^2\big)$. Hence, Assumption 5 does not preclude the optimal model and can avoid the prevalent shortcoming in existing MA literature.

Theorem 3. Assume Assumption 1-5 and (2.3). If

$$\eta_n^d = \inf_{\mathbf{w}\in\mathcal{H}_n} L_n^d(\mathbf{w}); \ \lim_{n\to\infty} N\eta_n^d \to \infty,$$

then we have

(i)

$$\|\hat{\mathbf{w}}_{S_n} - \hat{\mathbf{w}}_{S_n}^d\|_2 \xrightarrow{a.s.} 0,$$

(ii)

$$\lim_{n\to\infty} \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{S_n}^d)} \xrightarrow{p} 1.$$

Thus, the SMA is asymptotically optimal without the integration order information satisfying (4.1) and (4.2), where

$$\hat{\mathbf{w}}_{S_n} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} S_n(\mathbf{w}),$$
$$\hat{\mathbf{w}}_{S_n}^d := \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} S_n(\mathbf{w}).$$

Theorem 3 asserts that the MSPE of the model averaging predictor with the weight selected by Shibata's condition $S_n(.)$ can achieve the best compromise between model complexity $\sigma^2(\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}/N)$, and goodness-of-fit, $\|\mathbf{a}-\mathbf{a}(\mathbf{w},d)\|_z^2$ under a one-step-ahead forecast framework.

We also explore the relationship of SMA with other MA criteria, which is limited among the existing MA studies. Define Mallow's model averaging (MMA) criterion as:

$$C_n(\mathbf{w}) = N\hat{\sigma}_w^2 + (\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w} - N)\hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is some consistent estimator of σ^2, which does not depend on $\mathbf{w}$.
Similarly, define Akaike model averaging (AMA) criterion as:

$$A_n(\mathbf{w}) = \log(\hat{\sigma}_w^2) + \frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N},$$

where $\hat{\sigma}_w^2$, $\Pi_{\min}(K_n)$, and $\Pi_{\max}(K_n)$ are defined as in the (4.3). Similar to the SMA, MMA and AMA are the model averaging generalization of Mallows' C_p and AIC. For example, while $\mathbf{w}$ shrinks to $\mathbf{w}_k$, where $\mathbf{w}_k \in \mathcal{H}_n$ is the weight vector with the k-th element equals 1 and other elements are zeros. Then, $A_n(\mathbf{w})$ will deduce to the well-known Akaike information criterion $A_n(k) := \log(\hat{\sigma}^2(k)) + 2k/N$.

Remark 1. Note that

$$\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w} = \sum_{1 \le i,j \le K_n} w_i w_j (i+j) = 2\sum_{k=1}^{K_n} w_k k,$$

since $\sum_{i=1}^{K_n} w_i = 1$. Thus, optimize

$$C_n(\mathbf{w}) = N\hat{\sigma}_w^2 + (\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w} - N)\hat{\sigma}^2,$$

is equivalent to optimize

$$C_n(\mathbf{w}) = N\hat{\sigma}_w^2 + (2\sum_{k=1}^{K_n} w_k k)\hat{\sigma}^2,$$

which is exactly the MMA criterion proposed by Hansen (2007).

Theorem 4. Assume Assumption 1-5 and (2.3). If

$$\eta_n^d = \inf_{\mathbf{w} \in \mathcal{H}_n} L_n^d(\mathbf{w});\ \lim_{n\to\infty} N\eta_n^d \to \infty.$$

For Mallow's model averaging (MMA) criterion:

$$\|\hat{\mathbf{w}}_{C_n} - \hat{\mathbf{w}}_{C_n}^d\|_2 \xrightarrow{a.s.} 0,$$

and

$$\lim_{n\to\infty} \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{C_n}^d)} \xrightarrow{p} 1.$$

For Akaike model averaging (AMA) criterion:

$$\|\hat{\mathbf{w}}_{A_n} - \hat{\mathbf{w}}_{A_n}^d\|_2 \xrightarrow{a.s.} 0,$$

and

$$\lim_{n\to\infty} \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{A_n}^d)} \xrightarrow{p} 1.$$

The $C_n(\mathbf{w})$ and $A_n(\mathbf{w})$ are both asymptotically optimal without the integration order information satisfying (4.1) and (4.2), where

$$\hat{\mathbf{w}}_{A_n} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} A_n(\mathbf{w}), \quad \hat{\mathbf{w}}^d_{A_n} := \arg\min_{\mathbf{w}\in\mathcal{H}^d_n} A_n(\mathbf{w}).$$
$$\hat{\mathbf{w}}_{C_n} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} C_n(\mathbf{w}), \quad \hat{\mathbf{w}}^d_{C_n} := \arg\min_{\mathbf{w}\in\mathcal{H}^d_n} C_n(\mathbf{w}).$$

Theorem 4 shows that the MSPE of the model averaging predictor with the weight selected by AMA or MMA can achieve the best compromise between model complexity and goodness-of-fit as SMA. Thus, those three criteria are asymptotically equivalent and their differences (up to a monotone transformation) are negligible relative to the oracle minimum predictive risk of MA as shown in Lemma 10 and the proof of Theorem 4 in the Appendix. Shibata (1980), Ing and Wei (2005), and Ing et al. (2012) present the similar results under MS. Theorem 4 is the extension to MA under general AR(∞) model.

5 Conclusion

In this paper, we study model averaging prediction under an integrated AR(∞) framework and obtain the uniformly asymptotic expression for the mean squared prediction error (MSPE) of the averaging predictor. The MSPE can be decomposed into three components: non-stationary integration order, model complexity, and goodness-of-fit via the expression. The asymptotic expression provides theoretical justification that the diverse model intersections from the MA approach can decrease the model complexity and misspecification if the model misspecification terms from different models are separable. Regarding the predictive risk reduction by MA, it can be shown that the magnitude of MA improvement has the same order as the oracle minimum risk of MS under algebraic-decay case, while the magnitude is negligible under exponential-decay case. And the risk reduction conclusion holds for selected MS and MA if the data-driven MS and MA are asymptotic efficient. To pick the best choice of weights, we propose Shibata model averaging (SMA) criterion and show that, even without the integration order information, the weight estimator of SMA and its variants including AMA and MMA criteria are asymptotically optimal. It would be an interesting future research topic to extend the method to allow structural changes or time trend and to develop inference methods for the weight estimator.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

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

Ando, T. and Li, K.-C. (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics*, 45(6):2654–2679.

Brillinger, D. R. (2001). *Time series: data analysis and theory*. SIAM.

Cheng, T.-C. F., Ing, C.-K., and Yu, S.-H. (2015). Toward optimal model averaging in regression models with time series errors. *Journal of Econometrics*, 189(2):321–334.

Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2):280–293.

Gao, Y., Zhang, X., Wang, S., and Zou, G. (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics*, 192(1):139–151.

Greenaway-McGrevy, R. (2015). Evaluating panel data forecasts under independent realization. *Journal of Multivariate Analysis*, 136:108–125.

Greenaway-McGrevy, R. (2019). Asymptotically efficient model selection for panel data forecasting. *Econometric Theory*, 35(4):842–899.

Greenaway-McGrevy, R. (2022). Forecast combination for vars in large n and t panels. *International Journal of Forecasting*, 38(1):142–164.

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

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

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

Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *The Annals of Statistics*, 35(3):1238–1277.

Ing, C.-K. (2020). Model selection for high-dimensional linear regression with dependent observations. *The Annals of Statistics*, 48(4):1959–1980.

Ing, C.-K. and Sin, C.-Y. (2006). On prediction errors in regression models with nonstationary regressors. *Lecture Notes-Monograph Series*, 60–71.

Ing, C.-K., Sin, C.-Y., and Yu, S.-H. (2010). Prediction errors in nonstationary autoregressions of infinite order. *Econometric Theory*, 26(3):774–803.

Ing, C.-K., Sin, C.-Y., and Yu, S.-H. (2012). Model selection for integrated autoregressive processes of infinite order. *Journal of Multivariate Analysis*, 106:57–71.

Ing, C.-K. and Wei, C.-Z. (2003). On same-realization prediction in an infinite-order autoregressive process. *Journal of Multivariate Analysis*, 85(1):130–155.

Ing, C.-K. and Wei, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *The Annals of Statistics*, 33(5):2423–2474.

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

Liao, J., Zou, G., Gao, Y., and Zhang, X. (2021). Model averaging prediction for time series models with a diverging number of parameters. *Journal of Econometrics*, 223(1):190–221.

Liao, J.-C. and Tsay, W.-J. (2020). Optimal multistep var forecast averaging. *Econometric Theory*, 36(6):1099–1126.

Liu, Q. and Okui, R. (2013). Heteroscedasticity-robust cp model averaging. *The Econometrics Journal*, 16(3):463–472.

Mallows, C. (1973). Some comments on C_p. *Technometrics*, 15(4):661–675.

Peng, J. and Yang, Y. (2022). On improvability of model selection by model averaging. *Journal of Econometrics*, 229(2):246–262.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, 147–164.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133.

Sun, Y., Hong, Y., Lee, T.-H., Wang, S., and Zhang, X. (2021). Time-varying model averaging. *Journal of Econometrics*, 222(2):974–992.

Wei, C.-Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *The Annals of Statistics*, 1667–1682.

Xu, W. and Zhang, X. (2022). From model selection to model averaging: A comparison for nested linear models. *arXiv preprint arXiv:2202.11978*.

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

Yang, Y. (2007). Prediction/estimation with simple linear models: Is it really that simple? *Econometric Theory*, 23(1):1–36.

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

Zhang, X. and Liu, C.-A. (2022). Model averaging prediction by k-fold cross-validation. *Journal of Econometrics, Forthcoming*.

Zhang, X., Wan, A. T., and Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, 174(2):82–94.

Appendix A: Mathematical Proof

In the Appendix, define:

$$
\begin{aligned}
B_{1n}(k,d) &= \left\{\frac{U_{n,n}^{'}(d)}{\sqrt{N}}\hat{\Omega}_n^{-1}(k)\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}U_{j,n}(d)\epsilon_{j+1,k-d}\right\}1(d\geq 1),\\
f_{1,n}(d) &= \left\{\frac{U_{n,n}^{'}(d)}{\sqrt{N}}\left[N^{-1}\sum_{j=K_n}^{n-\sqrt{n}-1}U_{j,n}(d)U_{j,n}^{'}(d)\right]^{-1}\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}U_{j,n}(d)\epsilon_{j+1}\right\}1(d\geq 1),\\
f_{1,n}^{*}(d) &= \left\{\frac{U_{n,n}^{*'}(d)}{\sqrt{N}}\left[N^{-1}\sum_{j=K_n}^{n-\sqrt{n}-1}U_{j,n}(d)U_{j,n}^{'}(d)\right]^{-1}\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}U_{j,n}(d)\epsilon_{j+1}\right\}1(d\geq 1),\\
B_{2n}(k-d) &= \left\{\frac{\mathbf{z}_n^{'}(k-d)}{\sqrt{N}}\Gamma^{-1}(k-d)\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}\mathbf{z}_j(k-d)\epsilon_{j+1,k-d}\right\}1(k>d),\\
f_{2,n}(k-d) &= \left\{\frac{\mathbf{z}_n^{'}(k-d)}{\sqrt{N}}\Gamma^{-1}(k-d)\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}\mathbf{z}_j(k-d)\epsilon_{j+1}\right\}1(k>d),\\
f_{2,n}^{*}(k-d) &= \left\{\frac{\mathbf{z}_n^{*'}(k-d)}{\sqrt{N}}\Gamma^{-1}(k-d)\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}\mathbf{z}_j(k-d)\epsilon_{j+1}\right\}1(k>d),\\
f_{2,n,\infty}^{*}(k-d) &= \left\{\frac{\mathbf{z}_n^{*'}(k-d)}{\sqrt{N}}\Gamma^{-1}(k-d)\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}\mathbf{z}_{j,\infty}(k-d)\epsilon_{j+1}\right\}1(k>d),\\
S_n^{*}(k-d) &= \sum_{i=1}^{\sqrt{n}/2}(a_i-a_i(k-d))z_{n+1-i}^{**},\ \ z_{n+1-i}^{**}=\sum_{j=0}^{\sqrt{n}/2}b_j\epsilon_{n+1-i-j},\\
U_{j,n}(v) &= \left(y_j(d)/N^{d-(1/2)},...,y_j(d-v+1)/N^{d-v+(1/2)}\right)^{'},\\
U_{n,n}^{*}(d) &= \left(N^{-d+1/2}\sum_{j=\sqrt{n}}^{n-1}\kappa_j(d)\epsilon_{n-j},...,N^{-1/2}\sum_{j=\sqrt{n}}^{n-1}\kappa_j(1)\epsilon_{n-j}\right)^{'},\\
\mathbf{z}_n^{*}(k) &= \left(\sum_{j=0}^{\sqrt{n}-K_n}b_j\epsilon_{n-j},...,\sum_{j=0}^{\sqrt{n}-K_n}b_j\epsilon_{n-k+1-j}\right)^{'},\ k\geq 1,\\
\kappa_j(1) &= \sum_{s=0}^{j}b_s,\ \ \text{and}\ \ \kappa_j(v)=\sum_{s=0}^{j}\kappa_s(v-1),\ \forall v\geq 2,\\
\Gamma(v) &= \mathrm{E}(\mathbf{z}_{t,\infty}(v)\mathbf{z}_{t,\infty}^{'}(v)),\ v\geq 1. \qquad \text{(A.1)}
\end{aligned}
$$

Lemma 1. For $K_n = o(n)$ and $0 \le k \le K_n$,

$$E\big(\sum_{k=0}^{K_n} w_k(\epsilon_{n+1,k} - \epsilon_{n+1})\big)^2 - \sum_{0 \le i,j \le K_n} w_i w_j \|a - a(\max\{i,j\})\|_z^2 = o(n^{-1}).$$

Proof.

$$\begin{aligned}
&E\big(\sum_{k=1}^{K_n} w_k(\epsilon_{n+1,k} - \epsilon_{n+1})\big)^2 - \sum_{0 \le i,j \le K_n} w_i w_j \|a - a(\max\{i,j\})\|_z^2 \\
&= \sum_{k=0}^{K_n} w_k^2 E(\epsilon_{n+1,k} - \epsilon_{n+1})^2 - \sum_{k=0}^{K_n} w_k^2 \|a - a(\max(k))\|_z^2 \\
&+ \sum_{k \ne l} w_k w_l \Big\{ E\big[(\epsilon_{n+1,k} - \epsilon_{n+1})(\epsilon_{n+1,l} - \epsilon_{n+1})\big] - \|a - a(\max\{k,l\})\|_z^2 \Big\} \\
&= (I) + (II).
\end{aligned}$$

Note that by Lemma B.5 of Ing et al. (2010), (I) is $o(n^{-1})$. It suffices to prove (II) is $o(n^{-1})$ as well. Denote $a_i - a_i(k)$ by $\gamma_i(k)$, and since

$$\epsilon_{n+1,k} - \epsilon_{n+1} = \sum_{i=1}^{n} \gamma_i(k) z_{n+1-i} = \sum_{i=1}^{n} \gamma_i(k)(z_{n+1-i} - z_{n+1-i,\infty}) + \sum_{i=1}^{n} \gamma_i(k) z_{n+1-i,\infty}.$$

Then,

$$|(II)| \le |(III)| + |(IV)| + |(V)| + |(VI)|,$$

where

$$\begin{aligned}
(III) &:= \sum_{k \ne l} w_k w_l \big\{ E\big[(\sum_{i=1}^{n} \gamma_i(k) z_{n+1-i,\infty})(\sum_{i=1}^{n} \gamma_i(l) z_{n+1-i,\infty})\big] - \|a - a(\max\{k,l\})\|_z^2 \big\} \\
(IV) &:= \sum_{k \ne l} w_k w_l E\Big\{ \big[\sum_{i=1}^{n} \gamma_i(k)(z_{n+1-i} - z_{n+1-i,\infty})\big]\big[\sum_{i=1}^{n} \gamma_i(l)(z_{n+1-i} - z_{n+1-i,\infty})\big] \Big\} \\
(V) &:= \sum_{k \ne l} w_k w_l E\Big\{ \big[\sum_{i=1}^{n} \gamma_i(k)(z_{n+1-i} - z_{n+1-i,\infty})\big]\big[\sum_{i=1}^{n} \gamma_i(l)(z_{n+1-i,\infty})\big] \Big\} \\
(VI) &:= \sum_{k \ne l} w_k w_l E\Big\{ \big[\sum_{i=1}^{n} \gamma_i(k)(z_{n+1-i,\infty})\big]\big[\sum_{i=1}^{n} \gamma_i(l)(z_{n+1-i} - z_{n+1-i,\infty})\big] \Big\}.
\end{aligned}$$

By Cauchy-Schwarz inequality and (B.17) of Ing et al. (2010),

$$|(IV)| = o(n^{-1}),$$

and since

$$z_{n+1-i,\infty} - z_{n+1-i} = \sum_{j=n-i}^{\infty} b_j \epsilon_{n+1-i-j},$$

then,

$$\begin{aligned}
&\sum_{k\neq l} w_k w_l E\Big\{\big[\sum_{i=1}^{n} \gamma_i(k)(z_{n+1-i,\infty})\big]\big[\sum_{i=1}^{n} \gamma_i(l)(z_{n+1-i} - z_{n+1-i,\infty})\big]\Big\} \\
=&\sum_{k\neq l} w_k w_l E\Big\{\big[\sum_{i=1}^{n} \gamma_i(k)(z_{n+1-i} - z_{n+1-i,\infty})\big]\big[\sum_{i=1}^{n} \gamma_i(l)(z_{n+1-i} - z_{n+1-i,\infty})\big]\Big\}.
\end{aligned}$$

Hence, by (IV)

$$|(V)| = o(n^{-1}), \quad |(VI)| = o(n^{-1}).$$

To show $|(III)| = o(n^{-1})$, first noting that

$$E\big[(\sum_{i=1}^{\infty} \gamma_i(k) z_{n+1-i,\infty})(\sum_{i=1}^{\infty} \gamma_i(l) z_{n+1-i,\infty})\big] = \|a - a(\max\{k,l\})\|_z^2,$$

by (3.2) of Ing and Wei (2003). Thus,

$$\begin{aligned}
|(III)| &= |\sum_{k\neq l} w_k w_l \big\{E\big[(\sum_{i=1}^{n} \gamma_i(k) z_{n+1-i,\infty})(\sum_{i=1}^{n} \gamma_i(l) z_{n+1-i,\infty})\big] - \|a - a(\max\{k,l\})\|_z^2\big\}| \\
&= |\sum_{k\neq l} w_k w_l \big\{E\big[(\sum_{i=1}^{n} \gamma_i(k) z_{n+1-i,\infty})(\sum_{i=1}^{n} \gamma_i(l) z_{n+1-i,\infty})\big]\big\} \\
&- E\big[(\sum_{i=1}^{\infty} \gamma_i(k) z_{n+1-i,\infty})(\sum_{i=1}^{\infty} \gamma_i(l) z_{n+1-i,\infty})\big]\big\}| \\
&\leq |\sum_{k\neq l} w_k w_l \big\{E\big[(\sum_{i=n+1}^{\infty} \gamma_i(k) z_{n+1-i,\infty})(\sum_{i=1}^{n} \gamma_i(l) z_{n+1-i,\infty})\big]\big\}| \\
&+ |\sum_{k\neq l} w_k w_l \big\{E\big[(\sum_{i=1}^{n} \gamma_i(k) z_{n+1-i,\infty})(\sum_{i=n+1}^{\infty} \gamma_i(l) z_{n+1-i,\infty})\big]\big\}| \\
&+ |\sum_{k\neq l} w_k w_l \big\{E\big[(\sum_{i=n+1}^{\infty} \gamma_i(k) z_{n+1-i,\infty})(\sum_{i=n+1}^{\infty} \gamma_i(l) z_{n+1-i,\infty})\big]\big\}| \\
&= (VII) + (VIII) + (IX).
\end{aligned}$$

By (2.2), $\sum_{j=1}^{\infty} |ja_j| < \infty$,

$$\begin{aligned}
& E\big[(\sum_{i=n+1}^{\infty} \gamma_i(k) z_{n+1-i,\infty})(\sum_{i=n+1}^{\infty} \gamma_i(l) z_{n+1-i,\infty})\big] \\
& = \chi_0 \sum_{j=n+1}^{\infty} a_j^2 + \sum_{n+1 \leq i,j,\, i\neq j}^{\infty} a_i a_j \chi_{|i-j|} \\
& = o(n^{-2}),
\end{aligned}$$

where $\chi_{i-j} = \mathrm{E}(z_{i,\infty} z_{j,\infty})$. Thus, $(IX) = o(n^{-1})$.

For $(VIII)$, choose $0 < \rho < 1$ such that $\rho n > K_n$. Then,

$$\begin{aligned}
(VIII) &= \gamma_1(k) \sum_{i=n+1}^{\infty} \gamma_i(k)\chi_{i-1} + \gamma_2(k) \sum_{i=n+1}^{\infty} \gamma_i(k)\chi_{i-2} +, ..., + \gamma_n(k) \sum_{i=n+1}^{\infty} \gamma_i(k)\chi_{i-n} \\
&= \gamma_1(k) \sum_{i=n+1}^{\infty} \gamma_i(k)\chi_{i-1} +, ..., + \gamma_{\rho n}(k) \sum_{i=n+1}^{\infty} \gamma_i(k)\chi_{i-\rho n} \\
&+ \gamma_{\rho n+1}(k) \sum_{i=n+1}^{\infty} \gamma_i(k)\chi_{i-(\rho n+1)} +, ..., + \gamma_n(k) \sum_{i=n+1}^{\infty} \gamma_i(k)\chi_{i-n}.
\end{aligned}$$

By (2.2),

$$\begin{aligned}
& \gamma_{\rho n+1}(k) \sum_{i=n+1}^{\infty} \gamma_i(k)\chi_{i-(\rho n+1)} +, ..., + \gamma_n(k) \sum_{i=n+1}^{\infty} \gamma_i(k)\chi_{i-n} \\
& \leq C(\sum_{i=n+1}^{\infty} |\gamma_i(k)|)(\sum_{i=\rho n+1}^{\infty} |\gamma_i(k)|) \\
& = o(n^{-2}),
\end{aligned} \tag{A.2}$$

$$\begin{aligned}
\chi_{n+1-(\rho n)} = \chi_{(1-\rho)n+1} &= \mathrm{E}(z_{t,\infty} z_{t-(1-\rho)n-1,\infty}) \\
&= \mathrm{E}\big[(\sum_{j=0}^{\infty} b_j \epsilon_{t-j})(\sum_{j=0}^{\infty} b_j \epsilon_{t-\rho)n-1-j})\big] \\
&\leq C \sum_{j=(1-\rho)n+1}^{\infty} |b_j| \\
&= o(n^{-1}),
\end{aligned}$$

and,

$$\begin{aligned}
&\gamma_1(k)\sum_{i=n+1}^{\infty}\gamma_i(k)\chi_{i-1}+,...,+\gamma_{\rho n}(k)\sum_{i=n+1}^{\infty}\gamma_i(k)\chi_{i-\rho n}\\
&\leq C(\rho n)(\sum_{i=n+1}^{\infty}|a_i|)(\sum_{j=(1-\rho)n+1}^{\infty}|b_j|)\\
&= o(n^{-1}). \qquad (A.3)
\end{aligned}$$

By $(A.2)$ and $(A.3)$, $(VIII) = o(n^{-1})$. Similarly, $(VII) = o(n^{-1})$ as well. Since $(I) - (IX)$ are $o(n^{-1})$, the statement of Lemma 1 holds.

□

Lemma 2. For $K_n^{\max\{4d-1,3\}} = o(n)$, $\max\{1, d\} \leq k \leq K_n$ and $\mathbf{w} \in \mathcal{H}_n^d$,

(i)

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_{2,n}(k-d)-f_{2,n}^*(k-d))\big]^2}{L_n^d(\mathbf{w})}\right| = 0,$$

(ii)

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_{2,n}^*(k-d)-f_{2,n,\infty}^*(k-d))\big]^2}{L_n^d(\mathbf{w})}\right| = 0,$$

where $f_{2,n}(k-d)$, $f_{2,n}^*(k-d)$, $f_{2,n,\infty}^*(k-d)$, are defined after (A.1).

Proof.

To show (i), it suffices to show

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\mathrm{E}\left|\frac{\sqrt{N}}{\sqrt{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_{2,n}(k-d)-f_{2,n}^*(k-d))\Big]\right|^2 = 0. \quad (A.4)$$

Observe that

$$|\sqrt{\frac{N}{k-d}}[f_{2,n}(k-d)-f_{2,n}^*(k-d)]| \leq |A_1(k-d)+A_2(k-d)|,$$

where

$$A_1(k-d) = \left\{(\mathbf{z}_n'(k-d)-\mathbf{z}_n^{*'}(k-d))\Gamma^{-1}(k-d)\frac{1}{\sqrt{N(k-d)}}\sum_{j=K_n}^{n-1}\mathbf{z}_j(k-d)\epsilon_{j+1}\right\}1(k>d),$$

$$A_2(k-d) = \left\{\mathbf{z}_n^{*'}(k-d)\Gamma^{-1}(k-d)\frac{1}{\sqrt{N(k-d)}}\sum_{j=n-\sqrt{n}}^{n-1}\mathbf{z}_j(k-d)\epsilon_{j+1}\right\}1(k>d).$$

For any $p \geq 2$, by Hölder inequality,

$$\mathrm{E}(|A_1(k-d)|^p) \leq \mathrm{E}(\|a_1(k-d)\|^{3p})^{1/3}\mathrm{E}(\|a_2(k-d)\|^{3p})^{1/3}\mathrm{E}(\|a_3(k-d)\|^{3p})^{1/3},$$

where

$$a_1(k-d) = \big(z_n - z_n^*, ..., z_{n-k+d+1} - z_{n-k+d+1}^*\big)' = \Big(\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j\epsilon_{n-j}, ..., \sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j\epsilon_{n-k+d+1-j}\Big)',$$

$$a_2(k-d) = \Gamma^{-1}(k-d), \quad a_3(k-d) = [N(k-d)]^{-1/2}\sum_{j=K_n}^{n-1}\mathbf{z}_j(k-d)\epsilon_{j+1}.$$

By Lemma B.3 of Ing et al. (2010), (2.3), and assumptions, for all $d < k \leq K_n$,

$$\mathrm{E}(\|a_1(k-d)\|^{3p}) \leq C\Big[(k-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j^2\Big]^{3p/2},$$

$$\mathrm{E}(\|a_2(k-d)\|^{3p}) \leq C, \quad \mathrm{E}(\|a_3(k-d)\|^{3p}) \leq C.$$

Hence,

$$\mathrm{E}(|A_1(k-d)|^p) \leq C\Big[(k-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j^2\Big]^{p/2} \leq C((K_n-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j^2)^{p/2}. \tag{A.5}$$

Similarly,

$$\mathrm{E}(|A_2(k-d)|^p) \leq \mathrm{E}(\|b_1(k-d)\|^{3p})^{1/3}\mathrm{E}(\|a_2(k-d)\|^{3p})^{1/3}\mathrm{E}(\|b_2(k-d)\|^{3p})^{1/3},$$

where

$$b_1(k-d) = \mathbf{z}_n^{*'}(k-d), \quad b_2(k-d) = [N(k-d)]^{-1/2}\sum_{j=n-\sqrt{n}}^{n-1}\mathbf{z}_j(k-d)\epsilon_{j+1}.$$

By Lemma B.3 of Ing et al. (2010),

$$\mathrm{E}(\|b_1(k-d)\|^{3p}) \leq C(k-d)^{3p/2}, \quad \mathrm{E}(\|a_3(k-d)\|^{3p}) \leq C(\sqrt{N})^{3p/2}.$$

$$\mathrm{E}(|A_2(k-d)|^p) \leq C(\frac{k-d}{\sqrt{N}})^{p/2} \leq C(\frac{K_n-d}{\sqrt{N}})^{p/2}. \tag{A.56}$$

Then by (A.5) and (A.6),

$$\mathrm{E}\Big|\Big(\sqrt{\frac{N}{i}}[f_{2,n}(i)-f_{2,n}^*(i)]\Big)\Big(\sqrt{\frac{N}{j}}[f_{2,n}(j)-f_{2,n}^*(j)]\Big)\Big|^p \leq C\Big\{[(K_n-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j^2]^{p/2} + (\frac{K_n-d}{\sqrt{N}})^{p/2}\Big\}, \tag{A.7}$$

and for any $\mathbf{w} \in \mathcal{H}_n^d$,

$$\frac{\sum_{\max(1,d)\le i,j\le K_n} w_{i-d}w_{j-d}\sqrt{i}\sqrt{j}}{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d} = \frac{\sum_{\max(1,d)\le i,j\le K_n} w_i w_j \sqrt{i-d}\sqrt{j-d}}{\sum_{\max(1,d)\le i,j\le K_n} w_i w_j \min\{i,j\}-d} \le \sqrt{K_n-d}. \tag{A.8}$$

Therefore, by (A.7) and (A.8),

$$\begin{aligned}
&\mathrm{E}\Big|\frac{\sqrt{N}}{\sqrt{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_{2,n}(k-d)-f^*_{2,n}(k-d))\Big]\Big|^2 \\
&\le C\frac{\sum_{\max(1,d)\le i,j\le K_n} w_i w_j\sqrt{i-d}\sqrt{j-d}}{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}\Big\{\Big[(K_n-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j^2\Big]+\Big(\frac{K_n-d}{\sqrt{N}}\Big)\Big\} \\
&\le C\sqrt{K_n-d}\Big\{\Big[(K_n-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j^2\Big]+\Big(\frac{K_n-d}{\sqrt{N}}\Big)\Big\} \\
&\le C\sum_{j=\sqrt{n}-K_n+1}^{\infty}|jb_j|^2+\sqrt{\frac{K_n^3}{N}}.
\end{aligned} \tag{A.9}$$

(A.4) holds by (A.9), (2.3) and Assumption 4.

To show (ii), it suffices to show

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\mathrm{E}\Big|\frac{\sqrt{N}}{\sqrt{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f^*_{2,n}(k-d)-f^*_{2,n,\infty}(k-d))\Big]\Big|^2=0. \tag{A.10}$$

$$f^*_{2,n}(k-d)-f^*_{2,n,\infty}(k-d)=\Big\{\frac{\mathbf{z}_n^{*'}(k-d)}{\sqrt{N}}\Gamma^{-1}(k-d)\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}\tilde{\mathbf{z}}_{j,\infty}(k-d)\epsilon_{j+1}\Big\}1(k>d),$$

where

$$\tilde{\mathbf{z}}_{t,\infty}(v)=\big(\tilde{z}_{t,\infty},\ldots,\tilde{z}_{t-v+1,\infty}\big)'=\big(z_{t,\infty}-z_t,\ldots,z_{t-v+1,\infty}-z_{t-v+1}\big)'=\Big(\sum_{j=t}^{\infty}b_j\epsilon_{t-j},\ldots,\sum_{j=t-v+1}^{\infty}b_j\epsilon_{t-v+1-j}\Big)'.$$

Then, by Hölder inequality, Lemma B.3 of Ing et al. (2010), and Lemma 2 of Wei (1987),

$$\begin{aligned}
&\mathrm{E}\Big|\sqrt{N}\sum_{k=\max\{1,d\}}^{K_n} w_k[f^*_{2,n}(k-d)-f^*_{2,n,\infty}(k-d)]\Big|^p\\
&\le \sum_{k=\max\{1,d\}}^{K_n} w_k\mathrm{E}\Big|\sqrt{N}[f^*_{2,n}(k-d)-f^*_{2,n,\infty}(k-d)]\Big|^p\\
&\le \sum_{k=\max\{1,d\}}^{K_n} w_k(\mathrm{E}\|\mathbf{z}^{*'}_n(k-d)\|^{3p})^{1/3}(\mathrm{E}\|\Gamma^{-1}(k-d)\|^{3p})^{1/3}(\mathrm{E}\|\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}\tilde{\mathbf{z}}_{j,\infty}(k-d)\epsilon_{j+1}\|^{3p})^{1/3}\\
&\le C\max_{1\le k\le K_n-d} k^{p/2}k^{p/2}(N^{-1}\sum_{t=K_n}^{n-\sqrt{n}-1}\mathrm{E}(\tilde{\mathbf{z}}_{t,\infty})^2)^{p/2}\\
&\le CK_n^p(N^{-1}\sum_{t=K_n}^{n-\sqrt{n}-1}\sum_{i=t}^{\infty}b_i^2)^{p/2}\\
&\le C(\sum_{i=K_n}^{\infty}|ib_i|^2)^{p/2}. \qquad (A.11)
\end{aligned}$$

Then, (A.10) holds by (A.11) and (2.3).

□

Lemma 3. For $K_n^2=o(n)$,

(i)

$$\lim_{n\to\infty}\max_{1\le k\le K_n}\left|\mathrm{E}(N(f^*_{2,n,\infty}(k))^2)-k\sigma^2\right|=0$$

(ii)

$$\lim_{n\to\infty}\max_{1\le k,l\le K_n}\left|\mathrm{E}(Nf^*_{2,n,\infty}(k)f^*_{2,n,\infty}(l))-\min(k,l)\sigma^2\right|=0 \qquad (A.12)$$

where $f^*_{2,n,\infty}(k-d)$, are define after (A.1).

Proof. We only prove (ii), since (i) is the special case of (ii). Without loss of generality, we assume that $k<l$. Define

$$\Gamma_{k,l}(0)=\mathrm{E}(\mathbf{z}_{t,\infty}(k)\mathbf{z}'_{t,\infty}(l))=\big(\Gamma(k),\Gamma(k,l-k)\big)$$

$$\Gamma_{l,k}(0)=\mathrm{E}(\mathbf{z}_{t,\infty}(l)\mathbf{z}'_{t,\infty}(k))=\begin{pmatrix}\Gamma(k)\\ \Gamma(l-k,k)\end{pmatrix},$$

$$\Gamma^*_{l,k}(0)=\mathrm{E}(\mathbf{z}^*_t(l)\mathbf{z}^{*'}_t(k))=\begin{pmatrix}\Gamma^*(k)\\ \Gamma^*(l-k,k)\end{pmatrix},$$

where

$$\Gamma^*(k) = \mathrm{E}(\mathbf{z}_t^*(k)\mathbf{z}_t^{*'}(k)),$$

is $k \times k$ matrix, $\Gamma_{k,l}(0)$ is $k \times l$ matrix, and $\Gamma_{l,k}(0)$, $\Gamma^*_{l,k}(0)$ are $l \times k$ matrices.

Then, observe that

$$\mathrm{E}(Nf^*_{2,n,\infty}(k)f^*_{2,n,\infty}(l)) = \mathrm{tr}(\Gamma^*_{l,k}(0)\Gamma^{-1}(k)\Gamma_{k,l}(0)\Gamma^{-1}(l))\,\frac{N-\sqrt{n}}{N}\sigma^2,$$

and by Woodbury matrix identity and partitioned matrix inversion formula,

$$\begin{aligned}
&\mathrm{tr}(\Gamma^*_{l,k}(0)\Gamma^{-1}(k)\Gamma_{k,l}(0)\Gamma^{-1}(l)) \\
&= \mathrm{tr}([\Gamma^*_{l,k}(0)-\Gamma_{l,k}(0)]\Gamma^{-1}(k)\Gamma_{k,l}(0)\Gamma^{-1}(l)) + \mathrm{tr}(\Gamma_{l,k}(0)\Gamma^{-1}(k)\Gamma_{k,l}(0)\Gamma^{-1}(l)) \\
&= \mathrm{tr}([\Gamma^*_{l,k}(0)-\Gamma_{l,k}(0)]\Gamma^{-1}(k)\Gamma_{k,l}(0)\Gamma^{-1}(l)) + \min(k,l) \\
&\le C\|\Gamma^{-1}(k)\|\mathrm{tr}(\Gamma^*(k)-\Gamma(k)) + \min(k,l) \\
&\le C\sum_{j=\sqrt{n}-K_n+1}^{\infty} jb_j^2 + \min(k,l)
\end{aligned}$$

$$|\mathrm{E}(Nf^*_{2,n,\infty}(k)f^*_{2,n,\infty}(l)) - \min(k,l)\sigma^2| \le C\Big(\frac{\sum_{j=\sqrt{n}-K_n+1}^{\infty}|jb_j|^2}{\sqrt{n}-K_n} + \frac{K_n}{\sqrt{n}}\Big). \tag{A.13}$$

Then, (A.12) holds by (A.13), (2.3), and Assumption 4.

□

Lemma 4. For $K_n^{\max\{4d-1,3\}} = o(n)$, $\max\{1,d\} \le k \le K_n$ and $\mathbf{w} \in \mathcal{H}_n^d$,

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))\big]}{L_n^d(\mathbf{w})}\right| = 0, \tag{A.14}$$

where $f_{2,n}(k-d)$ and $S_n(k-d)$ are defined after (A.1) and (3.1), respectively.

Proof. Define

$$f^\star_{2,n}(k-d) = \left\{\frac{\mathbf{z}_n^{\star'}(k-d)}{\sqrt{N}}\Gamma^{-1}(k-d)\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}\mathbf{z}_j(k-d)\epsilon_{j+1}\right\}1(k>d),$$

$$S_n^*(k-d) = \sum_{i=1}^{\sqrt{n}/2}(a_i - a_i(k-d))z^{**}_{n+1-i},$$

$$\mathbf{z}_n^\star(k) = \Big(\sum_{j=0}^{\sqrt{n}/2-K_n} b_j\epsilon_{n-j}, \ldots, \sum_{j=0}^{\sqrt{n}/2-K_n} b_j\epsilon_{n-k+1-j}\Big)', k \ge 1,$$

$$z^{**}_{n+1-i} = \sum_{j=0}^{\sqrt{n}/2} b_j\epsilon_{n+1-i-j}.$$

Since for all $1 \le u, v \le K_n - d$, ,$z_n^\star(u)$ is independent of $(S_n(v) - S_n^*(v), \sum_{j=K_n}^{n-\sqrt{n}-1} z_j(v)\epsilon_{j+1})$ and $\sum_{j=K_n}^{n-\sqrt{n}-1} z_j(u)\epsilon_{j+1}$ is independent from $(S_n^*(v), z_n^\star(v))$,

$$\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}^\star(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k[S_n(k-d) - S_n^*(k-d)])\big] = 0,$$
$$\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}^\star(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n^*(k-d))\big] = 0.$$

Then,

$$\begin{aligned}
&\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))\big] \\
&= \mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k[f_{2,n}(k-d) - f_{2,n}^\star(k-d)])(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))\big] \\
&+ \mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}^\star(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k[S_n(k-d) - S_n^*(k-d)])\big] \\
&+ \mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}^\star(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n^*(k-d))\big] \\
&= \mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k[f_{2,n}(k-d) - f_{2,n}^\star(k-d)])(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))\big],
\end{aligned}$$

therefore,

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_n^d} \Big| \frac{\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))\big]}{L_n^d(\mathbf{w})} \Big| \\
&\le \sup_{\mathbf{w}\in\mathcal{H}_n^d} \Big| \frac{\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k[f_{2,n}(k-d) - f_{2,n}^\star(k-d)])(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))\big]}{L_n^d(\mathbf{w})} \Big| \\
&\le \sup_{\mathbf{w}\in\mathcal{H}_n^d} \frac{\mathrm{E}^{1/2}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_{2,n}(k-d) - f_{2,n}^\star(k-d))\big]^2}{L_n^d(\mathbf{w})} \\
&\times \sup_{\mathbf{w}\in\mathcal{H}_n^d} \frac{\mathrm{E}^{1/2}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2}{L_n^d(\mathbf{w})} \\
&\le C\{\sum_{j=\sqrt{n}/2-K_n+1}^{\infty} |jb_j|^2 + \sqrt{\frac{K_n^3}{N}}\}^{1/2},
\end{aligned} \tag{A.15}$$

where the last inequality is insured by Lemma 1 and with the same argument as Lemma 2 (i). Then, (A.14) holds by (A.15), (2.3), and Assumption 4.

□

Lemma 5. For $K_n^{\max\{4d-1,3\}} = o(n)$, $\max\{1,d\} \le k \le K_n$ and $\mathbf{w} \in \mathcal{H}_n^d$,

$$\lim_{n\to\infty} \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_{1,n}(d) + \sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2}{L_n^d(\mathbf{w})} - 1 \right| = 0, \tag{A.16}$$

where $f_{1,n}(d)$, $f_{2,n}(k-d)$, and $S_n(k-d)$ are defined after (A.1) and (3.1).

Proof. Let

$$(I) = \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_{1,n}(d) + \sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2}{L_n^d(\mathbf{w})} - 1 \right|.$$

Then,

$$(I) \le (II) + (III) + (IV) + (V) + (VI) + (VII), \tag{A.17}$$

where

$$(II) = \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{\mathrm{E}(f_{1,n}(d))^2 - \frac{d(d+1)\sigma^2}{N}}{L_n^d(\mathbf{w})} \right|,$$

$$(III) = \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{\mathrm{E}(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d))^2 - \sigma^2 \sum_{\max\{1,d\}\le i,j\le K_n} w_i w_j (\min\{i,j\}-d)}{L_n^d(\mathbf{w})} \right|,$$

$$(IV) = \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{\mathrm{E}(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))^2 - \sum_{\max\{1,d\}\le i,j\le K_n} w_i w_j \|a - a(\max\{i,j\}-d)\|_z^2}{L_n^d(\mathbf{w})} \right|,$$

$$(V) = \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{2\mathrm{E}\big[f_{1,n}(d) \sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d)\big]}{L_n^d(\mathbf{w})} \right|,$$

$$(VI) = \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{2\mathrm{E}\big[f_{1,n}(d) \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]}{L_n^d(\mathbf{w})} \right|,$$

$$(VII) = \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{2\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d) \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]}{L_n^d(\mathbf{w})} \right|.$$

By Lemma 2 of Ing et al. (2010),

$$\lim_{n\to\infty} (II) = 0. \tag{A.18}$$

By Lemma 2 and 3,

$$\lim_{n\to\infty} (III) = 0. \tag{A.19}$$

By Lemma 1,

$$\lim_{n\to\infty}(IV)=0. \tag{A.20}$$

By Lemma 4,

$$\lim_{n\to\infty}(VII)=0. \tag{A.21}$$

By (B.37)- (B.39), Lemma 1, Lemma B.1, B.3 of Ing et al. (2010) and Hölder's inequality,

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\mathrm{E}\left|\frac{f_{1,n}(d)\sum_{k=\max\{1,d\}}^{K_n}w_kf_{2,n}(k-d)-f^*_{1,n}(d)\sum_{k=\max\{1,d\}}^{K_n}w_kf^*_{2,n}(k-d)}{L_n^d(\mathbf{w})}\right|=0$$

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\mathrm{E}\left|\frac{f_{1,n}(d)\sum_{k=\max\{1,d\}}^{K_n}w_kS_n(k-d)-f^*_{1,n}(d)\sum_{k=\max\{1,d\}}^{K_n}w_kS^*_n(k-d)}{L_n^d(\mathbf{w})}\right|=0,$$

and the facts that for all $d\le k\le K_n$

$$\mathrm{E}(f^*_{1,n}(d)f^*_{2,n}(k-d))=\mathrm{E}(f^*_{1,n}(d)S^*_n(k-d))=0.$$

We obtain

$$\lim_{n\to\infty}(V)=0, \tag{A.22}$$

and

$$\lim_{n\to\infty}(VI)=0. \tag{A.23}$$

By (A.17) - (A.23), (A.16) holds.

□

Lemma 6. For $K_n^{\max\{4d-1,3\}}=o(n)$, $\max\{1,d\}\le k\le K_n$ and $\mathbf{w}\in\mathcal{H}_n^d$,

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{E(f_n(k,d),S_n(k-d),\mathbf{w})-E(F_n(k,d),S_n(k-d),\mathbf{w})}{L_n^d(\mathbf{w})}\right|=0, \tag{A.24}$$

where

$$E(f_n(k,d),S_n(k-d),\mathbf{w})=\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n}w_kf_n(k)+\sum_{k=\max\{1,d\}}^{K_n}w_kS_n(k-d)\big]^2$$

$$E(F_n(k,d),S_n(k-d),\mathbf{w})=\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n}w_kF_n(k,d)+\sum_{k=\max\{1,d\}}^{K_n}w_kS_n(k-d)\big]^2,$$

$F_n(k,d)=f_{1,n}(d)+f_{2,n}(k-d)$. And $f_{1,n}(d)$, $f_{2,n}(k-d)$, $f_n(k)$, and $S_n(k-d)$ are defined after (A.1) and (3.1).

Proof. For any $\mathbf{w} \in \mathcal{H}_n^d$, $B_n(k-d) := B_{1n}(k,d) + B_{2n}(k-d)$, $B_{1n}(k,d)$, $B_{2n}(k-d)$ defined after (A.1).

$$
\begin{aligned}
&\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_n(k) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2 \\
&= \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_n(k) - \sum_{k=\max\{1,d\}}^{K_n} w_k B_n(k-d)\big]^2 \\
&+ \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k B_n(k-d) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2 \\
&+ \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_n(k) - \sum_{k=\max\{1,d\}}^{K_n} w_k B_n(k-d)\big]\big[\sum_{k=\max\{1,d\}}^{K_n} w_k B_n(k-d) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big] \\
&= (I) + (II) + (III).
\end{aligned} \tag{A.25}
$$

By Lemma B1, B3, B4, B6, Hölder's inequality, Theorem 1 (ii), (A.26), (A.28) of Ing et al. (2010),

$$
\begin{aligned}
\frac{(I)}{L_n^d(\mathbf{w})} &\le \frac{\sum_{k=\max\{1,d\}}^{K_n} w_k \mathrm{E}\big[f_n(k) - w_k B_n(k-d)\big]^2}{L_n^d(\mathbf{w})} \\
&\le \frac{C\sum_{k=\max\{1,d\}}^{K_n} w_k \frac{k}{N}\frac{k^2}{N}}{L_n^d(\mathbf{w})} \le \frac{C}{N L_n^d(\mathbf{w})}\frac{K_n^3}{N},
\end{aligned} \tag{A.26}
$$

$$
\begin{aligned}
(II) &= \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k B_n(k-d) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2 \\
&= \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k B_n(k-d) - \sum_{k=\max\{1,d\}}^{K_n} w_k F_n(k-d)\big]^2 \\
&+ \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k B_n(k) - \sum_{k=\max\{1,d\}}^{K_n} w_k F_n(k-d)\big]\big[\sum_{k=\max\{1,d\}}^{K_n} w_k F_n(k-d) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big] \\
&+ \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k F_n(k,d) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2 \\
&= (IV) + (V) + (VI).
\end{aligned} \tag{A.27}
$$

Since

$$(IV) \leq \sum_{k=\max\{1,d\}}^{K_n} w_k \mathrm{E}\big[f_{1,n}(d) - B_{1n}(k,d)\big]^2 + \sum_{k=\max\{1,d\}}^{K_n} w_k \mathrm{E}\big[f_{2,n}(k-d) - B_{2n}(k,d)\big]^2,$$

by (B.43) - (B.45) of Ing et al. (2010),

$$\frac{(IV)}{L_n^d(\mathbf{w})} \leq \frac{C}{NL_n^d(\mathbf{w})}, \tag{A.28}$$

and by Cauchy-Schwarz inequality, sufficiently large N, and Lemma 5,

$$\frac{(V)}{L_n^d(\mathbf{w})} \leq \frac{C}{(NL_n^d(\mathbf{w}))^{1/2}}. \tag{A.29}$$

By Cauchy-Schwarz inequality again and above decomposition of (II),

$$\frac{(III)}{L_n^d(\mathbf{w})} \leq \frac{C}{NL_n^d(\mathbf{w})}\left(\frac{K_n^3}{N}\right)^{1/2}. \tag{A.30}$$

Then,

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{E(f_n(k,d),S_n(k-d),\mathbf{w}) - E(F_n(k,d),S_n(k-d),\mathbf{w})}{L_n^d(\mathbf{w})}\right| \\
&= \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{E(f_n(k,d),S_n(k-d),\mathbf{w}) - (VI)}{L_n^d(\mathbf{w})}\right| \\
&\leq \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{(I)+(II)+(III)+(IV)+(V)}{L_n^d(\mathbf{w})}\right|.
\end{aligned} \tag{A.31}$$

Then, by (A.25) - (A.31), (A.24) holds.

□

Proof of Theorem 1. Observe that for any $\mathbf{w} \in \mathcal{H}_n^d$,

$$\mathrm{E}(y_{n+1} - \hat{y}_{n+1}(\mathbf{w}))^2 - \sigma^2 = \mathrm{E}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_n(k) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\Big]^2,$$

and

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_n(k) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2}{L_n^d(\mathbf{w})} - 1\right| \\
&\leq \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{E(f_n(k,d),S_n(k-d),\mathbf{w}) - E(F_n(k,d),S_n(k-d),\mathbf{w})}{L_n^d(\mathbf{w})}\right| \\
&+ \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_{1,n}(d) + \sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2}{L_n^d(\mathbf{w})} - 1\right|.
\end{aligned}$$

Then, Theorem 1 holds by Lemma 5 and 6.

□

Proof of Corollary 1. Without loss of generality, randomly choose two AR models, AR(k_1) and AR(k_2), where $\max(1, d) \leq k_1 < k_2 \leq K_n$, and consider conducting model averaging on these two models only. The associated $\mathbf{w}$ equals to $(0, ..., w_{k_1}, ..., w_{k_2}, ..., 0) \in \mathcal{H}_n^d$, and $w_{k_1} + w_{k_2} = 1$. Then, by Theorem 1, the model averaging predictor of AR(k_1) and AR(k_2) is strictly less than the model selection predictors if

$$\begin{aligned} & w_{k_1}^2 \frac{k_1}{N} + (1 - w_{k_1})^2 \frac{k_2}{N} + 2w_{k_1}(1 - w_{k_1}) \frac{k_1}{N} \\ & \qquad\qquad + \\ & w_{k_1}^2 \|\mathbf{a} - \mathbf{a}(k_1 - d)\|_z^2 + (1 - w_{k_1})^2 \|\mathbf{a} - \mathbf{a}(k_2 - d)\|_z^2 + 2w_{k_1}(1 - w_{k_1}) \|\mathbf{a} - \mathbf{a}(k_2 - d)\|_z^2 \\ & < \frac{k_1}{N} + \|\mathbf{a} - \mathbf{a}(k_1 - d)\|_z^2, \end{aligned} \tag{A.32}$$

and

$$\begin{aligned} & w_{k_1}^2 \frac{k_1}{N} + (1 - w_{k_1})^2 \frac{k_2}{N} + 2w_{k_1}(1 - w_{k_1}) \frac{k_1}{N} \\ & \qquad\qquad + \\ & w_{k_1}^2 \|\mathbf{a} - \mathbf{a}(k_1 - d)\|_z^2 + (1 - w_{k_1})^2 \|\mathbf{a} - \mathbf{a}(k_2 - d)\|_z^2 + 2w_{k_1}(1 - w_{k_1}) \|\mathbf{a} - \mathbf{a}(k_2 - d)\|_z^2 \\ & < \frac{k_2}{N} + \|\mathbf{a} - \mathbf{a}(k_2 - d)\|_z^2. \end{aligned} \tag{A.33}$$

By some algebraic manipulations, (A.32) and (A.33) can be reduced to

$$(1 - w_{k_1})^2 \, \frac{k_2 - k_1}{N} < (1 - w_{k_1}^2) \big[\, \|\mathbf{a} - \mathbf{a}(k_1 - d)\|_z^2 - \|\mathbf{a} - \mathbf{a}(k_2 - d)\|_z^2 \, \big], \tag{A.34}$$

and

$$w_{k_1}^2 \big[\, \|\mathbf{a} - \mathbf{a}(k_1 - d)\|_z^2 - \|\mathbf{a} - \mathbf{a}(k_2 - d)\|_z^2 \, \big] < 2w_{k_1}(1 - w_{k_1}) \frac{k_2 - k_1}{N}. \tag{A.35}$$

By the quadratic formula, (A.34) implies w_{k_1} must be within the following interval:

$$w_{k_1} \in \left(\frac{C(k_1, k_2, n) - B(k_1, k_2)}{C(k_1, k_2, n) + B(k_1, k_2)} \, , \, 1 \right),$$

and (A.35) implies w_{k_1} must be within the interval below:

$$w_{k_1} \in \left(0 \, , \, \frac{2\, C(k_1, k_2, n)}{C(k_1, k_2, n) + B(k_1, k_2)} \right),$$

where

$$C(k_1,k_2,n) := \frac{k_2-k_1}{N}, \quad B(k_1,k_2) := \|\mathbf{a}-\mathbf{a}(k_1-d)\|_z^2 - \|\mathbf{a}-\mathbf{a}(k_2-d)\|_z^2.$$

Since $C(k_1,k_2,n)$ is always greater than zero, if $B(k_1,k_2) > 0$, then

$$\left(\frac{C(k_1,k_2,n)-B(k_1,k_2)}{C(k_1,k_2,n)+B(k_1,k_2)}\ ,\ 1 \right) \cap \left(0\ ,\ \frac{2\,C(k_1,k_2,n)}{C(k_1,k_2,n)+B(k_1,k_2)} \right) \cap [0,1] \neq \emptyset,$$

there exists a weighting vector $\mathbf{w}^\circ_{k_1,k_2} := (0,...,w^\circ_{k_1},...,w^\circ_{k_2},...,0) \in \mathcal{H}^d_n$ with $w^\circ_{k_1} + w^\circ_{k_2} = 1$ such that

$$L^d_n(\mathbf{w}^\circ_{k_1,k_2}) < \min\left(L^d_n(\mathbf{w}_{k_1}), L^d_n(\mathbf{w}_{k_2})\right),$$

where $\mathbf{w}_{k_1}$ and $\mathbf{w}_{k_2}$ are vertices of $\mathcal{H}^d_n$ corresponding to model selection predictors of AR(k_1) and AR(k_2), respectively. Note that we do not restrict the relation between $C(k_1,k_2,n)$ and $B(k_1,k_2)$. Either $C(k_1,k_2,n) \geq B(k_1,k_2)$ or $C(k_1,k_2,n) \leq B(k_1,k_2)$ are allowed, where $C(k_1,k_2,n) \geq B(k_1,k_2)$ implies

$$\frac{k_2}{N} + \|\mathbf{a}-\mathbf{a}(k_2-d)\|_z^2 \geq \frac{k_1}{N} + \|\mathbf{a}-\mathbf{a}(k_1-d)\|_z^2,$$

AR(k_1) generates smaller predictive risk than AR(k_2) and vice versa.

By condition (3.4), since there is a $k \in \{\max(1,d),...,K_n\}$ such that $|B(k,l)| > 0$, $\forall\, l \neq k$. We can repeat the above argument for all the pairs of AR(k) and AR(l) with fixed k, $\max(1,d) \leq l \leq K_n, l \neq k$. Then, for $\mathcal{H}^d_n$, there are $K_n - \max(1,d)$ number of pairs and weight vectors either $\mathbf{w}^\circ_{k,l} := (0,...,w^\circ_k,...,w^\circ_l,...,0)$ if $k < l$ or $\mathbf{w}^\circ_{k,l} := (0,...,w^\circ_l,...,w^\circ_k,...,0)$ if $l < k$. Denote $\mathcal{P}_n(\mathbf{w}^\circ_{k,l})$ be the collection of weight vectors $\mathbf{w}^\circ_{k,l}$, and

$$\mathbf{w}^\diamond_n := \arg\min_{\mathcal{P}_n(\mathbf{w}^\circ_{k,l})} L^d_n(\mathbf{w}).$$

Clearly, $\mathbf{w}^\diamond_n$ in $\mathcal{H}^d_n / V(\mathcal{H}^d_n)$, and

$$L^d_n(\mathbf{w}^\diamond_n) < \min(L^d_n(\mathbf{w}_k), L^d_n(\mathbf{w}_l)),$$

for all the pairs of AR(k) and AR(l), $\max(1,d) \leq l \leq K_n, l \neq k$. Hence,

$$L^d_n(\mathbf{w}^\diamond_n) < \min_{\mathbf{w}\in\mathcal{V}(\mathcal{H}^d_n)} L^d_n(\mathbf{w}).$$

□

Proof of Theorem 2.

(i) For any $\mathbf{w} = (w_1, w_2, ..., w_{K_n}) \in H_n^d$, define $\varphi_1 = 1$ and $\varphi_j = \sum_{j=2}^{K_n} w_j$, with some algebraic manipulations, (3.2) can be rewritten as

$$L_n^d(\mathbf{w}) = \frac{\sigma^2 d^2}{N} + \frac{\sigma^2 \max(1,d)}{N} + A_{K_n} + \sum_{j=\max(1,d)+1}^{K_n} \varphi_j^2 \frac{\sigma^2}{N} + \sum_{j=\max(1,d)+1}^{K_n} (1-\varphi_j)^2 [A_{j-1} - A_j]. \tag{A.36}$$

Then, for $\mathbf{w}_n^* \in H_n^d$ such that $L_n^d(\mathbf{w}_n^*) = \inf_{\mathbf{w}\in H_n^d} L_n^d(\mathbf{w})$, $L_n^d(\mathbf{w}_n^*)$ can be obtained by plugging

$$\varphi_j = \sum_{j=\max(1,d)+1}^{K_n} \left(\frac{A_{j-1} - A_j}{\frac{\sigma^2}{N} + A_{j-1} - A_j} \right),$$

into (A.36).

(ii) Since

$$L_n^d(\mathbf{w}_{k_n^*}) = \frac{\sigma^2 d^2}{N} + \sigma^2 \frac{k_n^*}{N} + A_{k_n^*},$$

and by the exponential-decay condition the argument as (A.1)- (A.5) in Ing and Wei (2005), we can get

$$k_n^* = O(\frac{1}{\alpha}\log(N)), \quad L_n^d(\mathbf{w}_{k_n^*}) = O(\frac{\frac{1}{\alpha}\log(N)}{N}). \tag{A.37}$$

Then,

$$\begin{aligned} \Delta_n = L_n^d(\mathbf{w}_{k_n^*}) - L_n^d(\mathbf{w}_n^*) &= \sum_{j=\max(1,d)+1}^{k_n^*} \left[\frac{\sigma^2}{N} \left(1 - \frac{A_{j-1} - A_j}{\frac{\sigma^2}{N} + A_{j-1} - A_j} \right) \right] \\ &\quad + \sum_{j=k_n^*+1}^{K_n} \left[(A_{j-1} - A_j) \left(1 - \frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N} + A_{j-1} - A_j} \right) \right] \\ &= (I) + (II). \end{aligned} \tag{A.38}$$

To show $\Delta_n = o(L_n^d(\mathbf{w}_{k_n^*}))$, it is sufficient to show that

$$(I) = o(\frac{\log(N)}{N}), \text{ and } (II) = o(\frac{\log(N)}{N}).$$

Since d is finite by Assumption 1,

$$A_j = C\exp(-\alpha(j-d)) = C\exp(-\alpha(j)), \ \ A_{j-1} - A_j = C\frac{\exp(\alpha(j))}{1-\exp(-\alpha)}.$$

$$
\begin{aligned}
(I) &= \frac{\sigma^2}{N} \sum_{j=\max(1,d)+1}^{k_n^*} \left(\frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N} + A_{j-1} - A_j} \right) \leq C\big(\frac{\sigma^2}{N}\big)^2 \sum_{j=\max(1,d)+1}^{k_n^*} \left(\frac{1}{A_{j-1} - A_j} \right) \\
&\leq C\big(\frac{\sigma^2}{N}\big)^2 \big(\frac{1}{1-\exp(-\alpha)}\big) \frac{\exp(\alpha k_n^*) - \exp(\alpha \max(1,d))}{e-1} \\
&= O(\frac{1}{N}),
\end{aligned} \tag{A.39}
$$

$$
\begin{aligned}
(II) &= \sum_{j=k_n^*+1}^{K_n} \left[(A_{j-1} - A_j) \left(1 - \frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N} + A_{j-1} - A_j} \right) \right] = C \sum_{j=k_n^*+1}^{K_n} \left[\left(\frac{(A_{j-1} - A_j)^2}{\frac{\sigma^2}{N} + A_{j-1} - A_j} \right) \right] \\
&\leq C \sum_{j=k_n^*+1}^{K_n} \left(A_{j-1} - A_j \right) \\
&\leq C \frac{1}{1-\exp(-\alpha)} \times \left(\exp(-\alpha k_n^*) + \ldots + \exp(-\alpha(K_n - 1)) \right) \\
&\leq C \exp(-\alpha k_n^*) \left[1 - \exp(-\alpha(K_n - k_n^* + 1)) \right] \\
&= O(\frac{1}{N}),
\end{aligned} \tag{A.40}
$$

then, by (A.37)-(A.40), $\Delta_n = o(L_n^d(\mathbf{w}_{k_n^*}))$.
(iii) By the algebraic-decay condition and the argument as (A.9) in Ing and Wei (2005), we can get

$$
k_n^* = O(N^{1/(\alpha+1)}), \quad L_n^d(\mathbf{w}_{k_n^*}) = O(N^{-\alpha/(\alpha+1)}). \tag{A.41}
$$

Since $\Delta_n \leq L_n^d(\mathbf{w}_{k_n^*})$, to show $\Delta_n = \Theta(L_n^d(\mathbf{w}_{k_n^*}))$, it is sufficient to show that

$$
(I) \geq c\, L_n^d(\mathbf{w}_{k_n^*}),
$$

where c is a positive constant greater than zero. Since $A_j = C(j-d)^{-\alpha}$,

$$
\begin{aligned}
(I) &= \frac{\sigma^2}{N} \sum_{j=\max(1,d)+1}^{k_n^*} \left(\frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N} + A_{j-1} - A_j} \right) = \frac{\sigma^2}{N} \sum_{j=\max(1,d)+1}^{k_n^*} \left(\frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N} + C(j-1-d)^{-\alpha} - C(j-d)^{-\alpha}} \right) \\
&\geq C\frac{\sigma^2}{N} \sum_{j=\max(1,d)+1}^{k_n^*} \left(\frac{\sigma^2 (j-1-d)^{\alpha}}{\sigma^2 (j-1-d)^{\alpha} + N\big[1 - \big(1 - \frac{1}{j-d}\big)^{\alpha}\big]} \right) \\
&\geq C\frac{\sigma^2}{N} \sum_{j=\max(1,d)+1}^{k_n^*} \left(\frac{\sigma^2 (j-1-d)^{\alpha}(j-d)}{\sigma^2 (j-1-d)^{\alpha}(j-d) + N} \right) \\
&\geq C\frac{\sigma^2}{N} \sum_{j=\max(1,d)+1}^{k_n^*} \left(\frac{\sigma^2 (j-1-d)^{\alpha+1}}{\sigma^2 (k_n^*)^{\alpha+1} + N} \right) \\
&\geq C\frac{\sigma^2}{N} \frac{1}{\sigma^2 (k_n^*)^{\alpha+1} + N} \sum_{j=\max(1,d)+1}^{k_n^*} (j-1-d)^{\alpha+1} \\
&\geq C\frac{\sigma^2}{N} \frac{1}{\sigma^2 (k_n^*)^{\alpha+1} + N} \big[(k_n^* - 1 - d)^{\alpha+2} \big] \geq C\frac{k_n^*}{N} = CN^{-\alpha/(\alpha+1)}, \qquad \text{(A.42)}
\end{aligned}
$$

where the second inequality is insured by

$$1 - (1-x)^p \leq Cx, \quad p > 0, \quad 0 < x < 1,$$

and the last inequality is by $k_n^* = O(N^{1/(\alpha+1)})$. By (A.41), (A.42), $\Delta_n \leq L_n^d(\mathbf{w}_{k_n^*})$, $\Delta_n = \Theta(L_n^d(\mathbf{w}_{k_n^*}))$ under algebraic-decay scenario.

□

To prove the Lemmas 7-9, we define

$$\begin{aligned}
A &:= \sum_{1\le i,j\le d-1} w_i w_j \min(i,j),\\
A_d &:= \sum_{1\le i,j\le d-1} w_i w_j d,\\
B &:= \sum_{d\le i,j\le K_n} w_i w_j \min(i,j),\\
C &:= \sum_{1\le i,j\le d-1} w_i w_j \max(i,j),\\
D &:= \sum_{d\le i,j\le K_n} w_i w_j \max(i,j),\\
E &:= \sum_{1\le i,j\le d-1} w_i w_j \hat{\sigma}^2(\max(i,j)),\\
E_d &:= \sum_{1\le i,j\le d-1} w_i w_j \hat{\sigma}^2(d),\\
F &:= \sum_{d\le i,j\le K_n} w_i w_j \hat{\sigma}^2(\max(i,j)),
\end{aligned}$$

where $\hat{\sigma}^2(k) = \frac{1}{N}\sum_{t=K_n}^{n-1}(y_{t+1} - \hat{y}_{t+1}(k))^2 = \frac{1}{N}\sum_{t=K_n}^{n-1}(y_{t+1} + \mathbf{y}_t'\hat{\mathbf{a}}_n(k))^2$.

Lemma 7. For any $1 \le k < d$ and $2 < q_1 < q$,

$$\Pr(\hat{w}_{S_{n,k}} > 0) = O(n^{-q_1/2}), \tag{A.43}$$

where $\hat{w}_{S_{n,k}}$ is the kth element of $\hat{\mathbf{w}}_{S_n}$, the selected weight by Shibata model averaging criteria.

Proof. While $\hat{w}_{S_{n,k}} > 0$, it means that there exists some $\mathbf{w} = (w_1, ..., w_k, ..., w_{K_n}) \in \mathcal{H}_n$, such that

$$\hat{\mathbf{w}}_{S_n} = \mathbf{w},\ w_k > 0,\ \text{for any } 1 \le k < d,$$

$$
\begin{aligned}
\Pr(\hat{w}_{S_{n,k}} > 0, 1 \le k < d) &= \Pr(w_k > 0, 1 \le k < d) \\
&= \Pr\big([N + A + B + C + D] \times [E + F] \\
&\qquad \le [N + A_d + B + A_d + D] \times [E_d + F] \big) \\
&= \Pr\big(N[E - E_d] \le \\
&\qquad [(A_d - A) + (A_d - C)][E_d + F] + [A + B + C + D][E_d - E] \big) \\
&\le \Pr\big(N[E - E_d] \le [(A_d - A) + (A_d - C)][E_d + F] \big) \\
&\le \Pr\big(N[E - E_d] \le 2A_d \hat{\sigma}^2(d) \big) \\
&\le \Pr\big(N[\sum_{1 \le i,j \le d-1} w_i w_j (\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d))] \le 2 \sum_{1 \le i,j \le d-1} w_i w_j \, d \, \hat{\sigma}^2(d) \big) \\
&\le \Pr\big(N[\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d)] \le 2d \, \hat{\sigma}^2(d) \big), \qquad \text{(A.44)}
\end{aligned}
$$

where the first to third inequalities are due to the fact that $\hat{\sigma}^2(k) \le \hat{\sigma}^2(l)$ for all $l < k$ and $\sum_{k=1}^{K_n} w_k = 1$. So $E_d + F \le \hat{\sigma}^2(d)$, $E_d - E \le 0$, and $\sum_{1 \le i,j \le d-1} w_i w_j (\hat{\sigma}^2(d-1) \le E$. Then, the (A.43) holds by (4.30) and Theorem 4.5 of Ing et al. (2012), and (A.44).

□

Lemma 8. For any $1 \le k < d$ and $2 < q_1 < q$,

$$\Pr(\hat{w}_{A_{n,k}} > 0) = O(n^{-q_1/2}), \qquad \text{(A.45)}$$

where $\hat{w}_{A_{n,k}}$ is the kth element of $\hat{\mathbf{w}}_{A_n}$, the selected weight by AIC-like model averaging criteria.

Proof. While $\hat{w}_{A_{n,k}} > 0$, it means that there exists some $\mathbf{w} = (w_1, ..., w_k, ..., w_{K_n}) \in \mathcal{H}_n$, such that

$$\hat{\mathbf{w}}_{A_n} = \mathbf{w}, \ w_k > 0, \ \text{for any } 1 \le k < d,$$

$$
\begin{aligned}
\Pr(\hat{w}_{A_{n,k}} > 0, 1 \le k < d) &= \Pr(w_k > 0, 1 \le k < d) \\
&= \Pr\Big(\log(E+F) + \frac{(A+B+C+D)}{N} \\
&\qquad \le \log(E_d+F) + \frac{(A_d+B+A_d+D)}{N} \Big) \\
&= \Pr\Big(\log(E+F) - \log(E_d+F) \le \frac{[(A_d-A)+(A_d-C)]}{N} \Big) \\
&\le \Pr\Big(\log(\frac{E+F}{E_d+F}) \le \frac{2d}{N} \Big) \\
&= \Pr\Big(\frac{E+F}{E_d+F} \le \exp(\frac{2d}{N}) \Big) \\
&= \Pr\Big(\frac{E-E_d}{E_d+F} \le \exp(\frac{2d}{N}) - 1 \Big) \\
&\le \Pr\Big(\frac{E-E_d}{E_d+F} \le \frac{2d}{N-2d} \Big) \\
&= \Pr\Big((N-2d)(E-E_d) \le 2d(E_d+F) \Big) \\
&\le \Pr\Big((N-2d)[\sum_{1\le i,j\le d-1} w_i w_j (\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d))] \le 2d\,\hat{\sigma}^2(d) \Big) \\
&\le \Pr\Big((N-2d)[\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d)] \le C\,\hat{\sigma}^2(d) \Big), \qquad \text{(A.46)}
\end{aligned}
$$

where the second inequality is insured by $\exp(x) - 1 \le \frac{x}{1-x}$ if $0 < x < 1$, and the last two inequalities hold since $\hat{\sigma}^2(k) \le \hat{\sigma}^2(l)$ for all $l < k$, $\sum_{k=1}^{K_n} w_k = 1$, and $\sum_{1\le i,j\le d-1} w_i w_j > 0$ by assumption. So $E_d + F \le \hat{\sigma}^2(d)$, $\sum_{1\le i,j\le d-1} w_i w_j (\hat{\sigma}^2(d-1) \le E$, and there exists a constant C bounds the left term inside the last probability Then, the (A.45) holds by (4.30) and Theorem 4.5 of Ing et al. (2012), and (A.46).

□

Lemma 9. For any $1 \le k < d$ and $2 < q_1 < q$,

$$
\Pr(\hat{w}_{C_{n,k}} > 0) = O(n^{-q_1/2}), \qquad \text{(A.47)}
$$

where $\hat{w}_{C_{n,k}}$ is the kth element of $\hat{\mathbf{w}}_{C_n}$, the selected weight by Mallow's model averaging criteria.

Proof. By Lemma 4.1 and (4.6) of Ing et al. (2012), for any $k \to \infty$, $\hat{\sigma}^2(k)$ is a consistent estimator of σ^2. Without loss of generality, let $\breve{\sigma}^2 = \hat{\sigma}^2(K_n)$. While $\hat{w}_{C_{n,k}} > 0$, it means that there exists some $\mathbf{w} = (w_1, ..., w_k, ..., w_{K_n}) \in \mathcal{H}_n$, such that

$$
\hat{\mathbf{w}}_{C_n} = \mathbf{w},\ w_k > 0,\ \text{for any } 1 \le k < d,
$$

$$
\begin{aligned}
&= \Pr\big([N + A + B + C + D] \times [E + F] \\
&= \Pr\big(N(E + F) + (A + B + C + D)\check{\sigma}^2 \\
&\qquad \le N(E_d + F) + (A_d + B + A_d + D)\check{\sigma}^2 \big) \\
&= \Pr\big(N[E - E_d] \le [(A_d - A) + (A_d - C)]\check{\sigma}^2 \big) \\
&\le \Pr\big(N[E - E_d] \le 2A_d\check{\sigma}^2 \big) \\
&\le \Pr\big(N[\sum_{1\le i,j\le d-1} w_i w_j (\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d))] \le 2 \sum_{1\le i,j\le d-1} w_i w_j \, d\, \check{\sigma}^2 \big) \\
&\le \Pr\big(N[\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d)] \le 2d\, \check{\sigma}^2 \big) \\
&\le \Pr\big(N[\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d)] \le 2d\, (\sigma^2(K_n) + \epsilon) \big) + \Pr\big(|\check{\sigma}^2 - \sigma^2(K_n)| > \epsilon \big), \qquad \text{(A.48)}
\end{aligned}
$$

where the second and third inequalities are due to the fact that $\hat{\sigma}^2(k) \le \hat{\sigma}^2(l)$ for all $l < k$ and $\sum_{k=1}^{K_n} w_k = 1$. So $E_d - E \le 0$, and $\sum_{1\le i,j\le d-1} w_i w_j (\hat{\sigma}^2(d-1) \le E$. Then, the (A.47) holds by (4.30) and Theorem 4.5 of Ing et al. (2012), and (A.48).

□

Proof of Theorem 3.
(i) since

$$
\begin{aligned}
\hat{\mathbf{w}}_{S_n} &= \hat{\mathbf{w}}_{S_n} 1(\hat{\mathbf{w}}_{S_n} \in \mathcal{H}_n \backslash \mathcal{H}_n^d) + \hat{\mathbf{w}}_{S_n} 1(\hat{\mathbf{w}}_{S_n} \in \mathcal{H}_n^d) \\
&= \hat{\mathbf{w}}_{S_n} 1(\hat{\mathbf{w}}_{S_n} \in \mathcal{H}_n \backslash \mathcal{H}_n^d) + \hat{\mathbf{w}}_{S_n}^d,
\end{aligned}
$$

$$
\begin{aligned}
\Pr\big(\|\hat{\mathbf{w}}_{S_n} - \hat{\mathbf{w}}_{S_n}^d\|_2 \big) &= \Pr\big(\|\hat{\mathbf{w}}_{S_n}\|_2 1(\hat{\mathbf{w}}_{S_n} \in \mathcal{H}_n \backslash \mathcal{H}_n^d) \big) \\
&= \Pr\big((\sum_{k=1}^{K_n} \hat{w}_{S_{n,k}}^2)^{1/2}\, 1(\hat{\mathbf{w}}_{S_n} \in \mathcal{H}_n \backslash \mathcal{H}_n^d) \big) \\
&\le \Pr\big(\hat{\mathbf{w}}_{S_n} \in \mathcal{H}_n / \mathcal{H}_n^d \big) = \Pr(\hat{w}_{S_{n,k}} > 0, 1 \le k < d).
\end{aligned}
$$

By Lemma 7, we obtain

$$
\Pr\big(\|\hat{\mathbf{w}}_{S_n} - \hat{\mathbf{w}}_{S_n}^d\|_2 \big) = O(n^{-q_1/2}), \;\; 2 < q_1 < q,
$$

is summable and the claim holds by Borel-Cantelli Lemma.

(ii)

$\hat{\mathbf{w}}_{S_n}^d = \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} S_n(\mathbf{w})$, and $L_n^d(\mathbf{w}_n^*) = \inf_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w})$,

$$
\begin{aligned}
0 \geq S_n(\hat{\mathbf{w}}_{S_n}^d) - S_n(\mathbf{w}_n^*) &= NL_n^d(\hat{\mathbf{w}}_{S_n}^d) - NL_n^d(\mathbf{w}_n^*) - V_n(\hat{\mathbf{w}}_{S_n}^d, \mathbf{w}_n^*), \\
V_n(\hat{\mathbf{w}}_{S_n}^d, \mathbf{w}_n^*) &\geq NL_n^d(\hat{\mathbf{w}}_{S_n}^d) - NL_n^d(\mathbf{w}_n^*) \geq 0, \\
\sup_{\mathbf{w}\in\mathcal{H}_n^d} \Big| \frac{V_n(\mathbf{w}, \mathbf{w}_n^*)}{NL_n^d(\mathbf{w})} \Big| &\geq \frac{V_n(\hat{\mathbf{w}}_n, \mathbf{w}_n^*)}{NL_n^d(\hat{\mathbf{w}}_n)} \geq 1 - \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{S_n}^d)} \geq 0.
\end{aligned}
$$

Therefore, if

$$
\lim_{n\to\infty} \sup_{\mathbf{w}\in\mathcal{H}_n^d} \Big| \frac{V_n(\mathbf{w}, \mathbf{w}_n^*)}{NL_n^d(\mathbf{w})} \Big| \xrightarrow{p} 0, \tag{A.49}
$$

then,

$$
\lim_{n\to\infty} \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{S_n}^d)} \xrightarrow{p} 1.
$$

Inspired by (4.1) of Ing et al. (2012) and Theorem 1, for all $\mathbf{w} \in \mathcal{H}_n^d$, we decompose $S_n(\mathbf{w})$ as below:

$$
\begin{aligned}
S_n(\mathbf{w}) =& NL_n^d(\mathbf{w}) + \mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}(\hat{\sigma}_w^2 - \sigma^2) + \mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}(\hat{\sigma}_w^2 - \sigma^2) + (N + d - d^2)\sigma^2 \\
&+ \Bigg(\sum_{\max(1,d)\leq i,\, j\leq K_n} w_i w_j \big[(\max(i,j) - d)\sigma^2 \\
&\quad - \|N^{-1/2} \sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(\max(i,j))\epsilon_{j+1,\max(i,j)-d}\|^2_{\hat{\Omega}_n^{-1}(\max(i,j))}\big] \Bigg) \\
&+ \big(N \sum_{1\leq i,\, j\leq K_n} w_i w_j [\hat{\Sigma}_n^2(\max(i,j) - d) - \sigma^2(\max(i,j) - d)] \big),
\end{aligned} \tag{A.50}
$$

where $\hat{\Sigma}_n^2(l) = N^{-1}\sum_{j=K_n}^{n-1} \epsilon_{j+1,l}^2$, $\sigma^2(l) = \sigma^2 + \|\mathbf{a} - \mathbf{a}(l)\|_z^2$, and for vector $\mathbf{v}$ and positive definite matrix Q, $\|\mathbf{v}\|_Q^2 = \mathbf{v}'Q\mathbf{v}$. $\|\mathbf{a} - \mathbf{a}(k)\|_z^2$, $\epsilon_{j+1,k}$ and $\hat{\Omega}_n^{-1}(k)$ are defined after (2.4).

In view of (A.50), we can rewrite $\frac{S_n(\mathbf{w}) - S_n(\mathbf{w}_n^*)}{NL_n^d(\mathbf{w})}$ as

$$
\frac{S_n(\mathbf{w}) - S_n(\mathbf{w}_n^*)}{NL_n^d(\mathbf{w})} = 1 - \frac{NL_n^d(\mathbf{w}_n^*)}{NL_n^d(\mathbf{w})} - \frac{V_n(\mathbf{w}, \mathbf{w}_n^*)}{NL_n^d(\mathbf{w})},
$$

and $\frac{V_n(\mathbf{w},\mathbf{w}_n^*)}{NL_n^d(\mathbf{w})}$ can be decomposed into seven parts:

$$
\begin{aligned}
V_{1n}(\mathbf{w}) &= -\frac{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}(\hat{\sigma}_w^2-\sigma^2)}{NL_n^d(\mathbf{w})},\\
V_{2n}(\mathbf{w},\mathbf{w}_n^*) &= -\frac{\mathbf{w}_n^{*\prime}\Pi_{\min}(K_n)\mathbf{w}_n^*(\hat{\sigma}_{w_n^*}^2-\sigma^2)}{NL_n^d(\mathbf{w})},\\
V_{3n}(\mathbf{w}) &= -\frac{\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}(\hat{\sigma}_w^2-\sigma^2)}{NL_n^d(\mathbf{w})},\\
V_{4n}(\mathbf{w},\mathbf{w}_n^*) &= -\frac{\mathbf{w}_n^{*\prime}\Pi_{\max}(K_n)\mathbf{w}_n^*(\hat{\sigma}_{w_n^*}^2-\sigma^2)}{NL_n^d(\mathbf{w})},\\
V_{5n}(\mathbf{w}) &= -\frac{1}{NL_n^d(\mathbf{w})}\Bigg(\sum_{\max(1,d)\le i,\, j\le K_n} w_iw_j\big[\,(\max(i,j)-d)\sigma^2\\
&\qquad -\|N^{-1/2}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(\max(i,j))\epsilon_{j+1,\max(i,j)-d}\|^2_{\hat{\Omega}_n^{-1}(\max(i,j))}\,\big]\Bigg),\\
V_{6n}(\mathbf{w},\mathbf{w}_n^*) &= -\frac{1}{NL_n^d(\mathbf{w})}\Bigg(\sum_{\max(1,d)\le i,\, j\le K_n} w_{n,i}^*w_{n,j}^*\big[\,(\max(i,j)-d)\sigma^2\\
&\qquad -\|N^{-1/2}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(\max(i,j))\epsilon_{j+1,\max(i,j)-d}\|^2_{\hat{\Omega}_n^{-1}(\max(i,j))}\,\big]\Bigg),\\
V_{7n}(\mathbf{w},\mathbf{w}_n^*) &= -\frac{\sum_{\max(1,d)\le i,\, j\le K_n}(w_iw_j-w_{n,i}^*w_{n,j}^*)[\hat{\Sigma}^2(\max(i,j)-d)-\sigma^2(\max(i,j)-d)]}{L_n^d(\mathbf{w})},
\end{aligned}
$$

where

$$
\hat{\sigma}_{w^*}^2=\frac{1}{N}\sum_{t=K_n}^{n-1}(y_{t+1}-\hat{y}_{t+1}(\mathbf{w}_n^*))^2,
$$

$w_{n,k}^*$ is the kth element of $\mathbf{w}_n^*$.
Since

$$
\sup_{\mathbf{w}\in\mathcal{H}_n}\Big|\frac{V_n(\mathbf{w},\mathbf{w}_n^*)}{NL_n^d(\mathbf{w})}\Big|\le\sum_{i=1,3,5}\sup_{\mathbf{w}\in\mathcal{H}_n}|V_{in}(\mathbf{w})|+\sum_{j=2,4,6,7}\sup_{\mathbf{w}\in\mathcal{H}_n}|V_{jn}(\mathbf{w},\mathbf{w}_n^*)|,
$$

if $\sup_{\mathbf{w}\in\mathcal{H}_n^d}|V_{in}(\mathbf{w})|=o_p(1)$ for $i=1,3,5$ and $\sup_{\mathbf{w}\in\mathcal{H}_n^d}|V_{jn}(\mathbf{w},\mathbf{w}_n^*)|=o_p(1)$ for $j=2,4,6,7$, then (A.49) automatically satisfies.

Observe that

$$
\begin{aligned}
\hat{\sigma}_w^2 &= \frac{1}{N}\sum_{t=K_n}^{n-1}(y_{t+1}+\sum_{k=1}^{K_n}w_k\mathbf{y}_t'\hat{\mathbf{a}}_n(k))^2=\frac{1}{N}\sum_{t=K_n}^{n-1}(\sum_{k=1}^{K_n}w_k[y_{t+1}+\mathbf{y}_t'\hat{\mathbf{a}}_n(k)])^2,\\
&= \sum_{1\le i,\, j\le K_n}w_iw_j\hat{\sigma}^2(\max(i,j)),
\end{aligned}
$$

where $\hat\sigma^2(k) = \frac{1}{N}\sum_{t=K_n}^{n-1}(y_{t+1} - \hat y_{t+1}(k))^2 = \frac{1}{N}\sum_{t=K_n}^{n-1}(y_{t+1} + \mathbf{y}_t'\hat{\mathbf{a}}_n(k))^2$.
And by (4.6) of Ing et al. (2012), for any $k \geq \max(1, d)$,

$$\hat\sigma^2(k) - \sigma^2 = [\hat\Sigma_n^2(k-d) - \sigma^2(k-d)] - \|N^{-1}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(k)\epsilon_{j+1,k-d}\|^2_{\hat\Omega_n^{-1}(k))} + \|\mathbf{a} - \mathbf{a}(k-d)\|_z^2,$$

then,

$$\begin{aligned}
|V_{1n}(\mathbf{w})| &= \Big|\frac{(\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w})\ \sum_{\max(1,d)\leq i,\, j\leq K_n} w_i w_j[\hat\sigma^2(\max(i,j)) - \sigma^2]}{NL_n^d(\mathbf{w})}\Big| \\
&= (\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w})\Big|\frac{\sum_{\max(1,d)\leq i,\, j\leq K_n} w_i w_j[\hat\Sigma_n^2(\max(i,j)-d) - \sigma^2(\max(i,j)-d)]}{NL_n^d(\mathbf{w})}\Big| \\
&+ (\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}) \\
&\quad\times\Big|\frac{\sum_{\max(1,d)\leq i,\, j\leq K_n} w_i w_j\|N^{-1}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(\max(i,j))\epsilon_{j+1,\max(i,j)-d}\|^2_{\hat\Omega_n^{-1}(\max(i,j)))}}{NL_n^d(\mathbf{w})}\Big| \\
&+ (\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w})\Big|\frac{\sum_{\max(1,d)\leq i,\, j\leq K_n} w_i w_j\|\mathbf{a} - \mathbf{a}(\max(i,j)-d)\|_z^2}{NL_n^d(\mathbf{w})}\Big| \\
&= (I) + (II) + (III),
\end{aligned}$$

by Lemma 4.1 and (4.8) of Ing et al. (2012), $\sum_{\max(1,d)\leq i,\, j\leq K_n} w_i w_j = 1$,

$$\begin{aligned}
(I) &= O_p(\frac{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}}{NL_n^d(\mathbf{w})}\frac{1}{\sqrt{N}}) = O_p(\frac{1}{\sqrt{N}}), \\
(II) &= O_p(\frac{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}}{NL_n^d(\mathbf{w})}\frac{K_n}{N}) = O_p(\frac{K_n}{N}),
\end{aligned}$$

and

$$(III) \leq C\frac{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}}{N} \leq C\frac{K_n}{N}.$$

Then,

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d}|V_{1n}(\mathbf{w})| = O_p(\frac{1}{\sqrt{N}} + \frac{K_n}{N} + \frac{K_n}{N}). \tag{A.51}$$

Similarly,

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d}|V_{2n}(\mathbf{w}, \mathbf{w}_n^*)| \leq |V_{2n}(\mathbf{w}_n^*, \mathbf{w}_n^*)| \leq \sup_{\mathbf{w}\in\mathcal{H}_n^d}|V_{1n}(\mathbf{w})|.$$

Thus, by (A.51),

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d}|V_{2n}(\mathbf{w}, \mathbf{w}_n^*)| = O_p(\frac{1}{\sqrt{N}} + \frac{K_n}{N} + \frac{K_n}{N}). \tag{A.52}$$

Similar to $V_{1n}(\mathbf{w})$, we can rewrite

$$
\begin{aligned}
|V_{3n}(\mathbf{w})| &= \Big|\frac{(\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w})\ \sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j[\hat{\sigma}^2_{\max(i,j)} - \sigma^2]}{NL_n^d(\mathbf{w})}\Big| \\
&= (\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w})\Big|\frac{\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j[\hat{\Sigma}^2_n(\max(i,j)-d) - \sigma^2(\max(i,j)-d)]}{NL_n^d(\mathbf{w})}\Big| \\
&+ (\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}) \\
&\quad \times \Big|\frac{\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j \|N^{-1}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(\max(i,j))\epsilon_{j+1,\max(i,j)-d}\|^2_{\hat{\Omega}_n^{-1}(\max(i,j)))}}{NL_n^d(\mathbf{w})}\Big| \\
&+ (\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w})\Big|\frac{\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j \|\mathbf{a} - \mathbf{a}(\max(i,j)-d)\|^2_z}{NL_n^d(\mathbf{w})}\Big| \\
&= (I^*) + (II^*) + (III^*),
\end{aligned}
$$

by Lemma 4.1 and (4.8) of Ing et al. (2012), $\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j = 1$,

$$
(I^*) = O_p\Big(\frac{\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}}{NL_n^d(\mathbf{w})}\frac{1}{\sqrt{N}}\Big) = O_p\Big(\frac{1}{NL_n^d(\mathbf{w})}\frac{K_n}{\sqrt{N}}\Big),
$$
$$
(II^*) = O_p\Big(\frac{\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}}{NL_n^d(\mathbf{w})}\frac{K_n}{N}\Big) = O_p\Big(\frac{1}{NL_n^d(\mathbf{w})}\frac{K_n^2}{N}\Big),
$$

and

$$
(III^*) \le C\frac{\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}}{N} \le C\frac{K_n}{N}.
$$

Then,

$$
\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{3n}(\mathbf{w})| = O_p\Big(\frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n}{\sqrt{N}} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N} + \frac{K_n}{N}\Big). \tag{A.53}
$$

Similar to the argument on $V_{2n}(\mathbf{w}, \mathbf{w}_n^*)$,

$$
\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{4n}(\mathbf{w}, \mathbf{w}_n^*)| = O_p\Big(\frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n}{\sqrt{N}} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N} + \frac{K_n}{N}\Big). \tag{A.54}
$$

To deal with $V_{5n}(\mathbf{w})$, define

$$
\hat{\Omega}_{d,n}(k) = \begin{cases} \hat{\Omega}_n(k), & 1 \le k \le d, \\ \begin{pmatrix} \Gamma(k-d) & \mathbf{0}_{(k-d)\times d} \\ \mathbf{0}_{d\times(k-d)} & \hat{\Omega}_n(d) \end{pmatrix}, & d < k \le K_n. \end{cases}
$$

Then, for any $d \leq k \leq K_n$,

$$
\begin{aligned}
&\Big|(k-d)\sigma^2 - \|N^{-1/2}\sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(k)\epsilon_{j+1,k-d}\|^2_{\hat{\Omega}_n^{-1}(k)}\Big| \\
&\leq \Big|(k-d)\sigma^2 - \|N^{-1/2}\sum_{j=K_n}^{n-1} \mathbf{z}_j(k-d)\epsilon_{j+1,k-d}\|^2_{\Gamma^{-1}(k-d)}\Big| 1(k>d) \\
&+ \|N^{-1/2}\sum_{j=K_n}^{n-1} U_{j,n}(d)\epsilon_{j+1,k-d}\|^2 \|\hat{\Omega}_n^{-1}(d)\| \\
&+ \|N^{-1/2}\sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(k)\epsilon_{j+1,k-d}\|^2 \|\hat{\Omega}_n^{-1}(k) - \hat{\Omega}_{d,n}(k)\|,
\end{aligned}
$$

$$
\begin{aligned}
|V_{5n}(\mathbf{w})| \leq & \frac{1}{NL_n^d(\mathbf{w})}\Big| \sum_{\max(1,d)\leq i,\, j\leq K_n} w_i w_j \big[(\max(i,j)-d)\sigma^2 \\
& - \|N^{-1/2}\sum_{j=K_n}^{n-1} \mathbf{z}_j(\max(i,j)-d)\epsilon_{j+1,\max(i,j)-d}\|^2_{\Gamma^{-1}(\max(i,j)-d)} \big] 1(\max(i,j)>d)\Big| \\
& + \frac{1}{NL_n^d(\mathbf{w})}\Big(\sum_{\max(1,d)\leq i,\, j\leq K_n} w_i w_j \|N^{-1/2}\sum_{j=K_n}^{n-1} U_{j,n}(d)\epsilon_{j+1,\max(i,j)-d}\|^2 \|\hat{\Omega}_n^{-1}(d)\|\Big) \\
& + \frac{1}{NL_n^d(\mathbf{w})}\Big(\sum_{\max(1,d)\leq i,\, j\leq K_n} w_i w_j \|N^{-1/2}\sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(\max(i,j))\epsilon_{j+1,\max(i,j)-d}\|^2 \\
& \times \|\hat{\Omega}_n^{-1}(\max(i,j)) - \hat{\Omega}_{d,n}(\max(i,j))\| \Big) \\
& = (I^\circ) + (II^\circ) + (III^\circ).
\end{aligned}
$$

By (2.3), Lemma 4.2 of Ing et al. (2012), Lemmas B.1, B.3, B.4, B.6, and Theorem 1 of Ing et al. (2010), and some algebraic manipulation,

$$
\begin{aligned}
(I^\circ) &= O_p(\frac{K_n^{1/2}}{NL_n^d(\mathbf{w})}), \\
(II^\circ) &= O_p(\frac{1}{NL_n^d(\mathbf{w})}), \\
(III^\circ) &= O_p(\frac{1}{NL_n^d(\mathbf{w})}\frac{K_n^2}{N^{1/2}}),
\end{aligned}
$$

$$
\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{5n}(\mathbf{w})| = O_p(\frac{K_n^{1/2}}{NL_n^d(\mathbf{w})} + \frac{1}{NL_n^d(\mathbf{w}_n^*)} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N^{1/2}}). \tag{A.55}
$$

Similarly,

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{6n}(\mathbf{w},\mathbf{w}_n^*)| = O_p(\frac{K_n^{1/2}}{NL_n^d(\mathbf{w})} + \frac{1}{NL_n^d(\mathbf{w}_n^*)} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N^{1/2}}). \tag{A.56}$$

Since $\sum_{k=\max(1,d)}^{K_n} w_k = 1$ and $\sum_{k=\max(1,d)}^{K_n} w_k^* = 1$, we can decompose

$$\begin{aligned}
|V_{7n}(\mathbf{w},\mathbf{w}_n^*)| \leq & \Big|\frac{\sum_{\max(1,d)\leq i\, j\leq K_n} w_i w_j[\hat{\Sigma}^2(\max(i,j)) - \sigma^2(\max(i,j)) - \{\hat{\Sigma}^2(K_n) - \sigma^2(K_n)\}]}{L_n^d(\mathbf{w})}\Big| \\
& + \Big|\frac{\sum_{\max(1,d)\leq i\, j\leq K_n} w_{n,i}^* w_{n,j}^*[\hat{\Sigma}^2(\max(i,j)) - \sigma^2(\max(i,j)) - \{\hat{\Sigma}^2(K_n) - \sigma^2(K_n)\}]}{L_n^d(\mathbf{w})}\Big| \\
\leq & \Big|\frac{\sum_{\max(1,d)\leq i\, j\leq K_n} w_i w_j[\hat{\Sigma}^2(\max(i,j)) - \sigma^2(\max(i,j)) - \{\hat{\Sigma}^2(K_n) - \sigma^2(K_n)\}]}{L_n^d(\mathbf{w})}\Big| \\
& + \Big|\frac{\sum_{\max(1,d)\leq i\, j\leq K_n} w_{n,i}^* w_{n,j}^*[\hat{\Sigma}^2(\max(i,j)) - \sigma^2(\max(i,j)) - \{\hat{\Sigma}^2(K_n) - \sigma^2(K_n)\}]}{L_n^d(\mathbf{w}_n^*)}\Big|).
\end{aligned}$$

By (4.4) of Ing et al. (2012),

$$\begin{aligned}
|V_{7n}(\mathbf{w},\mathbf{w}_n^*)| &= O_p(\frac{\sum_{\max(1,d)\leq i\, j\leq K_n} w_i w_j \|\mathbf{a}(\max(i,j)) - \mathbf{a}(K_n)\|_z}{N^{1/2}L_n^d(\mathbf{w})}) \\
&\leq O_p(\frac{\sum_{\max(1,d)\leq i\, j\leq K_n} w_i w_j \|\mathbf{a} - \mathbf{a}(\max(i,j))\|_z}{(L_n^d(\mathbf{w}))^{1/2}} \frac{1}{(NL_n^d(\mathbf{w}))^{1/2}}) \\
&\leq O_p(\frac{1}{(NL_n^d(\mathbf{w}))^{1/2}}),
\end{aligned}$$

then,

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{7n}(\mathbf{w},\mathbf{w}_n^*)| = O_p(\frac{1}{(NL_n^d(\mathbf{w}_n^*))^{1/2}}). \tag{A.57}$$

Using (A.51)-(A.57), $\lim_{n\to\infty} N\eta_n^d \to \infty$, and Assumptions 4 and 5, we have

$$\sup_{\mathbf{w}\in\mathcal{H}_n} \Big|\frac{V_n(\mathbf{w},\mathbf{w}_n^*)}{NL_n^d(\mathbf{w})}\Big| \leq \sum_{i=1,3,5} \sup_{\mathbf{w}\in\mathcal{H}_n} |V_{in}(\mathbf{w})| + \sum_{j=2,4,6,7} \sup_{\mathbf{w}\in\mathcal{H}_n} |V_{jn}(\mathbf{w},\mathbf{w}_n^*)| = o_p(1).$$

Thus, (A.49) is satisfied and (ii) of Theorem 2 holds.

□

Lemma 10.

Suppose there is another model averaging weights criterion $\tilde{S}_n(\mathbf{w})$, which is a function of

model averaging weights.
Define

$$G_n(\mathbf{w}) = S_n(\mathbf{w}) - g(\tilde{S}_n(\mathbf{w})),$$

where $g(.)$ is a increasing function, and $S_n(\mathbf{w})$ is the Shibata model averaging criterion. Assume Assumption 1-4 hold. If

$$\eta_n^d = \inf_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w}) = L_n^d(\mathbf{w}_n^*);\ \lim_{n\to\infty} N\eta_n \to \infty,$$

and

$$\lim_{n\to\infty} \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left|\frac{G_n(\mathbf{w}) - G_n(\mathbf{w}_n^*)}{NL_n^d(\mathbf{w})}\right| \xrightarrow{p} 0, \tag{A.58}$$

then,

$$\lim_{n\to\infty} \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{\tilde{S}_n}^d)} \xrightarrow{p} 1,$$

where

$$\hat{\mathbf{w}}_{\tilde{S}_n}^d = \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} \tilde{S}_n(\mathbf{w}),\quad L_n^d(\mathbf{w}_n^*) = \inf_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w}).$$

Lemma 10 is quite similar to Theorem 4.2 in Shibata (1980) but under more general framework (model average setting).

Proof.
Since $\hat{\mathbf{w}}_{\tilde{S}_n}^d = \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} \tilde{S}_n(\mathbf{w})$, $L_n^d(\mathbf{w}_n^*) = \inf_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w})$, and $g(.)$ is increasing function. Then,

$$\begin{aligned}
0 &\geq g(\tilde{S}_n(\hat{\mathbf{w}}_{\tilde{S}_n}^d)) - g(\tilde{S}_n(\mathbf{w}_n^*)) = S_n(\hat{\mathbf{w}}_{\tilde{S}_n}^d) - G_n(\hat{\mathbf{w}}_{\tilde{S}_n}^d) - (S_n(\mathbf{w}_n^*) - G_n(\mathbf{w}_n^*)) \\
&= NL_n^d(\hat{\mathbf{w}}_{\tilde{S}_n}^d) - NL_n^d(\mathbf{w}_n^*) - V_n(\hat{\mathbf{w}}_{\tilde{S}_n}^d, \mathbf{w}_n^*) - (G_n(\hat{\mathbf{w}}_{\tilde{S}_n}^d) - G_n(\mathbf{w}_n^*))), \\
&(G_n(\hat{\mathbf{w}}_{\tilde{S}_n}^d) - G_n(\mathbf{w}_n^*)) + V_n(\hat{\mathbf{w}}_{\tilde{S}_n}^d, \mathbf{w}_n^*) \geq NL_n^d(\hat{\mathbf{w}}_{\tilde{S}_n}^d) - NL_n(\mathbf{w}_n^*) \geq 0, \\
&\sup_{\mathbf{w}\in\mathcal{H}_n^d} \left|\frac{G_n(\mathbf{w}) - G_n(\mathbf{w}_n^*))}{NL_n^d(\mathbf{w})}\right| + \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left|\frac{V_n(\mathbf{w}, \mathbf{w}_n^*)}{NL_n^d(\mathbf{w})}\right| \geq \frac{V_n(\hat{\mathbf{w}}_{\tilde{S}_n}^d, \mathbf{w}_n^*)}{NL_n^d(\hat{\mathbf{w}}_{\tilde{S}_n}^d)} \geq 1 - \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{\tilde{S}_n}^d)} \geq 0.
\end{aligned}$$

Then, by (A.49) and (A.58), we can get

$$\lim_{n\to\infty} \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{\tilde{S}_n}^d)} \xrightarrow{p} 1.$$

□

Proof of Theorem 4.

For Mallow's model averaging (MMA) criterion:

Without loss of generality, let $\check{\sigma}^2 = \hat{\sigma}^2(K_n)$. The selected weights of Mallow's model averaging $\hat{\mathbf{w}}_{C_n}$, $\hat{\mathbf{w}}_{C_n} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} C_n(\mathbf{w})$ satisfies (4.1) by the arguments as the proofs of (i) of Theorem 2 and Lemma 9.

To show $\hat{\mathbf{w}}^d_{C_n} := \arg\min_{\mathbf{w}\in\mathcal{H}^d_n} C_n(\mathbf{w})$ satisfying (4.2), we check the condition (A.46) in Lemma 10 holds. The difference between Shibata model averaging (SMA) criterion and Mallow's model averaging (MMA) C_n is

$$G_n(\mathbf{w}) = S_n(\mathbf{w}) - C_p(\mathbf{w}) = (\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2 - \hat{\sigma}^2_w) + N\check{\sigma}^2.$$

Then,

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{H}^d_n} |G_n(\mathbf{w}) - G_n(\mathbf{w}^*_n)| \le &\sup_{\mathbf{w}\in\mathcal{H}^d_n} |(\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2 - \hat{\sigma}^2_w)| \\ &+ \sup_{\mathbf{w}\in\mathcal{H}^d_n} |(\mathbf{w}^{*'}_n[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}^*_n)(\check{\sigma}^2 - \hat{\sigma}^2_{w^*})|. \quad (A.59)\end{aligned}$$

First, we will show

$$\sup_{\mathbf{w}\in\mathcal{H}^d_n} \Big|\frac{(\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2 - \hat{\sigma}^2_w)}{NL^d_n(\mathbf{w})}\Big| = o_p(1). \quad (A.60)$$

Since

$$\begin{aligned}&\sup_{\mathbf{w}\in\mathcal{H}^d_n} \Big|\frac{(\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2 - \hat{\sigma}^2_w)}{NL^d_n(\mathbf{w})}\Big| \\ \le &\sup_{\mathbf{w}\in\mathcal{H}^d_n} \Big|\frac{(\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w})(\sum_{\max(1,d)\le i,\, j\le K_n} w_iw_j\hat{\sigma}^2(\max(i,j)) - \sigma^2)}{NL^d_n(\mathbf{w})}\Big| \\ &+ \sup_{\mathbf{w}\in\mathcal{H}^d_n} \Big|\frac{(\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w})(\sum_{\max(1,d)\le i,\, j\le K_n}[\check{\sigma}^2 - \sigma^2])}{NL^d_n(\mathbf{w})}\Big| \\ &+ \sup_{\mathbf{w}\in\mathcal{H}^d_n} \Big|\frac{(\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w})(\sum_{\max(1,d)\le i,\, j\le K_n} w_iw_j\hat{\sigma}^2(\max(i,j)) - \sigma^2)}{NL^d_n(\mathbf{w})}\Big| \\ &+ \sup_{\mathbf{w}\in\mathcal{H}^d_n} \Big|\frac{(\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w})(\sum_{\max(1,d)\le i,\, j\le K_n}[\check{\sigma}^2 - \sigma^2])}{NL^d_n(\mathbf{w})}\Big| \\ &\le C\big(\sup_{\mathbf{w}\in\mathcal{H}^d_n} |V_{1n}(\mathbf{w})| + |\check{\sigma}^2 - \sigma^2| + \sup_{\mathbf{w}\in\mathcal{H}^d_n} |V_{3n}(\mathbf{w})| + \frac{\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}}{NL^d_n(\mathbf{w})}|\check{\sigma}^2 - \sigma^2|\big) \\ &\le C\Big(\sup_{\mathbf{w}\in\mathcal{H}^d_n} |V_{1n}(\mathbf{w})| + |\check{\sigma}^2 - \sigma^2| + \sup_{\mathbf{w}\in\mathcal{H}^d_n} |V_{3n}(\mathbf{w})| + O_p\big(\frac{1}{NL^d_n(\mathbf{w}^*_n)}\frac{K_n}{\sqrt{N}} + \frac{1}{NL^d_n(\mathbf{w}^*_n)}\frac{K^2_n}{N} + \frac{K_n}{N}\big)\Big),\end{aligned}$$

where the last inequality is insured by Lemma 4.1 and (4.8) of Ing et al. (2012), and similar arguments for (A.53). Then, by (A.51), (A.53), and $\check{\sigma}^2 = \hat{\sigma}^2(K_n)$ is consistent of σ^2, (A.60) holds.

And since

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}_n^{*'}[\Pi_{\min}(K_n)+\Pi_{\max}(K_n)]\mathbf{w}_n^*)(\check{\sigma}^2-\hat{\sigma}^2_{w^*})}{NL_n^d(\mathbf{w})}\Big| \\
&\leq \sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}_n^{*'}[\Pi_{\min}(K_n)+\Pi_{\max}(K_n)]\mathbf{w}_n^*)(\check{\sigma}^2-\hat{\sigma}^2_{w^*})}{NL_n^d(\mathbf{w}^*)}\Big| \\
&\leq \sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}'[\Pi_{\min}(K_n)+\Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2-\hat{\sigma}^2_{w})}{NL_n^d(\mathbf{w})}\Big|.
\end{aligned}$$

Then, by (A.60)

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}_n^{*'}[\Pi_{\min}(K_n)+\Pi_{\max}(K_n)]\mathbf{w}_n^*)(\check{\sigma}^2-\hat{\sigma}^2_{w^*})}{NL_n^d(\mathbf{w})}\Big| = o_p(1). \tag{A.61}$$

So, by (A.59)-(A.61), the difference of Shibata model averaging (SMA) criterion and Mallow's model averaging (MMA) satisfies the condition (A.58) of Lemma 10, implies $\hat{\mathbf{w}}^d_{C_n} := \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} C_n(\mathbf{w})$ satisfies (4.2). Thus, the model averaging weights selected by Mallow's model averaging (MMA) shares the property of asymptotically optimal without the integration order information in (4.1) and (4.2).

For Akaike model averaging (AMA) criterion:

Now we prove the model average weight selected by Akaike model averaging (AMA) criterion also shares the property of asymptotic efficiency in (4.1) and (4.2). First, the selected weights of Akaike model averaging $\hat{\mathbf{w}}_{A_n}$, $\hat{\mathbf{w}}_{C_n} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} C_n(\mathbf{w})$ satisfies (4.1) by the arguments as the proofs of (i) of Theorem 2 and Lemma 8.

To prove (4.2), we will check the condition (A.58) in Lemma 10 holds.
Set $g(x) = N\exp(x)$, then, the difference between Shibata's condition and transformation of AMA can be shown:

$$\begin{aligned}
G_n(\mathbf{w}) &= S_n(\mathbf{w}) - g(A_n(\mathbf{w})) \\
&= N\hat{\sigma}^2_w\Big(1+\frac{\mathbf{w}'[\Pi_{\min}(K_n)+\Pi_{\max}(K_n)]\mathbf{w}}{N} - \exp(\frac{\mathbf{w}'[\Pi_{\min}(K_n)+\Pi_{\max}(K_n)]\mathbf{w}}{N})\Big).
\end{aligned}$$

Then, for sufficiently large n,

$$
\begin{aligned}
|G_n(\mathbf{w}) - G_n(\mathbf{w}_n^*)| \leq & \Big| N\hat{\sigma}_w^2 (1 + \frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N} - \exp(\frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N})) \Big| \\
& + \Big| N\hat{\sigma}_{w^*}^2 (1 + \frac{\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^{*'}}{N} - \exp(\frac{\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*}{N})) \Big| \\
\leq & \Big| N\hat{\sigma}_w^2 (\frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N})^2 \Big| + \Big| N\hat{\sigma}_{w^*}^2 (\frac{\mathbf{w}_n^*[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*}{N})^2 \Big| \\
\leq & \Big| N(\hat{\sigma}_w^2 - \sigma^2)(\frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N})^2 \Big| \\
& + \Big| N\sigma^2 (\frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N})^2 \Big| \\
& + \Big| N(\hat{\sigma}_{w^*}^2 - \sigma^2)(\frac{\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*}{N})^2 \Big| \\
& + \Big| N\sigma^2 (\frac{\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*}{N})^2 \Big|, \qquad \text{(A.62)}
\end{aligned}
$$

where the second inequality holds by $|1 + x - \exp(x)| \leq |x|^2$ if $|x| \leq 1$. Then, by (A.62),

$$
\begin{aligned}
\sup_{\mathbf{w} \in \mathcal{H}_n^d} \Big| \frac{G_n(\mathbf{w}) - G_n(\mathbf{w}_n^*))}{NL_n^d(\mathbf{w})} \Big| \leq & C(\sup_{\mathbf{w} \in \mathcal{H}_n^d} |V_{1n}(\mathbf{w})| + \sup_{\mathbf{w} \in \mathcal{H}_n^d} |V_{2n}(\mathbf{w}, \mathbf{w}_n^*)| \\
& + \sup_{\mathbf{w} \in \mathcal{H}_n^d} |V_{3n}(\mathbf{w})| + \sup_{\mathbf{w} \in \mathcal{H}_n^d} |V_{4n}(\mathbf{w}, \mathbf{w}_n^*)| + \frac{K_n^2}{N} \frac{1}{NL_n^d(\mathbf{w}_n^*)}). \qquad \text{(A.63)}
\end{aligned}
$$

By (A.51)-(A.54), (A.63) and Assumption 4,

$$
\sup_{\mathbf{w} \in \mathcal{H}_n^d} \Big| \frac{G_n(\mathbf{w}) - G_n(\mathbf{w}_n^*))}{NL_n^d(\mathbf{w})} \Big| = o_p(1),
$$

the difference of SMA and transformation of AMA satisfies the condition (A.58) in Lemma 10. Hence, $\hat{\mathbf{w}}_{A_n}^d := \arg\min_{\mathbf{w} \in \mathcal{H}_n^d} A_n(\mathbf{w})$ satisfies (4.2). Thus, the model averaging weights selected by Akaike model averaging (AMA) criterion can achieve asymptotic efficiency in (4.1) and (4.2).

□

Proof of Corollary 2. Since

$$\begin{aligned}\frac{\hat{\Delta}_n}{L_n^d(\hat{\mathbf{w}}_{MS_n}^d)} &= 1 - \frac{L_n^d(\hat{\mathbf{w}}_{MAC_n}^d)}{L_n^d(\mathbf{w}_n^*)}\frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\mathbf{w}_{k_n^*})}\frac{L_n^d(\mathbf{w}_{k_n^*})}{L_n^d(\hat{\mathbf{w}}_{MS_n}^d)}\\ &= \frac{\Delta_n}{L_n^d(\mathbf{w}_{k_n^*})} + \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\mathbf{w}_{k_n^*})}\left(1 - \frac{L_n^d(\hat{\mathbf{w}}_{MAC_n}^d)}{L_n^d(\mathbf{w}_n^*)}\frac{L_n^d(\mathbf{w}_{k_n^*})}{L_n^d(\hat{\mathbf{w}}_{MS_n}^d)}\right)\\ &= \frac{\Delta_n}{L_n^d(\mathbf{w}_{k_n^*})} + o(1),\end{aligned}$$

where the last equality is insured by the asymptotic assumptions and $L_n^d(\mathbf{w}_n^*) = \Theta(L_n^d(\mathbf{w}_{k_n^*}))$. Thus, by Theorem 2,

$$\hat{\Delta}_n = o(L_n^d(\hat{\mathbf{w}}_{MS_n}^d)), \quad \hat{\Delta}_n = \Theta(L_n^d(\hat{\mathbf{w}}_{MS_n}^d))$$

under exponential and algebraic decay, respectively.

To prove

$$L_n^d(\hat{\mathbf{w}}_{MAC_n}^d) = \Theta(L_n^d(\hat{\mathbf{w}}_{MS_n}^d)),$$

observe that,

$$\frac{L_n^d(\hat{\mathbf{w}}_{MAC_n}^d)}{L_n^d(\hat{\mathbf{w}}_{MS_n}^d)} = \frac{L_n^d(\hat{\mathbf{w}}_{MAC_n}^d)}{L_n^d(\mathbf{w}_n^*)}\frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\mathbf{w}_{k_n^*})}\frac{L_n^d(\mathbf{w}_{k_n^*})}{L_n^d(\hat{\mathbf{w}}_{MS_n}^d)}.$$

By the asymptotic assumptions, it is sufficient to show

$$L_n^d(\mathbf{w}_n^*) = \Theta(L_n^d(\mathbf{w}_{k_n^*})).$$

By (3.6),

$$\begin{aligned}L_n^d(\mathbf{w}_n^*) &> \sum_{j=\max(1,d)+1}^{K_n} \frac{\frac{\sigma^2}{N}(A_{j-1}-A_j)}{\frac{\sigma^2}{N} + A_{j-1} - A_j} > \sum_{j=\max(1,d)+1}^{k_n^*} \frac{\frac{\sigma^2}{N}(A_{j-1}-A_j)}{\frac{\sigma^2}{N} + A_{j-1} - A_j}\\ &> C\sigma^2\frac{k_n^*}{N} \geq cL_n^d(\mathbf{w}_{k_n^*}),\end{aligned}$$

for some $c > 0$, where the third inequality holds by $\sigma^2/N < C(A_{j-1}-A_j)$, $j = \max(1,d)+1, ..., k_n^*$ for some large enough C under exponential and algebraic decay. Hence,

$$L_n^d(\mathbf{w}_n^*) = \Theta(L_n^d(\mathbf{w}_{k_n^*})).$$

□