Model Averaging Prediction for Possibly Nonstationary Autoregressions*

Tzu-Chi Lin† and Chu-An Liu‡

September 27, 2023

Abstract

As an alternative to model selection (MS), this paper considers model averaging (MA) for integrated autoregressive processes of infinite order. We derive a uniformly asymptotic expression for the mean squared prediction error (MSPE) of the averaging prediction with fixed weights and then propose a Mallows-type criterion to select the data-driven weights that minimize the MSPE asymptotically. We show that the proposed MA estimator and its variants, Shibata and Akaike MA estimators, are asymptotically optimal in the sense of achieving the lowest possible MSPE. We further demonstrate that MA can provide significant MSPE reduction over MS when the model misspecification bias is algebraic decay. These theoretical findings are supported by Monte Carlo simulations and real data analysis.

Keywords: Asymptotic improvability, Asymptotic optimality, Integrated autoregressive processes, Model averaging

JEL Classification: C22, C52, C53

*We thank the conference participants of STSC 2023, AMES 2023, EcoSta 2023, and COMPSTAT 2023 for their discussions and suggestions. All errors remain the authors'.

†Federal Reserve Bank of Philadelphia, 10 N Independence Mall W, Philadelphia, PA 19106, USA. Email: Simon.X.Lin@phil.frb.org.

‡Institute of Economics, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei, 115, Taiwan. Email: caliu@econ.sinica.edu.tw.

1 Introduction

Model selection (MS) has a long history in time series analysis. For an autoregressive process of infinite order (AR(∞)), estimation and inference are based on an approximation autoregressive model of order k. In the literature, several methods have been proposed to select the order k to achieve the best prediction, for example, the final prediction error (Akaike, 1970), Mallows' C_p (Mallows, 1973), Akaike information criterion (Akaike, 1974), and Shibata information criterion (Shibata, 1980). However, for a small perturbation of data, the MS method may choose a different model and result in diverse estimates and predictions. In addition, the selected model might lose some useful information contained in other models, and hence neglect the uncertainty across different models.

As an alternative to MS, model averaging (MA) is a smoothed extension of MS. Instead of choosing a single model, MA incorporates all available information by effectively averaging over all candidate models. The two main MA approaches are Bayesian model averaging and frequentist model averaging; see Claeskens et al. (2008), Moral-Benito (2015), and Steel (2020) for literature reviews of both approaches. In the past two decades, there is a rapidly growing development of frequentist MA methods. The central questions of concern in frequentist MA methods are how to assign the weights for candidate models, what are the theoretical properties of the selected weights, and does MA provide a significant improvement over MS? This paper studies the MA approaches and deals with these issues in a general autoregressive model.

In this paper, we consider the MA prediction for integrated autoregressive processes of infinite order. Our model framework is general enough to include the standard ARIMA(p, d, q) process as a special case, and can be applied to many stationary and nonstationary time series analysis. The main goal of this paper is to construct a one-step-ahead prediction based on a sequence of finite-order approximation AR(k) models for $1 \leq k \leq K_n$, where the maximum order K_n can go to infinity with the sample size n. Since the true data generating process is a dth-order integrated AR(∞), all candidate models are misspecified. Instead of the MS approach studied in Ing et al. (2010, 2012), we adopt an MA approach to construct an averaging prediction and study its asymptotic and finite-sample properties.

We first derive a uniformly asymptotic expression for the mean squared prediction error (MSPE) of the averaging prediction with fixed weights and demonstrate the bias-variance trade-off for the MA approach. We show that the MSPE of the averaging prediction can be decomposed into three components: the nonstationary estimation effect term, model complexity term, and model misspecification term. This result is not a trivial extension of Ing et al. (2010), because we need to take the interaction effect between any two candidate models into account when we characterize the model complexity and model misspecification of the MA approach. Based on the MSPE decomposition, we propose a Mallows-type criterion

to select the data-driven weights for nonstationary autoregressions of infinite order. The novel feature of the proposed method is that we introduce a penalty term to account for the model complexity and model misspecification simultaneously using both the minimum and maximum of the autoregressive orders between any two candidate models.

We provide two theoretical justifications for the proposed MA estimator: asymptotic optimality and asymptotic improvability. For asymptotic optimality, we first show that the proposed averaging prediction asymptotically assigns zero weight to the candidate model with the autoregressive order k less than the integration order d. We then show that without knowing the integration order, the proposed averaging prediction is asymptotically optimal in the sense of achieving the lowest possible MSPE in the class of MA estimators. This optimal result extends the asymptotic optimality of Ing et al. (2012) from MS to MA. In addition to the proposed Mallows-type criterion, we also extend Akaike (1974)'s and Shibata (1980)'s MS methods to the MA prediction and demonstrate the asymptotic optimality of these two related MA methods.

In the existing frequentist MA studies, a great amount of numerical evidence has shown that MA tends to perform better than MS in finite samples. However, little work has been done on examining the potential MSPE reduction of MA compared to MS. Recently, Peng and Yang (2022) and Xu and Zhang (2022) demonstrate that MA can provide significant squared prediction risk reduction over MS in nested linear models with orthonormal basis functions and linear nested regression models, respectively. In this paper, we relax the non-stochastic regression design of these two papers and establish the asymptotic improvability for a general autoregressive model. We first show that if there exists a candidate model whose misspecification bias is different from that of other models, then we can find at least one weight vector such that the MA has smaller MSPE than MS. We further show that when the model misspecification bias is algebraic decay, the MA methods can achieve significant MSPE reduction over the MS counterparts, but the magnitude of improvement is asymptotically negligible in the exponential-decay case. We demonstrate that asymptotic improvability holds for both fixed weights and data-driven weights.

In simulations, we examine the finite sample performance of the Mallows MA estimator and its variants, Shibata and Akaike MA estimators, in both algebraic-decay and exponential-decay cases. Monte Carlo simulations show that these MA methods perform quite well and produce similar empirical MSPEs. Compared with MS methods, the MA methods generally achieve lower empirical MSPEs in both cases. As the sample size increases, we can observe significant MSPEs improvement from MS to MA in the algebraic-decay case, but not in the exponential-decay case. Therefore, the simulation results are consistent with our theoretical findings. As an empirical illustration, we apply the proposed MA methods to the climate change prediction. Our empirical results show that the MA methods have lower empirical

MSPEs than the MS methods, but the improvement diminishes as the rolling window size increases. These findings are quite similar to those of the exponential-decay case in our simulation study.

For the frequentist MA approaches, numerous methods of weight selection have been proposed based on distinct criteria, for example, information criterion weighting (Buckland et al., 1997; Hjort and Claeskens, 2003), adaptive regression by mixing models (Yang, 2000, 2001; Yuan and Yang, 2005), Mallows model averaging (Hansen, 2007; Wan et al., 2010; Liu and Okui, 2013), jackknife model averaging (Hansen and Racine, 2012; Lu and Su, 2015; Ando and Li, 2014, 2017), plug-in averaging (Liu, 2015; Charkhi et al., 2016; Cheng et al., 2019), and others. In the time series context, the MS methods with asymptotic optimality have been studied in Shibata (1980), Ing and Wei (2003, 2005), Ing (2007, 2020), Ing et al. (2010, 2012), and Greenaway-McGrevy (2015, 2019). For the MA methods, asymptotic optimality has been investigated for the linear regression model with lagged dependent variables (Zhang et al., 2013), the regression model with time series errors (Cheng et al., 2015), the factor-augmented regression model (Cheng and Hansen, 2015), the longitudinal data model (Gao et al., 2016), the vector autoregressive (VAR) model (Liao et al., 2019; Liao and Tsay, 2020), the stationary AR(∞) process (Liao et al., 2021), the time-varying parameter regression models (Sun et al., 2021), and panel data VAR model (Greenaway-McGrevy, 2022). Most of these studies, however, are limited to the stationary or local stationary time series, which might not be applicable for data with non-stationary patterns in economics, finance, or climate change. Furthermore, there is no asymptotic comparison available between MS and MA in these studies.

The rest of this paper is organized as follows. Section 2 presents the model framework and the MA prediction. Section 3 introduces the Mallows model averaging criterion. Section 4 presents the uniformly asymptotic expression for the MSPE, asymptotic optimality, and asymptotic improvability. Section 5 discusses the related MA methods. Section 6 examines the finite sample properties of the proposed method. Section 7 provides the empirical study, and Section 8 concludes the paper. Proofs are included in the Appendix. Throughout this paper, we employ the following symbols. We use C to denote some positive constant that is independent of the sample size n, and C may represent different values in different equations. Let $\xrightarrow{p}$ and $\xrightarrow{a.s.}$ represent convergence in probability and almost surely, respectively. Let $\|\mathbf{v}\|_2$ be the Euclidean norm for vector $\mathbf{v}$ and $\|A\|^2 = \lambda_{\max}(A'A)$ be the maximum eigenvalue of matrix $A'A$.

2 Model Framework

In this paper, we follow the model setup of Ing et al. (2010, 2012) and assume that the observations $\{y_1, ..., y_n\}$ are generated from a dth-order integrated AR(∞) process as below:

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

where L is the backshift operator, $0 \leq d < \infty$ is an unknown integer, and $\{\epsilon_t, t = 0, \pm 1, \pm 2, ...\}$ are independent random variables with mean zero and variance σ^2. Note that ϵ_t does not necessarily come from the same distribution. This dth-order integrated AR(∞) process (2.1) includes the standard ARIMA(p, d, q) process as a special case and is general enough to be applied to many stationary and nonstationary time series analysis.

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

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

By Theorem 3.8.4 of Brillinger (2001), the stationary component (2.2) yields

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

Here, we follow Ing et al. (2010, 2012) and impose the initial condition $y_t = 0$ for $t \leq 0$.

Our goal is to construct a one-step-ahead prediction of y_{n+1} given the observed data $\{y_1, ..., y_n\}$. We consider a sequence of finite-order approximation models AR(1),..., AR(K_n), where the maximum order K_n can go to infinity with the sample size n. Because the true data generating process is AR(∞), each AR(k) model is misspecified and just an approximation model. For each AR(k) model, $1 \leq k \leq K_n$, the least squares estimator is defined as:

$$-\hat{\mathbf{a}}(k) = \left[\sum_{j=K_n}^{n-1} \mathbf{y}_j(k) \mathbf{y}_j^{'}(k) \right]^{-1} \sum_{j=K_n}^{n-1} \mathbf{y}_j(k) y_{j+1}, \tag{2.4}$$

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. Note that the asymptotic properties of least squares estimators of integrated autoregressive processes with a finite integration order have been well studied; see Kawashima (1980), Tiao and Tsay (1983), and Chan and Wei (1988). 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) model, the one-step-ahead prediction of y_{n+1} is

$$\hat{y}_{n+1}(k) = -\mathbf{y}_n^{'}(k) \hat{\mathbf{a}}(k). \tag{2.5}$$

We now consider the one-step-ahead averaging prediction. Let w_k be the weight corresponding to the AR(k) model and $\mathbf{w} = (w_1, ..., w_{K_n})'$ be a weight vector with $w_k \geq 0$ and $\sum_{k=1}^{K_n} w_k = 1$. That is, 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 predicted values of $\hat{y}_{n+1}(k)$, we construct an averaging prediction as

$$\hat{y}_{n+1}(\mathbf{w}) = \sum_{k=1}^{K_n} w_k \hat{y}_{n+1}(k). \tag{2.6}$$

3 Mallows Model Averaging Criterion

In this section, we propose a Mallows-type criterion to select the model weights for the averaging prediction for possibly nonstationary autoregressions. Define

$$\Pi_{\min}(K_n) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \dots & K_n \end{pmatrix} \text{ and } \Pi_{\max}(K_n) = \begin{pmatrix} 1 & 2 & \dots & K_n \\ 2 & 2 & \dots & K_n \\ \vdots & \vdots & \ddots & \vdots \\ K_n & K_n & \dots & K_n \end{pmatrix}, \tag{3.1}$$

where the (i,j)th element of $\Pi_{\min}(K_n)$ and $\Pi_{\max}(K_n)$ are $\min(i,j)$ and $\max(i,j)$ for $1 \leq i,\, j \leq K_n$, respectively.

The proposed averaging criterion is defined as:

$$C_n(\mathbf{w}) = N\hat{\sigma}_w^2 + (\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w} - N)\check{\sigma}^2, \tag{3.2}$$

where $N = n - K_n$, $\hat{\sigma}_w^2 = N^{-1}\sum_{t=K_n}^{n-1}(y_{t+1} - \hat{y}_{t+1}(\mathbf{w}))^2$, and $\check{\sigma}^2$ is some consistent estimator of σ^2 that does not depend on $\mathbf{w}$. For example, $\check{\sigma}^2$ can be constructed by $\hat{\sigma}^2(K_n) = N^{-1}\sum_{t=K_n}^{n-1}(y_{t+1} - \hat{y}_{t+1}(K_n))^2$. The data-driven weights based on the proposed averaging criterion are defined as

$$\hat{\mathbf{w}}_{\text{MMA}} = \arg\min_{\mathbf{w}\in\mathcal{H}_n} C_n(\mathbf{w}), \tag{3.3}$$

and the proposed one-step-ahead averaging prediction for y_{n+1} is

$$\hat{y}_{n+1}(\hat{\mathbf{w}}_{\text{MMA}}) = \sum_{k=1}^{K_n} \hat{w}_{\text{MMA},k}\hat{y}_{n+1}(k). \tag{3.4}$$

Observe that the proposed Mallows-type criterion $C_n(\mathbf{w})$ is a quadratic function of the weight vector. Therefore, the data-driven weights can be computed numerically via quadratic programming, and numerical algorithms of quadratic programming are available for most programming languages.

Note that $\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}$ and $\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}$ are derived from the MSPE of the averaging prediction, and these two terms characterize the model complexity and model misspecification, respectively; see the discussion after Theorem 1. Furthermore, since $\sum_{k=1}^{K_n} w_k = 1$, we have

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

Therefore, minimizing the proposed criterion (3.2) is equivalent to minimizing

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

which corresponds to the MMA criterion proposed by Hansen (2007).

4 Assumptions and Main Results

In this section, we present the asymptotic properties of the proposed averaging prediction. We first present the technical assumptions and provide the asymptotic expression for the MSPE of the averaging prediction. We then demonstrate the asymptotic optimality and asymptotic improvability of the proposed averaging prediction.

4.1 Assumptions

We state the assumptions required for main results.

Assumption 1. *d is a fixed nonnegative integer and bounded by some $\bar{d} < \infty$.*

Assumption 2. *Let $F_{t,m,\boldsymbol{v}_m}(\cdot)$ be the distribution function of the linear combination of innovations: $\boldsymbol{v}_m'\boldsymbol{\epsilon}_{t,m}$, where $\boldsymbol{\epsilon}_{t,m} = (\epsilon_t, ..., \epsilon_{t-m+1})'$ and $\boldsymbol{v}_m = (v_1, ..., v_m)' \in R^m$ with $\sum_{j=1}^m v_j^2 = 1$. For all $m \ge 1$, $m \le t < \infty$, there exist some real positive numbers α, δ, and C such that $F_{t,m,\boldsymbol{v}_m}(\cdot)$ satisfies the local Hölder condition of order α: $|F_{t,m,\boldsymbol{v}_m}(x) - F_{t,m,\boldsymbol{v}_m}(y)| \le C|x-y|^\alpha$, as $|x-y| \le \delta$.*

Assumption 3. $\sup_{0<t<\infty} E|\epsilon_t|^q < \infty$, $q = 1, 2, ...$

Assumption 4. $K_n^{\max\{4d-1,3\}} = o(n)$.

Assumption 1 implies that y_t is generated from an integrated autoregressive process with a finite integration order d for $0 \le d < \bar{d}$. Assumption 2 is the nonsingularity condition of Ing et al. (2010, 2012). It is used to establish the negative moment bounds of the minimum

eigenvalue of the Fisher information matrix. This condition holds for most continuous-type distributions, for example, the normal distribution; see Ing and Sin (2006) for more discussion.

Assumption 3 is the moment condition of ϵ_t and is identical to Condition (19) of Ing et al. (2010). Assumption 4 puts a bound on the number of models relative to the sample size. It also reflects the fact that the correlation among the time series y_t is higher when the integration order d is larger, and hence it may result in a smaller minimum eigenvalue of the information matrix defined in (2.4). Thus, the integration order d also limits the maximal order K_n used in MA.

Note that there is a trade-off between the moment condition on ϵ_t in Assumption 3 and divergence rate of K_n in Assumption 4. If we have a weaker moment condition in Assumption 3, then we will have a more restrictive condition on K_n in Assumption 4. For $d \geq 1$, our Assumption 4 is the same as the maximal order condition of Ing et al. (2010, 2012). For $d = 0$, we have $K_n^3 = o(n)$ in Assumption 4, while the maximal order condition in the MS case is $K_n^{2+\delta} = o(n)$ for some $\delta > 0$. Thus, our condition is slightly more restrictive than that of the MS case. This is a small price paid for the MA approach, because the MA approach selects the model weights from an uncountable set, while the MS approach compares the candidate models in a countable set.

4.2 MSPE

We first introduce some notation that we will use to characterize the asymptotic expression for the MSPE of the averaging prediction. 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 $\mathbf{a}(v) = (a_1(v), ..., a_v(v))' = \text{argmin}_{\mathbf{c}\in R^v} \text{E}(z_{t,\infty} + \mathbf{z}'_{t-1,\infty}(v)\mathbf{c})^2$. In the rest of the paper, we sometime use $\mathbf{a}(v)$ to denote an infinite dimensional vector with the ith element equal to $a_i(v), i = 1, 2, ...$, where $a_i(v) = 0$ if $i > v \geq 0$ or $v \leq 0$. Define $\|\mathbf{d}\|_z^2 = \sum_{1\leq i,j\leq\infty} d_i d_j \chi_{i-j}$, where $\chi_{i-j} = \text{E}(z_{i,\infty} z_{j,\infty})$, and $\mathbf{d} = (d_1, d_2, ...)'$ is an infinite dimensional vector which belongs to $l^2(\mathcal{Z}^+)$, that is, $\sum_{i\in\mathcal{Z}^+} d_i^2 < \infty$. Since $z_{t,\infty} + \sum_{i=1}^{\infty} a_i z_{t-i,\infty} = \epsilon_t$, we have, for all $v \geq 0$,

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

We next assume that the integration order d is known and present the asymptotic expression for the MSPE of the averaging prediction. The asymptotic optimality of the proposed averaging prediction with data-driven weights will be established for the unknown d in the next section. Define the MSPE of the averaging prediction as $\text{MSPE}(\mathbf{w}) =$

$\mathrm{E}(y_{n+1} - \hat{y}_{n+1}(\mathbf{w}))^2 - \sigma^2$. For a given value of the integration order d, we consider a sequence of finite-order approximation models starting from AR(max$(1, d)$) to AR(K_n). For $d \geq 1$, the most parsimonious candidate model is AR(d), which implies that we assign zero weight to the candidate model AR(k) for $1 \leq k < d$. Therefore, we consider the weight vector with $w_k \geq 0$, $\sum_{k=1}^{K_n} w_k = 1$, and $w_k = 0$ for $1 \leq k < d$. That is, the weight vector $\mathbf{w}$ belongs to the set $\mathcal{H}_n^d := \{\mathbf{w} \in [0,1]^{K_n} : \sum_{k=1}^{K_n} w_k = 1, w_k = 0 \text{ for } 1 \leq k < d\}$, and $\mathcal{H}_n^d$ is a subset of $\mathcal{H}_n$.

Theorem 1. *Suppose that Assumptions 1-4 hold, then we have*

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathcal{H}_n^d} \left| \frac{\mathrm{E}(y_{n+1} - \hat{y}_{n+1}(\boldsymbol{w}))^2 - \sigma^2}{L_n^d(\boldsymbol{w})} - 1 \right| = 0, \tag{4.1}$$

where

$$L_n^d(\boldsymbol{w}) = \sigma^2 \frac{d^2 + d}{N} + \sigma^2 \frac{\boldsymbol{w}'\Pi_{\min}(K_n)\boldsymbol{w} - d}{N} + \|\boldsymbol{a} - \boldsymbol{a}(\boldsymbol{w}, d)\|_z^2, \tag{4.2}$$

with

$$\sigma^2 \frac{\boldsymbol{w}'\Pi_{\min}(K_n)\boldsymbol{w} - d}{N} = \sigma^2 \frac{\sum_{1\leq i,j\leq K_n} w_i w_j \min(i,j) - d}{N}, \tag{4.3}$$

and

$$\begin{aligned} \|\boldsymbol{a} - \boldsymbol{a}(\boldsymbol{w}, d)\|_z^2 &= E\left[\sum_{v=1}^{K_n} w_v \left(\sum_{i=1}^{\infty} (a_i - a_i(v-d)) z_{t-i,\infty}\right)\right]^2 \\ &= \sum_{1\leq i,j\leq K_n} w_i w_j \|\boldsymbol{a} - \boldsymbol{a}(\max(i,j) - d)\|_z^2. \end{aligned} \tag{4.4}$$

Theorem 1 presents a uniformly asymptotic expression for the MSPE of the averaging prediction, and it shows that we can asymptotically decompose MSPE($\mathbf{w}$) into three terms. The first term is $N^{-1}\sigma^2(d^2 + d)$, which measures the estimation effect of the nonstationary component. The second term is $N^{-1}\sigma^2(\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w} - d)$, which measures the model complexity. As shown in (4.3), this term is penalized based on the smaller order between any two models. The third term is $\|\mathbf{a} - \mathbf{a}(\mathbf{w}, d)\|_z^2$, which measures the model misspecification. Unlike the model complexity term, the model misspecification term shown in (4.4) is approximated by the bigger order between any two models. Therefore, Theorem 1 demonstrates the bias-variance trade-off for the MA approach.

Note that Theorem 1 extends the uniformly asymptotic expression for the MSPE of the least squares predictor from MS to MA. Suppose that we set the weight vector $\mathbf{w}$ equal to the unit weight vector $\mathbf{w}_{1,k}$, where the kth element is one and others are zeros. Then, the averaging estimator simplifies to a selection estimator, and $\hat{y}_{n+1}(\mathbf{w}_{1,k})$ is equivalent to the

one-step-ahead prediction of y_{n+1} based on the AR(k) model, that is, $\hat{y}_{n+1}(\mathbf{w}_{1,k}) \equiv \hat{y}_{n+1}(k)$. For any $\mathbf{w}_{1,k}$, we can rewrite (4.3) and (4.4) as $N^{-1}\sigma^2(k-d)$ and $\|\mathbf{a}-\mathbf{a}(k-d)\|_z^2$, respectively. Hence, $L_n^d(\mathbf{w})$ in Theorem 1 can be simplified as follows:

$$L_n^d(\mathbf{w}_{1,k}) = \sigma^2\frac{d^2}{N} + \sigma^2\frac{k}{N} + \|\mathbf{a} - \mathbf{a}(k-d)\|_z^2, \tag{4.5}$$

which corresponds to the uniformly asymptotic expression for the MSPE of $\hat{y}_{n+1}(k)$ in Theorem 2 of Ing et al. (2010). Furthermore, when $d = 0$, the autoregressive process is stationary, and (4.5) corresponds to Theorem 3 of Ing and Wei (2003).

We next compare the uniformly asymptotic expression between (4.2) and (4.5) to characterize the potential MSPE improvement from MS to MA. Suppose that we have only two candidate models, AR(i) and AR(j), with $i < j$. Then the AR(j) model has a larger model complexity term, but a smaller model misspecification term, since $\|\mathbf{a}-\mathbf{a}(j-d)\|_z^2 \leq \|\mathbf{a}-\mathbf{a}(v-d)\|_z^2$ for any $v \leq j$. Note that the model complexity term and model misspecification term of the MS approach come from the same model, while these two terms of the MA approach are calculated based on $\min(i,j)$ in (4.3) and $\|\mathbf{a} - \mathbf{a}(\max(i,j) - d)\|_z^2$ in (4.4), respectively. Therefore, we can construct a convex combination between the model AR(i) and AR(j) such that the MSPE of the MA approach is smaller than that of the MS approach.

Based on the results of Theorem 1, we can further provide the conditions under which there exists at least one weight vector such that the MA approach achieves strictly lower MSPE than the MS approach. Observe that

$$\begin{aligned}\mathcal{H}_n^d &= \{\mathbf{w} \in [0,1]^{K_n} : \sum_{k=1}^{K_n} w_k = 1, w_k = 0 \text{ for } 1 \leq k < d\} \\ &= \{\mathbf{w} \in [0,1]^{K_n - \max(1,d)+1} : \sum_{k=\max(1,d)}^{K_n} w_k = 1\}.\end{aligned}$$

For the MS case, the candidate models AR($\max(1,d)$),..., AR(K_n) correspond one-to-one to vertices of $\mathcal{H}_n^d$. Define $\mathcal{V}(\mathcal{H}_n^d)$ as the set of all the vertices in $\mathcal{H}_n^d$ and let $\mathcal{H}_n^d \backslash \mathcal{V}(\mathcal{H}_n^d)$ be the weight set $\mathcal{H}_n^d$ excluding the vertices $\mathcal{V}(\mathcal{H}_n^d)$.

Corollary 1. *Suppose that Assumptions 1-4 hold. If there is a $k \in \{\max(1,d), ..., K_n\}$ such that*

$$\|\boldsymbol{a} - \boldsymbol{a}(k-d)\|_z^2 \neq \|\boldsymbol{a} - \boldsymbol{a}(l-d)\|_z^2, \;\; \forall\, l \in \{\max(1,d), ..., K_n\}, \;\; l \neq k, \tag{4.6}$$

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

$$\inf_{\boldsymbol{w} \in \mathcal{H}_n^d \backslash \mathcal{V}(\mathcal{H}_n^d)} L_n^d(\boldsymbol{w}) \leq L_n^d(\boldsymbol{w}^\diamond) < \min_{\boldsymbol{w} \in \mathcal{V}(\mathcal{H}_n^d)} L_n^d(\boldsymbol{w}).$$

Corollary 1 shows that when the model misspecification bias of one particular model is different from that of any other candidate models, there exists at least one weight vector such that the MSPE of the associated MA estimator is strictly less than that of any MS estimator. In this case, the optimal weight vector that minimizes $L_n^d(\mathbf{w})$ will assign non-zero weights on at least two models, which implies that MA can further reduce the MSPE of MS.

We now present the uniformly asymptotic expression for the MSPE of the averaging prediction with the optimal weights. Let $\mathbf{w}_n^* = \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w})$ denote the weights that minimize $L_n^d(\mathbf{w})$ over the set $\mathcal{H}_n^d$. Then, $L_n^d(\mathbf{w}_n^*)$ is the minimum MSPE of the MA estimator.

Corollary 2. *Let $A_j = \|\boldsymbol{a} - \boldsymbol{a}(j-d)\|_z^2$. Suppose that Assumptions 1-4 hold, then we have*

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

Like Theorem 1, we can also decompose $L_n^d(\mathbf{w}_n^*)$ into three components. Ideally, one would aim to estimate the weights $\mathbf{w}_n^*$ by minimizing $L_n^d(\mathbf{w})$ directly. Unfortunately, this is infeasible because the minimization of $L_n^d(\mathbf{w})$ depends on the unknown model misspecification bias. Instead of minimizing the unknown $L_n^d(\mathbf{w})$, we select the data-driven weights by minimizing the proposed Mallows-type criterion and demonstrate that the empirical weights asymptotically minimize the MSPE.

4.3 Asymptotic optimality

In practice, the integration order d is unknown. Therefore, we construct the averaging prediction by combining all finite-order AR models starting from AR(1) to AR(K_n) for any $0 \le d < \bar{d}$. In this section, we first show that the proposed averaging prediction asymptotically assigns zero weight to the model with the autoregressive order less than d. We then show that the proposed averaging prediction is asymptotically optimal when the integration order d is unknown.

For any data-driven MA criterion, $\mathrm{MA}_n(\mathbf{w})$, let $\hat{\mathbf{w}}_{\mathrm{MA}} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} \mathrm{MA}_n(\mathbf{w})$ and $\hat{\mathbf{w}}_{\mathrm{MA}}^d := \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} \mathrm{MA}_n(\mathbf{w})$ be the weights obtained by minimizing the criterion $\mathrm{MA}_n(\mathbf{w})$ over the weight set $\mathcal{H}_n$ and $\mathcal{H}_n^d$, respectively. A data-driven MA criterion is said to be asymptotically optimal for a dth-order integrated AR(∞) process with an unknown d if the data-driven weights satisfy

$$\|\hat{\mathbf{w}}_{\mathrm{MA}} - \hat{\mathbf{w}}_{\mathrm{MA}}^d\|_2 \xrightarrow{a.s.} 0, \tag{4.7}$$

and

$$\lim_{n\to\infty} \frac{L_n^d(\hat{\mathbf{w}}_{\mathrm{MA}}^d)}{L_n^d(\mathbf{w}_n^*)} \xrightarrow{p} 1. \tag{4.8}$$

Condition (4.7) states that when d is unknown, the weights obtained by minimizing the criterion $\mathrm{MA}_n(\mathbf{w})$ over the unrestrictive set $\mathcal{H}_n$ will converge almost surely to the weights obtained by minimizing the same criterion $\mathrm{MA}_n(\mathbf{w})$ over the restrictive set $\mathcal{H}_n^d$. This condition implies that the weight on the candidate model with the autoregressive order less than d will converge almost surely to zero. Recall that $L_n^d(\mathbf{w})$ is the uniformly asymptotic expression for the MSPE of the averaging prediction derived in Theorem 1. Thus, Condition (4.8) states that the data-driven weights achieve the lowest possible MSPE on the set $\mathcal{H}_n^d$ asymptotically.

Let $\xi_n^d := L_n^d(\mathbf{w}_n^*) = \inf_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w})$ denote the minimum MSPE in the class of averaging estimators with weights belonging to the set $\mathcal{H}_n^d$ derived in Corollary 2. We next state the additional assumption for the asymptotic optimality.

Assumption 5. $K_n^{1/2}(N\xi_n^d)^{-1} \longrightarrow 0$.

Assumption 5 puts a bound on the maximum order K_n relative to the sample size N, and it specifies that $K_n^{1/2}$ grows at a rate slower than $N\xi_n^d$. As pointed out in Cheng et al. (2015) and Liao et al. (2021), many MA approaches require a stronger assumption on K_n and it may result in inferior prediction due to the preclusion of the optimal model. Unlike the conditions used in the existing studies, for example, Condition (11) of Ando and Li (2014), Assumption 2(a) of Liao and Tsay (2020), Assumption 2 of Zhang (2021), and Condition (4.3) of Liao et al. (2021), Assumption 5 is weaker and quite mild, and hence does not preclude the optimal model.

Let $\hat{\mathbf{w}}_{\mathrm{MMA}} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} C_n(\mathbf{w})$ and $\hat{\mathbf{w}}_{\mathrm{MMA}}^d := \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} C_n(\mathbf{w})$ be the MMA weights obtained by minimizing the criterion $C_n(\mathbf{w})$ over the weight set $\mathcal{H}_n$ and $\mathcal{H}_n^d$, respectively. The following theorem shows the asymptotic optimality of the proposed averaging criterion when d is unknown.

Theorem 2. *Suppose that Assumptions 1-5 hold, then we have*

$$\|\hat{\boldsymbol{w}}_{\mathrm{MMA}} - \hat{\boldsymbol{w}}_{\mathrm{MMA}}^d\|_2 \xrightarrow{a.s.} 0 \ \ \textit{and} \ \ \lim_{n\to\infty} \frac{L_n^d(\hat{\boldsymbol{w}}_{\mathrm{MMA}}^d)}{L_n^d(\boldsymbol{w}_n^*)} \xrightarrow{p} 1.$$

Theorem 2 shows that the proposed MA prediction achieves the lowest possible MSPE. This optimal result extends the asymptotic optimality of Ing et al. (2012) from MS to MA. In addition, for the MA methods, the result extends the asymptotic optimality in Theorem 1 of Liao et al. (2021) from the stationary autoregressive process to a dth-order integrated AR(∞) process with an unknown d.

4.4 Asymptotic improvability

In this section, we provide an asymptotic comparison for the MSPE between MA and MS. Recall that $\hat{y}_{n+1}(\mathbf{w}_{1,k}) = \hat{y}_{n+1}(k)$, where $\mathbf{w}_{1,k}$ is the unit weight vector that assigns the whole

weight on the kth element. Let $\mathbf{w}_{1,k}^* := \arg\min_{\mathbf{w}\in\mathcal{V}(\mathcal{H}_n^d)} L_n^d(\mathbf{w})$ denote the optimal unit weight vector that minimizes $L_n^d(\mathbf{w})$ over the set of all the vertices in $\mathcal{H}_n^d$. Thus, $L_n^d(\mathbf{w}_{1,k}^*)$ is the minimum MSPE of the MS estimator. Like $\mathbf{w}_n^*$, the optimal unit weight vector $\mathbf{w}_{1,k}^*$ also depends on the sample size n. Here we suppress the sample size n from $\mathbf{w}_{1,k}^*$ for notational simplicity.

We follow Peng and Yang (2022) and examine the potential MSPE reduction of MA compared to MS as follows:

$$\Delta_n = L_n^d(\mathbf{w}_{1,k}^*) - L_n^d(\mathbf{w}_n^*). \tag{4.8}$$

We then investigate the magnitude of Δ_n relative to $L_n^d(\mathbf{w}_{1,k}^*)$ in the following two cases:

(i) Algebraic-decay case: $\|\mathbf{a} - \mathbf{a}(v)\|_z^2 = Cv^{-\alpha}$,

(ii) Exponential-decay case: $\|\mathbf{a} - \mathbf{a}(v)\|_z^2 = C\exp(-\alpha(v))$,

where α is a positive constant. Both the algebraic-decay and exponential-decay are frequently used in time series analysis. The above two scenarios are simplified but have the same optimal orders of the MS estimator as the examples 1 and 2 in Ing and Wei (2005). Following Peng and Yang (2022), we use the symbols $\succeq$ and $\asymp$, where $a_n \succeq b_n$ means $b_n = O(a_n)$, and $a_n \asymp b_n$ means both $a_n \succeq b_n$ and $b_n \succeq a_n$.

Theorem 3. *Suppose that Assumptions 1-4 hold, then we have*

(i) $\Delta_n \asymp L_n^d(\boldsymbol{w}_{1,k}^*)$ *for the algebraic-decay case,*

(ii) $\Delta_n = o(L_n^d(\boldsymbol{w}_{1,k}^*))$ *for the exponential-decay case.*

From Corollary 1, we show that if there exists a candidate model whose misspecification bias is different from that of other models, then Δ_n is greater than zero. Theorem 3 further provides a measurement on the potential MSPE improvement from MS to MA. Under a dth-order integrated AR(∞) model, if the model misspecification bias is algebraic decay as the model dimension increases, the magnitude of potential MSPE reduction has the same order as that of the minimum MSPE of MS. However, for the exponential-decay case, the magnitude is asymptotically negligible. These results are consistent with the existing findings such as Peng and Yang (2022) and Xu and Zhang (2022), in which they consider a non-stochastic regression design. In contrast to their framework, our results are established for a general autoregressive model with broader applicability.

We next compare the MSPE of data-driven MA and MS approaches. Similar to the definition of $\hat{\mathbf{w}}_{\text{MA}}^d$ in the previous section, we use $\hat{\mathbf{w}}_{\text{MS}}^d := \arg\min_{\mathbf{w}\in\mathcal{V}(\mathcal{H}_n^d)} \text{MA}_n(\mathbf{w})$ to denote the unit weight vector that minimizes the criterion $\text{MA}_n(\mathbf{w})$ over the set of all the vertices in $\mathcal{H}_n^d$. Therefore, for any data-driven MS approach, $L_n^d(\hat{\mathbf{w}}_{\text{MS}}^d)$ is the MSPE of the associated MS estimator.

Corollary 3. *Let* $\hat{\Delta}_n := L_n^d(\hat{\boldsymbol{w}}_{\text{MS}}^d) - L_n^d(\hat{\boldsymbol{w}}_{\text{MA}}^d)$. *Suppose that Assumptions 1-5 hold. If*

$$\frac{L_n^d(\hat{\boldsymbol{w}}_{\text{MS}}^d)}{L_n^d(\boldsymbol{w}_{1,k}^*)} \xrightarrow{p} 1 \quad \textit{and} \quad \frac{L_n^d(\hat{\boldsymbol{w}}_{\text{MA}}^d)}{L_n^d(\boldsymbol{w}_n^*)} \xrightarrow{p} 1, \tag{4.9}$$

then we have

(i) $\hat{\Delta}_n \asymp L_n^d(\hat{\boldsymbol{w}}_{\text{MS}}^d)$ *for the algebraic-decay case,*

(ii) $\hat{\Delta}_n = o(L_n^d(\hat{\boldsymbol{w}}_{\text{MS}}^d))$ *for the exponential-decay case.*

Furthermore, we have $L_n^d(\hat{\boldsymbol{w}}_{\text{MA}}^d) \asymp L_n^d(\hat{\boldsymbol{w}}_{\text{MS}}^d)$ *in both cases.*

Corollary 3 shows that if both data-driven MA and MS approaches are asymptotic optimal, then the asymptotic improvability in Theorem 3 will hold for the chosen weight $\hat{\mathbf{w}}_{\text{MA}}^d$ and selected model $\hat{\mathbf{w}}_{\text{MS}}^d$. Note that under Assumptions 1-5, Theorem 3.1 of Ing et al. (2012) demonstrates the asymptotic optimality of AIC and its equivalent MS criteria. In the next section, we will discuss the asymptotic optimality for the MA counterparts of these MS criteria.

5 Other Model Averaging Criteria

In this section, we discuss the relationship between the proposed MMA criterion and other data-driven MA criteria, and investigate their asymptotic properties. Inspired by Shibata (1980), we propose a Shibata model averaging (SMA) criterion as follows:

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

where $\Pi_{\min}(K_n)$, $\Pi_{\max}(K_n)$, and $\hat{\sigma}_w^2$ are defined in (3.1) and (3.2). Note that SMA extends Shibata (1980)'s criterion from MS to MA, and it is closely related to AIC-type and Mallows-type MA criteria. Following the same idea, we define an Akaike model averaging (AMA) criterion as follows:

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

Like the discussion in Section 4.2, if we consider the unit weight vector $\mathbf{w}_{1,k}$, then $S_n(\mathbf{w}_{1,k})$ and $A_n(\mathbf{w}_{1,k})$ will correspond to the Shibata MS criterion, $S_n(k) := (N + 2k)\hat{\sigma}^2(k)$, and Akaike information criterion, $A_n(k) := \log(\hat{\sigma}^2(k)) + N^{-1}(2k)$, respectively, where $\hat{\sigma}^2(k) = N^{-1}\sum_{t=K_n}^{n-1}(y_{t+1} - \hat{y}_{t+1}(k))^2$. However, unlike the MMA criterion, both SMA and AMA criteria are not a quadratic function of the weight vector and cannot be solved by quadratic programming.

Let $\hat{\mathbf{w}}_{\text{SMA}} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} S_n(\mathbf{w})$ and $\hat{\mathbf{w}}_{\text{SMA}}^d := \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} S_n(\mathbf{w})$ be the SMA weights obtained by minimizing the criterion $S_n(\mathbf{w})$ over the weight set $\mathcal{H}_n$ and $\mathcal{H}_n^d$, respectively.

Similarly, let $\hat{\mathbf{w}}_{\text{AMA}} := \arg\min_{\mathbf{w}\in\mathcal{H}_n} A_n(\mathbf{w})$ and $\hat{\mathbf{w}}^d_{\text{AMA}} := \arg\min_{\mathbf{w}\in\mathcal{H}^d_n} A_n(\mathbf{w})$. The following theorem shows the asymptotic optimality of SMA and AMA criteria when d is unknown.

Theorem 4. *Suppose that Assumptions 1-5 hold. For the Shibata model averaging criterion, we have*

$$\|\hat{\boldsymbol{w}}_{\text{SMA}} - \hat{\boldsymbol{w}}^d_{\text{SMA}}\|_2 \xrightarrow{a.s.} 0 \ \ \textit{and} \ \ \lim_{n\to\infty} \frac{L^d_n(\hat{\boldsymbol{w}}^d_{\text{SMA}})}{L^d_n(\boldsymbol{w}^*_n)} \xrightarrow{p} 1.$$

For the Akaike model averaging criterion, we have

$$\|\hat{\boldsymbol{w}}_{\text{AMA}} - \hat{\boldsymbol{w}}^d_{\text{AMA}}\|_2 \xrightarrow{a.s.} 0 \ \ \textit{and} \ \ \lim_{n\to\infty} \frac{L^d_n(\hat{\boldsymbol{w}}^d_{\text{AMA}})}{L^d_n(\boldsymbol{w}^*_n)} \xrightarrow{p} 1.$$

Theorem 4 shows that both SMA and AMA are asymptotically optimal in the sense of achieving the lowest possible MSPE. Therefore, according to Theorem 2 and Theorem 4, MMA, SMA, and AMA are asymptotically equivalent in terms of achieving the minimum MSPE. Furthermore, the differences among these criteria (up to a monotone transformation) are negligible relative to the minimum MSPE of MA as shown in Lemma 8 and the proof of Theorem 4 in the Appendix. For MS methods, Shibata (1980), Ing and Wei (2005), and Ing et al. (2012) established similar results for these MS criteria. Our results extend the asymptotic equivalence between these criteria from MS to MA for a dth-order integrated AR(∞) model.

6 Simulations

In this section, we investigate the finite-sample performance of proposed averaging criteria in two simulation designs. The first design corresponds to the algebraic-decay case, and we consider the following process:

$$\big(1 + \sum_{j=1}^{100} a_j L^j\big)(1-L)^d y_t = \epsilon_t + 0.5\epsilon_{t-1},$$

where $\epsilon_t \sim i.i.d.N(0,1)$. We set $a_j = c(-1)^{j-1}j^{-\alpha}$, where $\alpha = 0.5$, 1, or 1.5, and the parameter c is varied on a grid from 0.1 to 0.9.

The second design corresponds to the exponential-decay case, and we consider the following ARIMA(1,d,1) process:

$$\big(1 + \phi L\big)(1-L)^d y_t = \epsilon_t + \theta\epsilon_{t-1},$$

where $\epsilon_t \sim i.i.d.N(0,1)$. The coefficient ϕ is varied on a grid from -0.8 to 0.8, and the coefficient θ is set to be 0.25, 0.5, or 0.75. In both simulation designs, the integration order d is set to be 0, 1, or 2. The sample size is varied between $n = 100$, 200, 500, and 1000, and the number of models is $K_n = [3n^{1/3}]$, where $[a]$ is the nearest integer of a.

We consider the following MS and MA estimators:

1. Akaike information criterion MS estimator (labeled AIC).
2. Bayesian information criterion MS estimator (labeled BIC).
3. Mallows' C_p MS estimator (labeled Cp).
4. Shibata information criterion MS estimator (labeled SIC).
5. Smoothed Bayesian information criterion MA estimator (labeled SBIC)
6. Akaike model averaging estimator (labeled AMA)
7. Shibata model averaging estimator (labeled SMA)
8. Mallows model averaging estimator (labeled MMA)

The AIC and BIC criteria are $A_n(k) = \log(\hat{\sigma}^2(k)) + N^{-1}(2k)$ and $B_n(k) = \log(\hat{\sigma}^2(k)) + N^{-1}(\log(N)k)$, respectively, where $\hat{\sigma}^2(k) = N^{-1}\sum_{t=K_n}^{n-1}(y_{t+1} - \hat{y}_{t+1}(k))^2$. The Cp and SIC criteria are $C_n(k) = N\hat{\sigma}^2(k) + 2k\check{\sigma}^2$ and $S_n(k) = (N + 2k)\hat{\sigma}^2(k)$, respectively. For the AIC, BIC, Cp, and SIC, we select the model with the smallest criterion value. The SBIC estimator is a simplified form of Bayesian model averaging with diffuse priors and is defined as $\hat{w}_k = \exp(-0.5NB_n(k))/\sum_{j=1}^{K_n}\exp(-0.5NB_n(j))$. The MMA estimator is defined in (3.2) and the AMA and SMA estimators are described in Section 5.

We evaluate the finite sample behavior of each method based on the following empirical MSPE: $\frac{1}{S}\sum_{s=1}^{S}\left(\frac{N}{\sigma^2}\left((y_{n+1}^{\{s\}} - \hat{y}_{n+1}^{\{s\}}(\hat{\mathbf{w}}^{\{s\}}))^2 - \sigma^2\right)\right)$, where $\hat{y}_{n+1}^{\{s\}}(\hat{\mathbf{w}}^{\{s\}})$ is the prediction based on each method in the sth replication. As pointed out in Hansen (2008), we subtract the error variance σ^2 because it is the common leading term of the MSPE across all candidate models. Here, the scaling N/σ^2 is used to ensure that results are scale-free. For σ^2, we use the same estimator $\hat{\sigma}^2(K_n) = N^{-1}\sum_{t=K_n}^{n-1}(y_{t+1}^{\{s\}} - \hat{y}_{t+1}^{\{s\}}(K_n))^2$ for all methods in each replication. The empirical MSPE is calculated by averaging across 50000 simulation replications. For ease of comparison, we divide the MSPE of each method by that of MMA and report the relative MSPE. Lower relative MSPE means better performance on predictions. When the relative MSPE exceeds one, it indicates that the specified method performs worse than MMA.

In Figure 1, we present the relative MSPEs of the various estimates for $d = 1$ in the algebraic-decay case. In each figure, the relative MSPEs are displayed for $\alpha = \{0.5, 1.0, 1.5\}$ and $n = \{100, 200, 500, 1000\}$ in 12 panels, and in each panel, the relative MSPEs are displayed for c between 0.1 and 0.9. The simulation results show that the MMA, AMA, and SMA have similar MSPEs in most situations and perform quite well. These three MA estimators have lower MSPEs than those of the AIC, Cp, and SIC, which is consistent with Corollary 3. The BIC is dominated by the SBIC, AMA, MMA, and SMA. The SBIC performs slightly better than the AMA, MMA, and SMA when $\alpha = 1.5$ and $n = 100$, but performs worse than the AMA, MMA, and SMA when α is small and n is large.

In Figure 2, we present the relative MSPEs of the various estimates for $d = 1$ in the exponential-decay case. In each figure, the relative MSPEs are displayed for $\theta =$

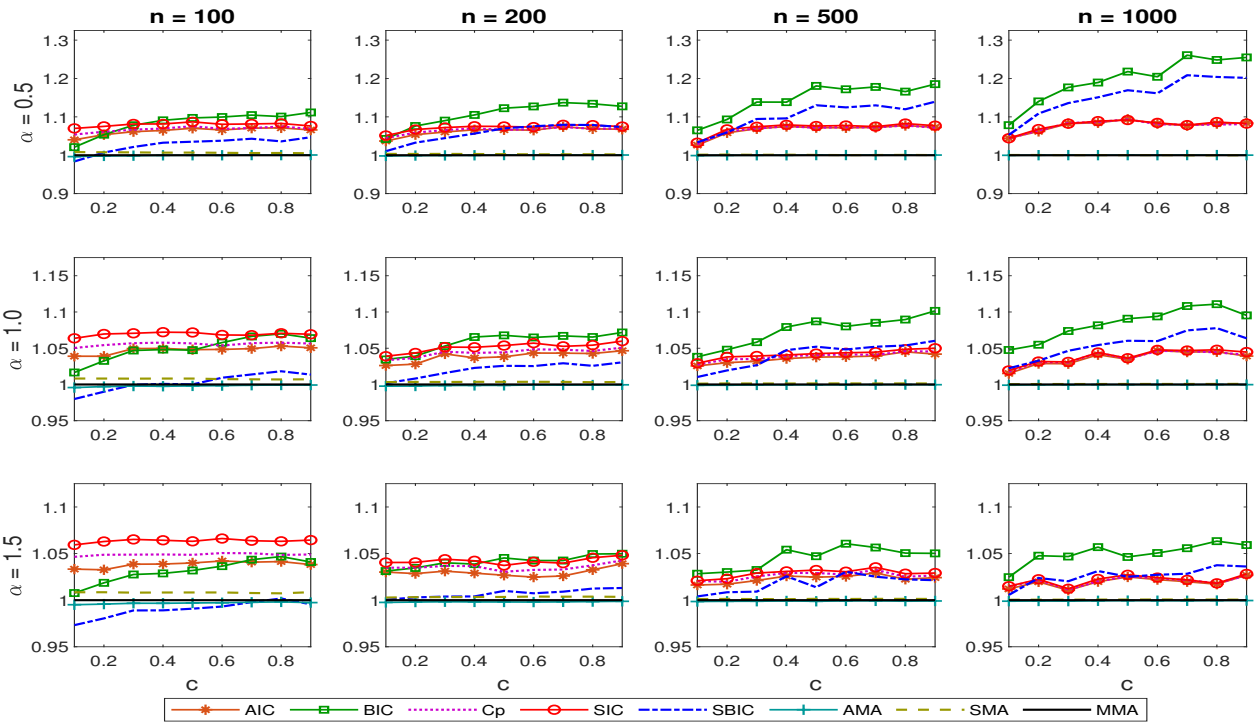

Figure 1: Relative MSPEs for $d = 1$ in the algebraic-decay case

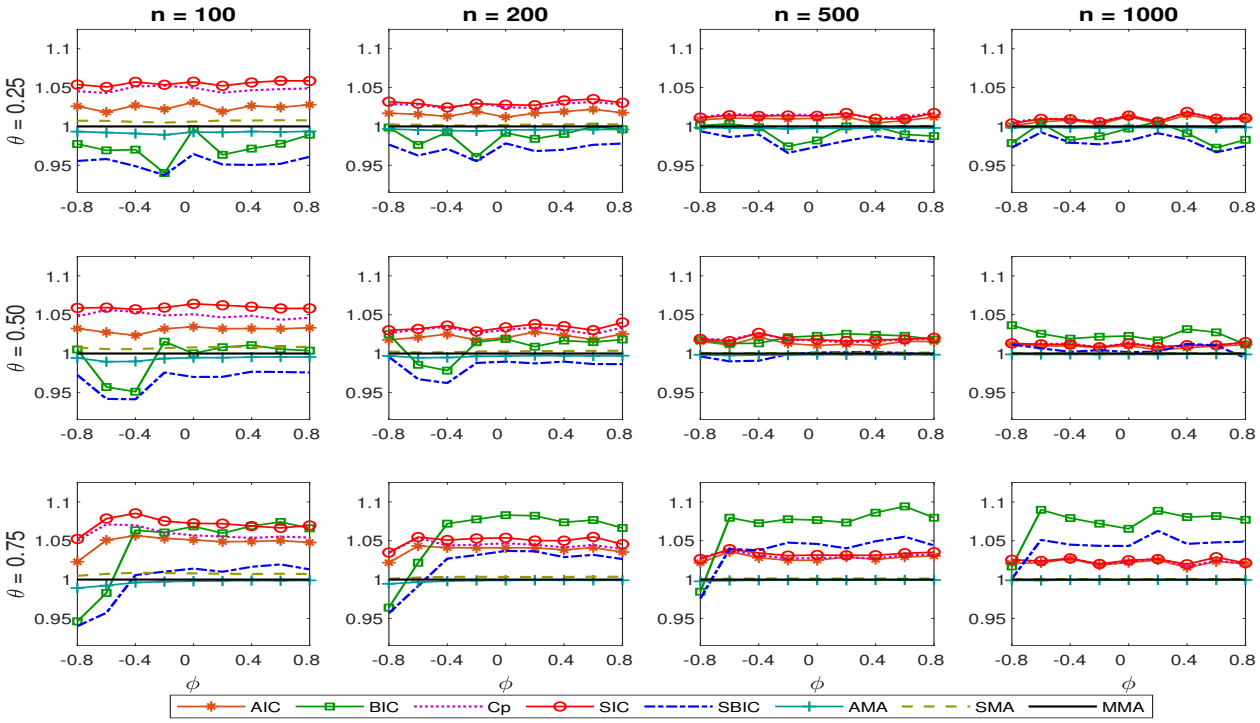

Figure 2: Relative MSPEs for $d = 1$ in the exponential-decay case

$\{0.25, 0.5, 0.75\}$ and $n = \{100, 200, 500, 1000\}$ in 12 panels, and in each panel, the relative MSPEs are displayed for ϕ between -0.8 and 0.8. The simulation results show that the MMA, AMA, and SMA still achieve lower MSPEs than those of AIC, Cp, and SIC in the exponential-decay case, but the efficiency gain of MA over MS diminishes as the sample size increases. The relative performance of the BIC, SBIC, and other estimators depends strongly on the sample size n and coefficients ϕ and θ. Both BIC and SBIC have larger MSPEs than other estimators when ϕ, θ, and n are large.

To illustrate the effect of the sample size on the MSPE in both algebraic-decay and exponential-decay cases, we present the relative MSPEs in Figures 3 and 4, in which the sample size increases from 100 to 2500 on a logarithmic scale. As the sample size increases, we can observe that the MSPEs of AIC, Cp, and SIC are getting close to those of AMA, MMA, and SMA in the exponential-decay case, but not in the algebraic-decay case, which is consistent with Corollary 3. Unlike these estimators, the relative MSPEs of BIC and SBIC increase as the sample size increases, when c is large in the algebraic-decay case and $\theta = 0.75$ in the exponential-decay case. Therefore, it shows that the BIC and SBIC are not asymptotically optimal in these cases. In the supplementary material, we also present the relative MSPEs for $d = 0$ and $d = 2$ in the algebraic-decay and exponential-decay cases, and the ranking of these estimators is quite similar to that for $d = 1$.

7 Empirical Example

In this section, we apply the proposed MA methods to the climate change prediction. We employ Rohde and Hausfather (2020)'s global land-ocean temperature dataset to study the Earth's surface temperature change between January 1850 to December 2021. Rohde and Hausfather (2020) constructed the Earth land and ocean temperature changes relative to a 1951-1980 baseline period; see their paper for a detailed description of the data construction. The monthly data consist of 2064 observations and are available at: https://doi.org/10.5281/zenodo.3634713. The time series plot of the land-ocean temperature changes between 1850-2021 is presented in Figure 5.

We calculate the one-step-ahead prediction of the land-ocean temperature changes using a rolling estimation scheme. We set the rolling window size to $n = 100$, 200, 500, or 1000, and the number of models as $K_n = [3n^{1/3}]$, where $[a]$ is the nearest integer of a. For each rolling window size n, we use observations $\{y_t\}_{t=b}^{n+b-1}$ in the training sample to estimate a sequence of AR(k) models and then apply the same MS and MA methods as those in the simulation study to construct the one-step-ahead prediction of y_{n+b} for $b = 1, ..., B$, where $B = n_0 - n$ and $n_0 = 2064$. We next evaluate these methods based on the following empirical MSPE: $\frac{1}{B}\sum_{b=1}^{B}\left(\frac{N}{\sigma^2}\left((y_{n+b} - \hat{y}_{n+b}(\hat{\mathbf{w}}_b))^2 - \sigma^2\right)\right)$, where $\hat{y}_{n+b}(\hat{\mathbf{w}}_b)$ is the prediction

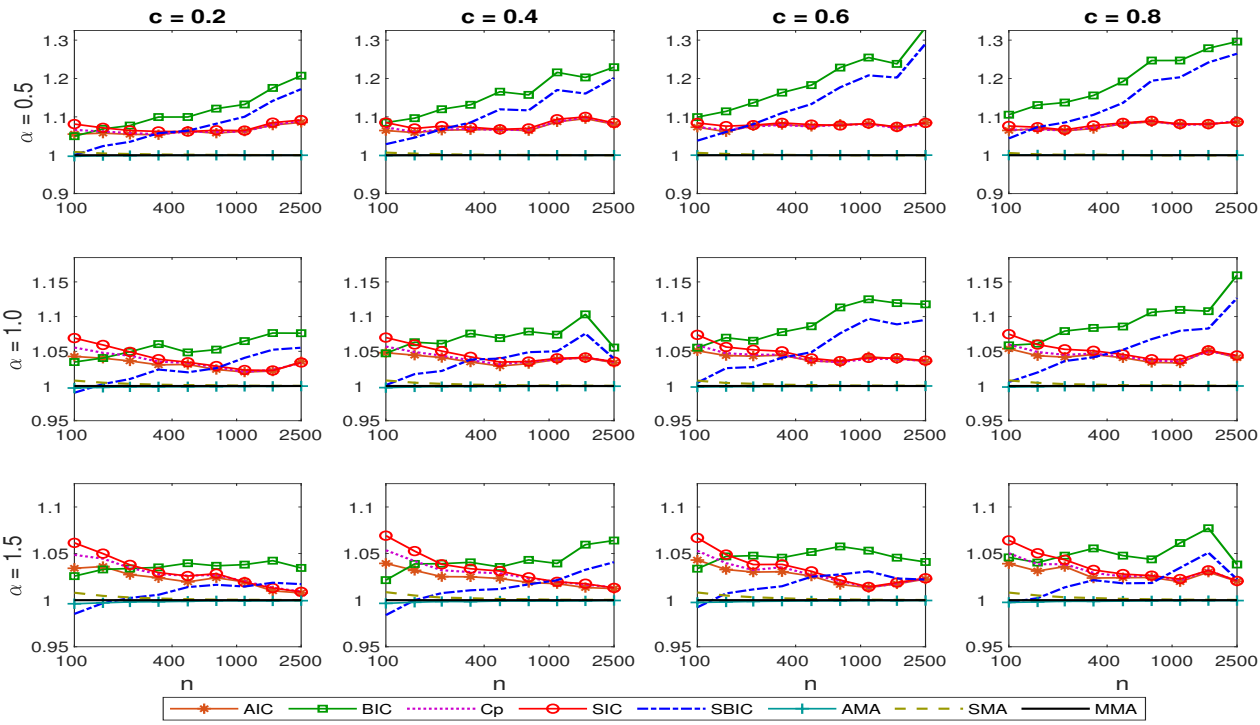

Figure 3: Relative MSPEs for the algebraic-decay case, $d = 1$, various sample sizes

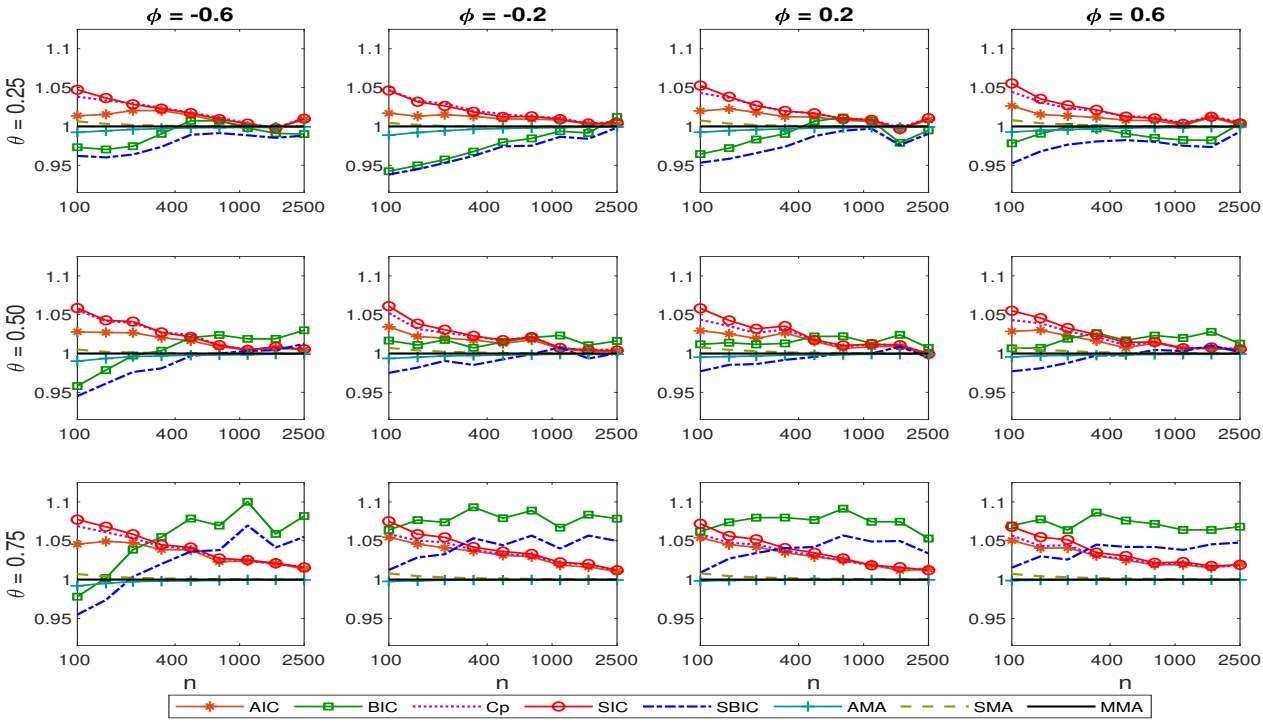

Figure 4: Relative MSPEs for the exponential-decay case, $d = 1$, various sample sizes

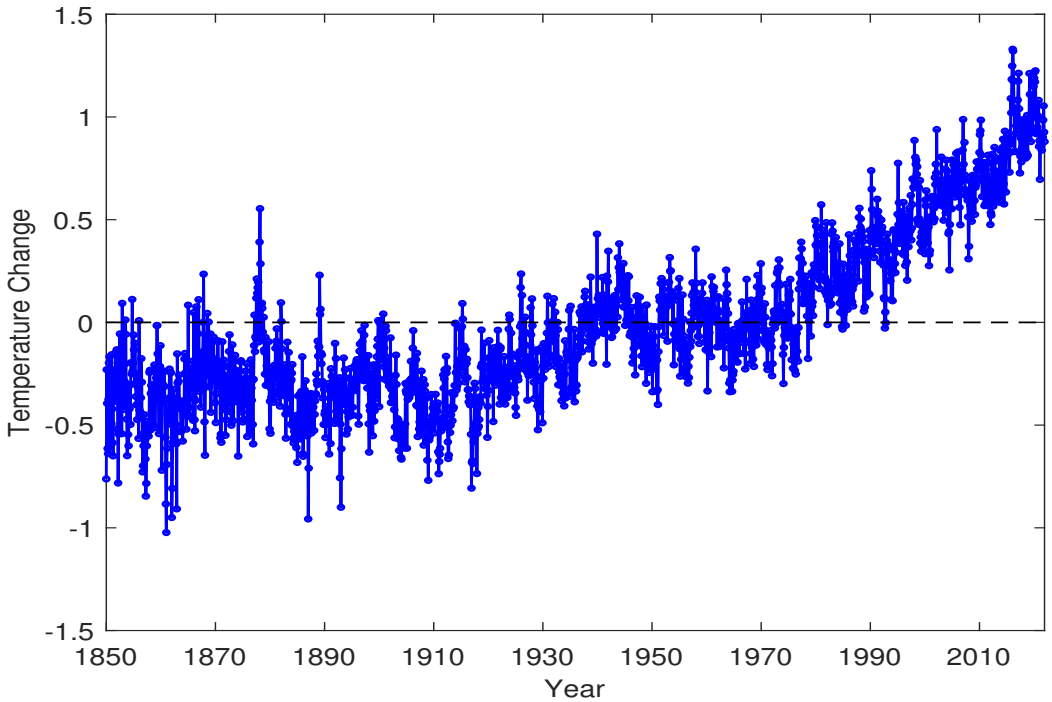

Figure 5: Temperature Change in Celsius

based on each method in the bth training sample. For ease of comparison, we divide the empirical MSPE of each method by that of MMA and report the relative MSPE. Thus, an entry greater than one indicates that the specified method performs worse than MMA.

Table 1 presents the relative MSPEs of MS and MA methods. The results show that AMA, MMA, and SMA achieve lower MSPEs than other methods in most scenarios. As the rolling window size increases, we observe that the MSPEs of AIC, Cp, and SIC approach those of AMA, MMA, and SMA, while the relative MSPEs of BIC and SBIC are increasing. The pattern of relative performance among these estimators in this empirical example is quite similar to that of the exponential-decay case in the simulation study.

Table 1: Relative MSPEs

n	AIC	BIC	Cp	SIC	SBIC	AMA	MMA	SMA
100	1.022	1.022	1.030	1.037	1.002	1.000	1.000	1.005
200	1.020	1.009	1.022	1.022	1.004	1.000	1.000	1.001
500	1.007	1.020	1.008	1.009	1.017	1.000	1.000	1.000
1000	1.003	1.044	1.003	1.003	1.040	1.000	1.000	1.000

8 Conclusion

In this paper, we study the MA prediction for integrated autoregressive processes of infinite order. We first derive a uniformly asymptotic expression for the MSPE of the averaging prediction with fixed weights and demonstrate the bias-variance trade-off for the MA approach. We then propose a Mallows-type criterion to select the data-driven weights and investigate two related MA methods, Shibata and Akaike MA estimators. We show that the proposed method and these two related methods are asymptotically optimal in the sense of achieving the lowest possible MSPE. We further demonstrate that the MA methods can provide significant MSPE reduction over the MS methods when the model misspecification bias is algebraic decay, but the magnitude of improvement is asymptotically negligible when the model misspecification bias is exponential decay. The theoretical properties of asymptotic optimality and asymptotic improvability are supported by the simulation study and real data analysis.

Appendix

A MSPE Decomposition and Supplementary Lemmas

We first provide the decomposition for the MSPE of the averaging prediction. As shown in Eq. (4) of Ing et al. (2010), the difference between y_{n+1} and $\hat{y}_{n+1}(k)$ can be decomposed as:

$$y_{n+1} - \hat{y}_{n+1}(k) = \Big\{ - N^{-1}\mathbf{s}_{n,n}^{'}(k)\Big[N^{-1}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(k)\mathbf{s}_{j,n}^{'}(k)\Big]^{-1}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(k)\epsilon_{j+1,k-d}\Big\} + \epsilon_{n+1,k-d}, \tag{A.1}$$

where $\mathbf{s}_{j,n}(k) = G_n(k)Q(k)\mathbf{y}_j(k)$, $G_n(k)$ is a $k \times k$ diagonal matrix defined as

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

$Q(k)$ is a $k \times k$ matrix such that

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

with $y_j(v) = (1-L)^{d-v} y_j$, and

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

Based on the decomposition of (A.1), the MSPE of the averaging prediction can be rewritten as

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

where

$$f_n(k) = \frac{\mathbf{s}'_{n,n}(k)}{\sqrt{N}} \Big[N^{-1} \sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(k) \mathbf{s}'_{j,n}(k) \Big]^{-1} \Big(\frac{1}{\sqrt{N}} \sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(k) \epsilon_{j+1,k-d} \Big),$$

and

$$S_n(k-d) = \epsilon_{n+1,k-d} - \epsilon_{n+1} = \sum_{i=1}^{n} (a_i - a_i(k-d)) z_{n+1-i}.$$

Note that $f_n(k)$ can be further decomposed into a non-stationary term $B_{1n}(k,d)$ and stationary term $B_{2n}(k-d)$ as follows:

$$B_{1n}(k,d) = \Big\{ \frac{U'_{n,n}(d)}{\sqrt{N}} \hat{\Omega}_n^{-1}(k) \frac{1}{\sqrt{N}} \sum_{j=K_n}^{n-\sqrt{n}-1} U_{j,n}(d) \epsilon_{j+1,k-d} \Big\} 1(d \geq 1),$$

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

where $\hat{\Omega}_n(k) = \Big[N^{-1} \sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(k) \mathbf{s}'_{j,n}(k) \Big]$ for $k \geq 1$, $\Gamma(v) = \mathrm{E}(\mathbf{z}_{t,\infty}(v) \mathbf{z}'_{t,\infty}(v))$ for $v \geq 1$, and $U_{j,n}(v) = \big((y_j(d)/N^{d-(1/2)}, ..., y_j(d-v+1)/N^{d-v+(1/2)}] \big)'$.

To take care of the dependence between future observations and the estimation sample, we use the following terms to approximate $B_{1n}(k,d)$

$$f_{1,n}(d) = \Big\{ \frac{U'_{n,n}(d)}{\sqrt{N}} \Big[N^{-1} \sum_{j=K_n}^{n-\sqrt{n}-1} U_{j,n}(d) U'_{j,n}(d) \Big]^{-1} \frac{1}{\sqrt{N}} \sum_{j=K_n}^{n-\sqrt{n}-1} U_{j,n}(d) \epsilon_{j+1} \Big\} 1(d \geq 1),$$

$$f^*_{1,n}(d) = \Big\{ \frac{U^{*\prime}_{n,n}(d)}{\sqrt{N}} \Big[N^{-1} \sum_{j=K_n}^{n-\sqrt{n}-1} U_{j,n}(d) U'_{j,n}(d) \Big]^{-1} \frac{1}{\sqrt{N}} \sum_{j=K_n}^{n-\sqrt{n}-1} U_{j,n}(d) \epsilon_{j+1} \Big\} 1(d \geq 1),$$

where $U^*_{n,n}(d) = \Big((N^{-d+1/2} \sum_{j=\sqrt{n}}^{n-1} \kappa_j(d) \epsilon_{n-j}, ..., N^{-1/2} \sum_{j=\sqrt{n}}^{n-1} \kappa_j(1) \epsilon_{n-j} \Big)'$, $\kappa_j(1) = \sum_{s=0}^{j} b_s$, and $\kappa_j(v) = \sum_{s=0}^{j} \kappa_s(v-1)$, $\forall v \geq 2$.

Similarly, we use the following terms to approximate $B_{2n}(k-d)$

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

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

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

where $\mathbf{z}_n^*(k) = \left(\sum_{j=0}^{\sqrt{n}-K_n} b_j \epsilon_{n-j}, ..., \sum_{j=0}^{\sqrt{n}-K_n} b_j \epsilon_{n-k+1-j} \right)'$ for $k \geq 1$. To approximate the model misspecification term $S_n(k-d)$, we use the following term

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

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

The following lemmas will be used in the proof of theorems and corollaries, and the proofs of these lemmas are included in the supplementary material.

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

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

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

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

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

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

$$(i) \quad \lim_{n\to\infty} \max_{1\leq k\leq K_n} \left| E(N(f_{2,n,\infty}^*(k))^2) - k\sigma^2 \right| = 0,$$

$$(ii) \quad \lim_{n\to\infty} \max_{1\leq k,l\leq K_n} \left| E(N f_{2,n,\infty}^*(k) f_{2,n,\infty}^*(l)) - \min(k,l)\sigma^2 \right| = 0. \tag{A.2}$$

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

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathcal{H}_n^d} \left| \frac{\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))\big]}{L_n^d(\boldsymbol{w})} \right| = 0. \tag{A.3}$$

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

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathcal{H}_n^d} \left| \frac{\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} (w_k f_{1,n}(d) + w_k f_{2,n}(k-d) + w_k S_n(k-d))\big]^2}{L_n^d(\boldsymbol{w})} - 1 \right| = 0. \tag{A.4}$$

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

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathcal{H}_n^d} \left| \frac{E(f_n(k,d), S_n(k-d), \boldsymbol{w}) - E(F_n(k,d), S_n(k-d), \boldsymbol{w})}{L_n^d(\boldsymbol{w})} \right| = 0, \tag{A.5}$$

where

$$E(f_n(k,d), S_n(k-d), \boldsymbol{w}) = \mathrm{E}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k f_n(k) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\Big]^2,$$

$$E(F_n(k,d), S_n(k-d), \boldsymbol{w}) = \mathrm{E}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k F_n(k,d) + \sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\Big]^2,$$

and $F_n(k,d) = f_{1,n}(d) + f_{2,n}(k-d)$.

Lemma 7. *Let* $\hat{w}_{\mathrm{MMA},k}$, $\hat{w}_{\mathrm{SMA},k}$, *and* $\hat{w}_{\mathrm{AMA},k}$ *be the* k*th element of* $\hat{\boldsymbol{w}}_{\mathrm{MMA}}$, $\hat{\boldsymbol{w}}_{\mathrm{SMA}}$, *and* $\hat{\boldsymbol{w}}_{\mathrm{AMA}}$, *respectively. For any* $1 \le k < d$ *and* $2 < q_1 < \max\{3, q\}$, *we have (i)* $Pr(\hat{w}_{\mathrm{MMA},k} > 0) = O(n^{-q_1/2})$, *(ii)* $Pr(\hat{w}_{\mathrm{SMA},k} > 0) = O(n^{-q_1/2})$, *and (iii)* $Pr(\hat{w}_{\mathrm{AMA},k} > 0) = O(n^{-q_1/2})$.

The following lemma extends Theorem 4.2 of Shibata (1980) from MS to MA.

Lemma 8.
Suppose there is another model averaging criterion $\tilde{S}_n(\boldsymbol{w})$, *which is a function of model averaging weights. Define* $G_n(\boldsymbol{w}) = C_n(\boldsymbol{w}) - g(\tilde{S}_n(\boldsymbol{w}))$, *where* $g(\cdot)$ *is a increasing function, and* $C_n(\boldsymbol{w})$ *is the Mallows model averaging criterion. Let* $\xi_n^d = \inf_{\boldsymbol{w}\in\mathcal{H}_n^d} L_n^d(\boldsymbol{w}) = L_n^d(\boldsymbol{w}_n^*)$ *and* $\hat{\boldsymbol{w}}_{\tilde{S}_n}^d = \arg\min_{\boldsymbol{w}\in\mathcal{H}_n^d} \tilde{S}_n(\boldsymbol{w})$. *Suppose that Assumptions 1-5 hold. If*

$$\lim_{n\to\infty} \sup_{\boldsymbol{w}\in\mathcal{H}_n^d} \left| \frac{G_n(\boldsymbol{w}) - G_n(\boldsymbol{w}_n^*)}{N L_n^d(\boldsymbol{w})} \right| \xrightarrow{p} 0, \tag{A.6}$$

then we have

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

B Proofs of Theorems and Corollaries

Proof of Theorem 1. Based on the MSPE decomposition in Appendix A, for any $\mathbf{w} \in \mathcal{H}_n^d$,

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

Observe that

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

Thus, Theorem 1 holds by Lemmas 5 and 6. This completes the proof. □

Proof of Corollary 1. Without loss of generality, we 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 construct the averaging prediction based on these two models only. Let $\mathbf{w}_{k_1,k_2}$ be the associated weight vector such that $\mathbf{w}_{k_1,k_2} = (0, ..., w_{k_1}, ..., w_{k_2}, ..., 0)' \in \mathcal{H}_n^d$ with $w_{k_1} + w_{k_2} = 1$. By Theorem 1, we have

$$\begin{aligned}
L_n^d(\mathbf{w}_{k_1,k_2}) = {} & w_{k_1}^2\frac{k_1}{N} + (1-w_{k_1})^2\frac{k_2}{N} + 2w_{k_1}(1-w_{k_1})\frac{k_1}{N} + w_{k_1}^2\|\mathbf{a}-\mathbf{a}(k_1-d)\|_z^2 \\
& + (1-w_{k_1})^2\|\mathbf{a}-\mathbf{a}(k_2-d)\|_z^2 + 2w_{k_1}(1-w_{k_1})\|\mathbf{a}-\mathbf{a}(k_2-d)\|_z^2.
\end{aligned}$$

Thus, the MSPE of the MA prediction is strictly less than the MS predictor if

$$L_n^d(\mathbf{w}_{k_1,k_2}) < \frac{k_1}{N} + \|\mathbf{a}-\mathbf{a}(k_1-d)\|_z^2, \tag{B.1}$$

and

$$L_n^d(\mathbf{w}_{k_1,k_2}) < \frac{k_2}{N} + \|\mathbf{a}-\mathbf{a}(k_2-d)\|_z^2. \tag{B.2}$$

By some algebra, (B.1) and (B.2) can be rewritten as

$$(1-w_{k_1})^2\,\frac{k_2-k_1}{N} < (1-w_{k_1}^2)\big[\,\|\mathbf{a}-\mathbf{a}(k_1-d)\|_z^2 - \|\mathbf{a}-\mathbf{a}(k_2-d)\|_z^2\,\big], \tag{B.3}$$

and

$$w_{k_1}^2\big[\,\|\mathbf{a}-\mathbf{a}(k_1-d)\|_z^2 - \|\mathbf{a}-\mathbf{a}(k_2-d)\|_z^2\,\big] < 2w_{k_1}(1-w_{k_1})\frac{k_2-k_1}{N}. \tag{B.4}$$

Based on the quadratic formula, (B.3) implies w_{k_1} must lie 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 (B.4) 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 $C(k_1,k_2,n) := N^{-1}(k_2 - k_1)$ and $B(k_1,k_2) := \|\mathbf{a} - \mathbf{a}(k_1 - d)\|_z^2 - \|\mathbf{a} - \mathbf{a}(k_2 - d)\|_z^2$.

Note that $C(k_1,k_2,n)$ is always greater than zero. If $B(k_1,k_2) > 0$, then

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

and there exists a weight 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 MS predictors of AR(k_1) and AR(k_2), respectively. Note that we do not restrict the relationship between $C(k_1,k_2,n)$ and $B(k_1,k_2)$, and either $C(k_1,k_2,n) \geq B(k_1,k_2)$ or $C(k_1,k_2,n) \leq B(k_1,k_2)$ is allowed. When $C(k_1,k_2,n) \geq B(k_1,k_2)$, we have

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

which implies that AR(k_1) generates a smaller MSPE than AR(k_2). Similarly, AR(k_2) generates a smaller MSPE than AR(k_1) when $C(k_1,k_2,n) \leq B(k_1,k_2)$.

By condition (4.6), 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 $K_n - \max(1,d)$ numbers 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 $\mathcal{P}_n(\mathbf{w}^{\circ}_{k,l})$ as 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 $\mathcal{H}^d_n \backslash V(\mathcal{H}^d_n)$ 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, $L^d_n(\mathbf{w}^{\diamond}_n) < \min_{\mathbf{w} \in \mathcal{V}(\mathcal{H}^d_n)} L^d_n(\mathbf{w})$. This completes the proof. □

Proof of Corollary 2. Define $\varphi_1 = 1$ and $\varphi_j = \sum_{j=2}^{K_n} w_j$ for any $\mathbf{w} = (w_1, w_2, ..., w_{K_n}) \in H^d_n$. By some algebra, (4.2) can be rewritten as

$$L^d_n(\mathbf{w}) = \frac{\sigma^2 d^2}{N} + \frac{\sigma^2 \max(1,d)}{N} + A_{K_n} + \sum_{j=\max(1,d)+1}^{K_n} \varphi_j^2 \frac{\sigma^2}{N} + \sum_{j=\max(1,d)+1}^{K_n} (1 - \varphi_j)^2 [A_{j-1} - A_j]. \tag{B.5}$$

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 (B.5). This completes the proof. □

Proof of Theorem 2. We first show $\|\hat{\mathbf{w}}_{\text{MMA}} - \hat{\mathbf{w}}_{\text{MMA}}^d\|_2 \xrightarrow{a.s.} 0$. Note that

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

Thus, we have

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

By Lemma 7 (i), we have $\Pr\big(\|\hat{\mathbf{w}}_{\text{MMA}} - \hat{\mathbf{w}}_{\text{MMA}}^d\|_2 \big) = O(n^{-q_1/2})$, $2 < q_1 < \max(3, q)$, which is summable. Then the result holds by Borel-Cantelli Lemma.

We next show that $\lim_{n\to\infty} L_n^d(\hat{\mathbf{w}}_{\text{MMA}}^d)/L_n^d(\mathbf{w}_n^*) \xrightarrow{p} 1$. Recall that $\hat{\mathbf{w}}_{\text{MMA}}^d = \arg\min_{\mathbf{w}\in\mathcal{H}_n^d} C_n(\mathbf{w})$ and $L_n^d(\mathbf{w}_n^*) = \inf_{\mathbf{w}\in\mathcal{H}_n^d} L_n^d(\mathbf{w})$. Thus, we have

$$\begin{aligned}0 \ge C_n(\hat{\mathbf{w}}_{\text{MMA}}^d) - C_n(\mathbf{w}_n^*) &= NL_n^d(\hat{\mathbf{w}}_{\text{MMA}}^d) - NL_n^d(\mathbf{w}_n^*) - V_n(\hat{\mathbf{w}}_{\text{MMA}}^d, \mathbf{w}_n^*), \\ V_n(\hat{\mathbf{w}}_{\text{MMA}}^d, \mathbf{w}_n^*) &\ge NL_n^d(\hat{\mathbf{w}}_{\text{MMA}}^d) - NL_n^d(\mathbf{w}_n^*) \ge 0, \\ \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{V_n(\mathbf{w}, \mathbf{w}_n^*)}{NL_n^d(\mathbf{w})} \right| &\ge \frac{V_n(\hat{\mathbf{w}}_{\text{MMA}}^d, \mathbf{w}_n^*)}{NL_n^d(\hat{\mathbf{w}}_{\text{MMA}}^d)} \ge 1 - \frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\hat{\mathbf{w}}_{\text{MMA}}^d)} \ge 0.\end{aligned}$$

Therefore, if

$$\lim_{n\to\infty} \sup_{\mathbf{w}\in\mathcal{H}_n^d} \left| \frac{V_n(\mathbf{w}, \mathbf{w}_n^*)}{NL_n^d(\mathbf{w})} \right| \xrightarrow{p} 0, \tag{B.6}$$

we have

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

For a vector $\mathbf{v}$ and a positive definite matrix Q, define $\|\mathbf{v}\|_Q^2 = \mathbf{v}'Q\mathbf{v}$. Inspired by Eq. (4.1) of Ing et al. (2012) and Theorem 1 in this paper, for all $\mathbf{w} \in \mathcal{H}_n^d$, $C_n(\mathbf{w})$ can be decomposed as below:

$$C_n(\mathbf{w}) = NL_n^d(\mathbf{w}) + \mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}(\check{\sigma}^2 - \sigma^2) + \mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}(\check{\sigma}^2 - \sigma^2) + (N + d - d^2)\sigma^2 + N\check{\sigma}^2$$

$$+\Big(\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j\big[(\max(i,j)-d)\sigma^2 - \|N^{-1/2}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(\max(i,j))\epsilon_{j+1,\max(i,j)-d}\|^2_{\hat\Omega_n^{-1}(\max(i,j))}\big]\Big)$$
$$+\big(N\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j[\hat\Sigma_n^2(\max(i,j)-d)-\sigma^2(\max(i,j)-d)]\big), \tag{B.7}$$

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 $\|\mathbf{a}-\mathbf{a}(k)\|_z^2$, $\epsilon_{j+1,k}$, and $\hat\Omega_n^{-1}(k)$ are defined after section 4.2 and (A.1).

In view of (B.7), we first rewrite $(NL_n^d(\mathbf{w}))^{-1}(C_n(\mathbf{w})-C_n(\mathbf{w}_n^*))$ as

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

Next, $(NL_n^d(\mathbf{w}))^{-1}V_n(\mathbf{w},\mathbf{w}_n^*)$ can be decomposed into seven parts:

$$V_{1n}(\mathbf{w})=-\frac{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}(\check\sigma^2-\sigma^2)}{NL_n^d(\mathbf{w})},\quad V_{2n}(\mathbf{w},\mathbf{w}_n^*)=-\frac{\mathbf{w}_n^{*\prime}\Pi_{\min}(K_n)\mathbf{w}_n^*(\check\sigma^2-\sigma^2)}{NL_n^d(\mathbf{w})},$$
$$V_{3n}(\mathbf{w})=-\frac{\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}(\check\sigma^2-\sigma^2)}{NL_n^d(\mathbf{w})},\quad V_{4n}(\mathbf{w},\mathbf{w}_n^*)=-\frac{\mathbf{w}_n^{*\prime}\Pi_{\max}(K_n)\mathbf{w}_n^*(\check\sigma^2-\sigma^2)}{NL_n^d(\mathbf{w})},$$
$$V_{5n}(\mathbf{w})=-\frac{1}{NL_n^d(\mathbf{w})}\Big(\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j\big[(\max(i,j)-d)\sigma^2 - \|N^{-1/2}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(\max(i,j))\epsilon_{j+1,\max(i,j)-d}\|^2_{\hat\Omega_n^{-1}(\max(i,j))}\big]\Big),$$
$$V_{6n}(\mathbf{w},\mathbf{w}_n^*)=-\frac{1}{NL_n^d(\mathbf{w})}\Big(\sum_{\max(1,d)\le i,\, j\le K_n} w_{n,i}^* w_{n,j}^*\big[(\max(i,j)-d)\sigma^2 - \|N^{-1/2}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(\max(i,j))\epsilon_{j+1,\max(i,j)-d}\|^2_{\hat\Omega_n^{-1}(\max(i,j))}\big]\Big),$$
$$V_{7n}(\mathbf{w},\mathbf{w}_n^*)=-\frac{\sum_{\max(1,d)\le i,\, j\le K_n}(w_iw_j-w_{n,i}^*w_{n,j}^*)[\hat\Sigma^2(\max(i,j)-d)-\sigma^2(\max(i,j)-d)]}{L_n^d(\mathbf{w})}.$$

Observe that

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

Thus, 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 (B.6) is automatically satisfied.

We now show each term is $o_p(1)$. By Eq. (4.6) of Ing et al. (2012), for any $k\ge\max(1,d)$,

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

Similar to Lemma 7 (i), without loss of generality, let $\check{\sigma}^2 = \hat{\sigma}^2(K_n)$. Then,

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

By Lemma 4.1 and Eq. (4.8) of Ing et al. (2012), $\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j = 1$, and $\|\mathbf{a} - \mathbf{a}(v)\|_z^2 \le \|\mathbf{a} - \mathbf{a}(l)\|_z^2$, $v \ge l$, we have

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

Then, it follows that

$$
\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{1n}(\mathbf{w})| = O_p\Big(\frac{1}{\sqrt{N}} + \frac{K_n}{N}\Big). \tag{B.8}
$$

Similarly,

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

Thus, by (B.8), we have

$$
\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{2n}(\mathbf{w}, \mathbf{w}_n^*)| = O_p\Big(\frac{1}{\sqrt{N}} + \frac{K_n}{N}\Big). \tag{B.9}
$$

Similar to $V_{1n}(\mathbf{w})$, we can rewrite $|V_{3n}(\mathbf{w})|$ as

$$
\begin{aligned}
|V_{3n}(\mathbf{w})| &= \Big| \frac{(\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}) \ \sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j [\hat{\sigma}^2(K_n) - \sigma^2]}{NL_n^d(\mathbf{w})} \Big|, \\
&= (\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}) \Big| \frac{\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j [\hat{\Sigma}_n^2(K_n - d) - \sigma^2(K_n - d)]}{NL_n^d(\mathbf{w})} \Big| \\
&+ (\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w})
\end{aligned}
$$

$$\times \Big| \frac{\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j \|N^{-1}\sum_{j=K_n}^{n-1} \mathbf{s}_{j,n}(K_n)\epsilon_{j+1,K_n-d}\|^2_{\hat\Omega_n^{-1}(K_n))}}{NL_n^d(\mathbf{w})}\Big|$$

$$+ (\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w})\Big|\frac{\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j \|\mathbf{a}-\mathbf{a}(K_n-d)\|_z^2}{NL_n^d(\mathbf{w})}\Big|$$

$$= (I^*) + (II^*) + (III^*),$$

By Lemma 4.1 and Eq. (4.8) of Ing et al. (2012), $\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j = 1$, and $\|\mathbf{a} - \mathbf{a}(v)\|_z^2 \le \|\mathbf{a}-\mathbf{a}(l)\|_z^2$, $v \ge l$, we have

$$(I^*) = O_p(\frac{\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}}{N^{3/2}L_n^d(\mathbf{w})}) = O_p(\frac{K_n}{N^{3/2}L_n^d(\mathbf{w})}),$$

$$(II^*) = O_p(\frac{K_n\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}}{N^2L_n^d(\mathbf{w})}) = O_p(\frac{K_n^2}{N^2L_n^d(\mathbf{w})}),$$

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

Then, it follows that

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{3n}(\mathbf{w})| = O_p(\frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n}{\sqrt{N}} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N} + \frac{K_n}{N}). \tag{B.10}$$

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

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{4n}(\mathbf{w},\mathbf{w}_n^*)| = O_p(\frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n}{\sqrt{N}} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N} + \frac{K_n}{N}). \tag{B.11}$$

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 $d \le k \le K_n$

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

$$\le \Big|(k-d)\sigma^2 - \|N^{-1/2}\sum_{j=K_n}^{n-1}\mathbf{z}_j(k-d)\epsilon_{j+1,k-d}\|^2_{\Gamma^{-1}(k-d)}\Big|1(k>d)$$

$$+ \|N^{-1/2}\sum_{j=K_n}^{n-1}U_{j,n}(d)\epsilon_{j+1,k-d}\|^2\|\hat\Omega_n^{-1}(d)\|$$

$$+ \|N^{-1/2}\sum_{j=K_n}^{n-1}\mathbf{s}_{j,n}(k)\epsilon_{j+1,k-d}\|^2\|\hat\Omega_n^{-1}(k) - \hat\Omega_{d,n}(k)\|.$$

Therefore, we have

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

By Eq. (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, we have

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

Then, it follows that

$$
\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{5n}(\mathbf{w})| = O_p(\frac{K_n^{1/2}}{NL_n^d(\mathbf{w}_n^*)} + \frac{1}{NL_n^d(\mathbf{w}_n^*)} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N^{1/2}}). \tag{B.12}
$$

Similarly,

$$
\sup_{\mathbf{w}\in\mathcal{H}_n^d} |V_{6n}(\mathbf{w},\mathbf{w}_n^*)| = O_p(\frac{K_n^{1/2}}{NL_n^d(\mathbf{w}_n^*)} + \frac{1}{NL_n^d(\mathbf{w}_n^*)} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N^{1/2}}). \tag{B.13}
$$

Since $\sum_{k=\max(1,d)}^{K_n} w_k = 1$ and $\sum_{k=\max(1,d)}^{K_n} w_k^* = 1$, $|V_{7n}(\mathbf{w},\mathbf{w}_n^*)|$ can be decomposed as

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

By Eq. (4.4) of Ing et al. (2012), we have

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

Then, it follows that

$$
\sup_{\mathbf{w}\in\mathcal{H}_n^d}|V_{7n}(\mathbf{w},\mathbf{w}_n^*)| = O_p\Big(\frac{1}{(NL_n^d(\mathbf{w}_n^*))^{1/2}}\Big). \tag{B.14}
$$

By (B.8)-(B.14), $\lim_{n\to\infty} N\xi_n^d \to \infty$, and Assumptions 4 and 5, we have

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

Thus, (B.6) is satisfied and $\lim_{n\to\infty} L_n^d(\hat{\mathbf{w}}_{\mathrm{MMA}}^d)/L_n^d(\mathbf{w}_n^*) \xrightarrow{p} 1$ holds. This completes the proof. □

Proof of Theorem 3. For (i), note that

$$
L_n^d(\mathbf{w}_{1,k}^*) = \frac{\sigma^2 d^2}{N} + \sigma^2\frac{k_n^*}{N} + A_{k_n^*}.
$$

Then, by the algebraic-decay condition and the argument as giving in Eq. (A.9) in Ing and Wei (2005), we have

$$
k_n^* = O(N^{1/(\alpha+1)}) \ \text{ and } \ L_n^d(\mathbf{w}_{1,k}^*) = O(N^{-\alpha/(\alpha+1)}), \tag{B.15}
$$

where $\mathbf{w}_{1,k}^*$ is the optimal unit weight vector defined after Section 4.4, and k_n^* is the optimal order for (4.5) under MS. In other words, the k_n^*th element of $\mathbf{w}_{1,k}^*$ equals one and others are zeros.

Observe that

$$
\begin{aligned}
\Delta_n &= L_n^d(\mathbf{w}_{1,k}^*) - L_n^d(\mathbf{w}_n^*)\\
&= \sum_{j=\max(1,d)+1}^{k_n^*}\Bigg[\frac{\sigma^2}{N}\Bigg(1-\frac{A_{j-1}-A_j}{\frac{\sigma^2}{N}+A_{j-1}-A_j}\Bigg)\Bigg] + \sum_{j=k_n^*+1}^{K_n}\Bigg[(A_{j-1}-A_j)\Bigg(1-\frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N}+A_{j-1}-A_j}\Bigg)\Bigg]\\
&= (I)+(II)
\end{aligned} \tag{B.16}
$$

and $\Delta_n \le L_n^d(\mathbf{w}_{1,k}^*)$. To show $\Delta_n \asymp L_n^d(\mathbf{w}_{1,k}^*)$, it is sufficient to show that $(I) \ge c\,L_n^d(\mathbf{w}_{1,k}^*)$, where c is a positive constant greater than zero. Since $A_j = C(j-d)^{-\alpha}$, we have

$$
(I) = \frac{\sigma^2}{N}\sum_{j=\max(1,d)+1}^{k_n^*}\Bigg(\frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N}+A_{j-1}-A_j}\Bigg) = \frac{\sigma^2}{N}\sum_{j=\max(1,d)+1}^{k_n^*}\Bigg(\frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N}+C(j-1-d)^{-\alpha}-C(j-d)^{-\alpha}}\Bigg)
$$

$$
\begin{aligned}
&\geq C\frac{\sigma^2}{N}\sum_{j=\max(1,d)+1}^{k_n^*}\left(\frac{\sigma^2(j-1-d)^\alpha}{\sigma^2(j-1-d)^\alpha+N\big[1-\big(1-\frac{1}{j-d}\big)^\alpha\big]}\right)\\
&\geq C\frac{\sigma^2}{N}\sum_{j=\max(1,d)+1}^{k_n^*}\left(\frac{\sigma^2(j-1-d)^\alpha(j-d)}{\sigma^2(j-1-d)^\alpha(j-d)+N}\right)\\
&\geq C\frac{\sigma^2}{N}\sum_{j=\max(1,d)+1}^{k_n^*}\left(\frac{\sigma^2(j-1-d)^{\alpha+1}}{\sigma^2(k_n^*)^{\alpha+1}+N}\right)\\
&\geq C\frac{\sigma^2}{N}\frac{1}{\sigma^2(k_n^*)^{\alpha+1}+N}\sum_{j=\max(1,d)+1}^{k_n^*}(j-1-d)^{\alpha+1}\\
&\geq C\frac{\sigma^2}{N}\frac{1}{\sigma^2(k_n^*)^{\alpha+1}+N}\big[\,(k_n^*-1-d)^{\alpha+2}\,\big]\geq C\frac{k_n^*}{N}=CN^{-\alpha/(\alpha+1)},
\end{aligned}
\tag{B.17}
$$

where the second inequality is insured by $1-(1-x)^p\leq Cx,\ \ p>0,\ \ 0<x<1$, and the last inequality holds by $k_n^*=O(N^{1/(\alpha+1)})$. By (B.15)-(B.17) and $\Delta_n\leq L_n^d(\mathbf{w}_{1,k}^*)$, we have $\Delta_n\asymp L_n^d(\mathbf{w}_{1,k}^*)$ under the algebraic-decay scenario.

For (ii), by the exponential-decay condition and the argument as Eq. (A.1)-(A.5) in Ing and Wei (2005), we have

$$
k_n^*=O(\frac{1}{\alpha}\log(N))\ \text{ and }\ L_n^d(\mathbf{w}_{1,k}^*)=O(\frac{\frac{1}{\alpha}\log(N)}{N}).
\tag{B.18}
$$

To show $\Delta_n=o(L_n^d(\mathbf{w}_{1,k}^*))$, it is sufficient to show that (I) and (II) in (B.16) are $o(N^{-1}\log(N))$. Since d is finite by Assumption 1, $A_j=C\exp(-\alpha(j-d))=C\exp(-\alpha(j))$, and $A_{j-1}-A_j=C(1-\exp(-\alpha))^{-1}\exp(\alpha(j))$, we have

$$
\begin{aligned}
(I)&=\frac{\sigma^2}{N}\sum_{j=\max(1,d)+1}^{k_n^*}\left(\frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N}+A_{j-1}-A_j}\right)\leq C\big(\frac{\sigma^2}{N}\big)^2\sum_{j=\max(1,d)+1}^{k_n^*}\left(\frac{1}{A_{j-1}-A_j}\right)\\
&\leq C\big(\frac{\sigma^2}{N}\big)^2\big(\frac{1}{1-\exp(-\alpha)}\big)\frac{\exp(\alpha k_n^*)-\exp(\alpha\max(1,d))}{e-1}=O(\frac{1}{N}),
\end{aligned}
\tag{B.19}
$$

and

$$
\begin{aligned}
(II)&=\sum_{j=k_n^*+1}^{K_n}\left[(A_{j-1}-A_j)\left(1-\frac{\frac{\sigma^2}{N}}{\frac{\sigma^2}{N}+A_{j-1}-A_j}\right)\right]=C\sum_{j=k_n^*+1}^{K_n}\left[\left(\frac{(A_{j-1}-A_j)^2}{\frac{\sigma^2}{N}+A_{j-1}-A_j}\right)\right]\\
&\leq C\sum_{j=k_n^*+1}^{K_n}\Big(A_{j-1}-A_j\Big)\leq C\frac{1}{1-\exp(-\alpha)}\ \times\ \Big(\exp(-\alpha k_n^*)+...+\exp(-\alpha(K_n-1))\Big)\\
&\leq C\exp(-\alpha k_n^*)\Big[1-\exp(-\alpha(K_n-k_n^*+1))\Big]=O(\frac{1}{N}).
\end{aligned}
\tag{B.20}
$$

Thus, by (B.18)-(B.20), we have $\Delta_n=o(L_n^d(\mathbf{w}_{1,k}^*))$. This completes the proof. □

Proof of Corollary 3. First, we show that $L_n^d(\mathbf{w}_n^*) \asymp L_n^d(\mathbf{w}_{1,k}^*)$ under either exponential or algebraic decay. By the representation in Corollary 2, we have

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

for some $c>0$, where the third inequality holds by $N^{-1}\sigma^2 < C(A_{j-1}-A_j)$, $j=\max(1,d)+1,...,k_n^*$ for some large enough C under either exponential or algebraic decay. Hence, $L_n^d(\mathbf{w}_n^*) \asymp L_n^d(\mathbf{w}_{1,k}^*)$ holds under either exponential or algebraic decay.

Next, observe that

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

where the last equality is insured by the condition (4.9) and the fact that $L_n^d(\mathbf{w}_n^*) \asymp L_n^d(\mathbf{w}_{1,k}^*)$. Thus, by Theorem 3, we have $\hat{\Delta}_n \asymp L_n^d(\hat{\mathbf{w}}_{\text{MS}}^d)$ and $\hat{\Delta}_n = o(L_n^d(\hat{\mathbf{w}}_{\text{MS}}^d))$ under algebraic and exponential decay, respectively.

We now show that $L_n^d(\hat{\mathbf{w}}_{\text{MA}}^d) \asymp L_n^d(\hat{\mathbf{w}}_{\text{MS}}^d)$. Observe that

$$\frac{L_n^d(\hat{\mathbf{w}}_{\text{MA}}^d)}{L_n^d(\hat{\mathbf{w}}_{\text{MS}}^d)} = \frac{L_n^d(\hat{\mathbf{w}}_{\text{MA}}^d)}{L_n^d(\mathbf{w}_n^*)}\frac{L_n^d(\mathbf{w}_n^*)}{L_n^d(\mathbf{w}_{1,k}^*)}\frac{L_n^d(\mathbf{w}_{1,k}^*)}{L_n^d(\hat{\mathbf{w}}_{\text{MS}}^d)}.$$

Then, by the condition (4.9) and $L_n^d(\mathbf{w}_n^*) \asymp L_n^d(\mathbf{w}_{1,k}^*)$, we have $L_n^d(\hat{\mathbf{w}}_{\text{MA}}^d) \asymp L_n^d(\hat{\mathbf{w}}_{\text{MS}}^d)$ under either exponential or algebraic decay. This completes the proof. □

Proof of Theorem 4.

Part I: Shibata model averaging criterion

The strategy of the proof is to show that Shibata model averaging weights $\hat{\mathbf{w}}_{\text{SMA}}$ satisfy (4.7) and (4.8). First, $\hat{\mathbf{w}}_{\text{SMA}}$ satisfy (4.7) by the same arguments as the proofs of the first part of Theorem 2 and Lemma 7 (ii). To show that $\hat{\mathbf{w}}_{\text{SMA}}^d$ satisfy (4.8), we will check the condition (A.6) in Lemma 8 holds. Note that the difference between Shibata model averaging criterion and Mallows model averaging criterion is

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

Then, it follows that

$$|G_n(\mathbf{w}) - G_n(\mathbf{w}_n^*)| \leq |(\mathbf{w}'[\Pi_{\min}(K_n)+\Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2-\hat{\sigma}_w^2)|$$

$$+ |(\mathbf{w}_n^{*\prime}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*)(\check{\sigma}^2 - \hat{\sigma}_{w^*}^2)|, \tag{B.21}$$

where

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

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

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

We next show that

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2 - \hat{\sigma}_w^2)}{NL_n^d(\mathbf{w})}\Big| = o_p(1). \tag{B.22}$$

Observe that

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

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

$$\begin{aligned}&\sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2 - \hat{\sigma}_w^2)}{NL_n^d(\mathbf{w})}\Big| \\ \le& \sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w})(\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j[\hat{\sigma}^2(\max(i,j)) - \sigma^2])}{NL_n^d(\mathbf{w})}\Big| \\ &+ \sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w})(\check{\sigma}^2 - \sigma^2)}{NL_n^d(\mathbf{w})}\Big| \\ &+ \sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w})(\sum_{\max(1,d)\le i,\, j\le K_n} w_i w_j[\hat{\sigma}^2(\max(i,j)) - \sigma^2])}{NL_n^d(\mathbf{w})}\Big| \\ &+ \sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{(\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w})(\check{\sigma}^2 - \sigma^2)}{NL_n^d(\mathbf{w})}\Big| \\ \le& C\Big(\sup_{\mathbf{w}\in\mathcal{H}_n^d}|U_{1n}(\mathbf{w})| + |\check{\sigma}^2 - \sigma^2| + \sup_{\mathbf{w}\in\mathcal{H}_n^d}|U_{2n}(\mathbf{w})| + \sup_{\mathbf{w}\in\mathcal{H}_n^d}\Big|\frac{\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}}{NL_n^d(\mathbf{w})}(\check{\sigma}^2 - \sigma^2)\Big|\Big),\end{aligned} \tag{B.23}$$

where

$$U_{1n}(\mathbf{w}) = \frac{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}(\hat{\sigma}_w^2 - \sigma^2)}{NL_n^d(\mathbf{w})} \text{ and } U_{2n}(\mathbf{w}) = \frac{\mathbf{w}'\Pi_{\max}(K_n)\mathbf{w}(\hat{\sigma}_w^2 - \sigma^2)}{NL_n^d(\mathbf{w})}.$$

By Eq. (4.6) of Ing et al. (2012), for any $k \geq \max(1, d)$, we have

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

Then, it follows that

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

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

$$(\tilde{I}) = O_p(\frac{1}{\sqrt{N}}), \ \ (\tilde{II}) = O_p(\frac{K_n}{N}), \ \text{ and } \ (I\tilde{I}I) \leq C\frac{K_n}{N}.$$

Then, it follows that

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d} |U_{1n}(\mathbf{w})| = O_p(\frac{1}{\sqrt{N}} + \frac{K_n}{N}). \tag{B.24}$$

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

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

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

$$(I^{**}) = O_p(\frac{K_n}{N^{3/2}L_n^d(\mathbf{w})}), \ \ (II^{**}) = O_p(\frac{K_n^2}{N^2L_n^d(\mathbf{w})}), \ \text{ and } \ (III^{**}) \leq C\frac{K_n}{N}.$$

Then, it follows that

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d} |U_{2n}(\mathbf{w})| = O_p(\frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n}{\sqrt{N}} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N} + \frac{K_n}{N}). \tag{B.25}$$

By (B.23)-(B.25), similar arguments for (B.25), and the fact that $\check{\sigma}^2 = \hat{\sigma}^2(K_n)$ is consistent for σ^2, we have

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_n^d} \Big| \frac{(\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2 - \hat{\sigma}_w^2)}{NL_n^d(\mathbf{w})} \Big| \\
&\leq C\big(O_p(\frac{1}{\sqrt{N}} + \frac{K_n}{N}) + o_p(1) + O_p(\frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n}{\sqrt{N}} + \frac{1}{NL_n^d(\mathbf{w}_n^*)}\frac{K_n^2}{N} + \frac{K_n}{N})\big).
\end{aligned}$$

Therefore, (B.22) holds.

Note that

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

Therefore, by (B.22), we have

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d} \Big| \frac{(\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*)(\check{\sigma}^2 - \hat{\sigma}_{w^*}^2)}{NL_n^d(\mathbf{w})} \Big| = o_p(1). \tag{B.26}$$

Thus, by (B.22) and (B.26), $S_n(\mathbf{w}) - C_n(\mathbf{w})$ satisfies the condition (A.6) of Lemma 8, which implies that $\hat{\mathbf{w}}_{\text{SMA}}^d$ satisfies (4.8). Therefore, without knowing the integration order, the Shibata model averaging estimator is asymptotically optimal in the sense of achieving (4.7) and (4.8).

Part II: Akaike model averaging criterion

Similar to the Shibata model averaging estimator, we will show that the Akaike model averaging weights $\hat{\mathbf{w}}_{\text{AMA}}$ satisfy (4.7) and (4.8). First, $\hat{\mathbf{w}}_{\text{AMA}}$ satisfy (4.7) by the same arguments as the proofs of the first part of Theorem 2 and Lemma 7 (iii). To show that $\hat{\mathbf{w}}_{\text{AMA}}^d$ satisfy (4.8), we will check whether the condition (A.6) in Lemma 8 holds. Let $g(x) = N\exp(x)$. The difference between the Mallows model averaging criterion and the transformation of the Akaike model averaging criterion is

$$G_n(\mathbf{w}) = C_n(\mathbf{w}) - g(A_n(\mathbf{w}))$$

$$= (\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2 - \hat{\sigma}_w^2) - N\check{\sigma}^2$$
$$+ N\hat{\sigma}_w^2\left(1 + \frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N} - \exp(\frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N})\right).$$

Then, it follows that

$$\begin{aligned}
&|G_n(\mathbf{w}) - G_n(\mathbf{w}_n^*)| \\
\leq&|(\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w})(\check{\sigma}^2 - \hat{\sigma}_w^2)| + |(\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*)(\check{\sigma}^2 - \hat{\sigma}_{w^*}^2)| \\
&+ \left|N\hat{\sigma}_w^2(1 + \frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N} - \exp(\frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N}))\right| \\
&+ \left|N\hat{\sigma}_{w^*}^2(1 + \frac{\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^{*'}}{N} - \exp(\frac{\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*}{N}))\right| \\
=&VV_{1n}(\mathbf{w}) + VV_{2n}(\mathbf{w}_n^*) + VV_{3n}(\mathbf{w}) + VV_{4n}(\mathbf{w}_n^*). \qquad \text{(B.27)}
\end{aligned}$$

By (B.21), (B.22), and (B.26), we have

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{VV_{1n}(\mathbf{w})}{NL_n^d(\mathbf{w})}\right| = o_p(1) \text{ and } \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{VV_{2n}(\mathbf{w}_n^*)}{NL_n^d(\mathbf{w})}\right| = o_p(1). \qquad \text{(B.28)}$$

For sufficiently large n, it follows that

$$\begin{aligned}
&\left|VV_{3n}(\mathbf{w}) + VV_{4n}(\mathbf{w}_n^*)\right| \\
\leq&\left|N\hat{\sigma}_w^2(\frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N})^2\right| + \left|N\hat{\sigma}_{w^*}^2(\frac{\mathbf{w}_n^*[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*}{N})^2\right| \\
\leq&\left|N(\hat{\sigma}_w^2 - \sigma^2)(\frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N})^2\right| + \left|N\sigma^2(\frac{\mathbf{w}'[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}}{N})^2\right| \\
&+ \left|N(\hat{\sigma}_{w^*}^2 - \sigma^2)(\frac{\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*}{N})^2\right| + \left|N\sigma^2(\frac{\mathbf{w}_n^{*'}[\Pi_{\min}(K_n) + \Pi_{\max}(K_n)]\mathbf{w}_n^*}{N})^2\right|, \qquad \text{(B.29)}
\end{aligned}$$

where the first inequality holds by $|1 + x - \exp(x)| \leq |x|^2$ if $|x| \leq 1$.

Therefore, by (B.29) and similar arguments on $V_{2n}(\mathbf{w}, \mathbf{w}_n^*)$, we have

$$\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{VV_{3n}(\mathbf{w}) + VV_{4n}(\mathbf{w}_n^*)}{NL_n^d(\mathbf{w})}\right| \leq C(\sup_{\mathbf{w}\in\mathcal{H}_n^d}|U_{1n}(\mathbf{w})| + \sup_{\mathbf{w}\in\mathcal{H}_n^d}|U_{2n}(\mathbf{w})| + \frac{K_n^2}{N}\frac{1}{NL_n^d(\mathbf{w}_n^*)}), \qquad \text{(B.30)}$$

where $U_{1n}(\mathbf{w})$ and $U_{2n}(\mathbf{w})$ are defined after (B.23).

Thus, by (B.24)-(B.25), (B.27)-(B.30), and Assumption 4, it follows that

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

which implies that $\hat{\mathbf{w}}_{\text{AMA}}^d$ satisfy (4.8) by Lemma 8. This completes the proof. □

References

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics 22*(1), 203–217.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

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

Ando, T. and K.-C. Li (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.

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

Chan, N. H. and C. Z. Wei (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *The Annals of Statistics*, 367–401.

Charkhi, A., G. Claeskens, and B. E. Hansen (2016). Minimum mean squared error model averaging in likelihood models. *Statistica Sinica 26*(2), 809–840.

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

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

Cheng, X., Z. Liao, and R. Shi (2019). On uniform asymptotic risk of averaging GMM estimators. *Quantitative Economics 10*(3), 931–979.

Claeskens, G., N. L. Hjort, et al. (2008). Model selection and model averaging. *Cambridge Books*.

Gao, Y., X. Zhang, S. Wang, and G. Zou (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. (2008). Least-squares forecast averaging. *Journal of Econometrics 146*(2), 342–350.

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

Hjort, N. L. and G. Claeskens (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 C.-Y. Sin (2006). On prediction errors in regression models with nonstationary regressors. *Lecture Notes-Monograph Series*, 60–71.

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

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

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

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

Kawashima, H. (1980). Parameter estimation of autoregressive integrated processes by least squares. *The Annals of Statistics*, 423–435.

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

Liao, J., G. Zou, Y. Gao, and X. Zhang (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 W.-J. Tsay (2020). Optimal multistep VAR forecast averaging. *Econometric Theory 36*(6), 1099–1126.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics 186*(1), 142–159.

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

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

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

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

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

Rohde, R. A. and Z. Hausfather (2020). The berkeley earth land/ocean temperature record. *Earth System Science Data 12*(4), 3469–3479.

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.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature 58*(3), 644–719.

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

Tiao, G. C. and R. S. Tsay (1983). Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *The Annals of Statistics*, 856–871.

Wan, A. T., X. Zhang, and G. Zou (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics 156*(2), 277–283.

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

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

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

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

Zhang, X. (2021). A new study on asymptotic optimality of least squares model averaging. *Econometric Theory 37*(2), 388–407.

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

Model Averaging Prediction for Possibly Nonstationary Autoregressions

Supplementary Material

Tzu-Chi Lin and Chu-An Liu

Federal Reserve Bank of Philadelphia and Academia Sinica

The supplementary material includes two parts. S.1 contains the proofs of supplementary lemmas, and S.2 provides additional simulation results.

S.1 Proofs of Supplementary Lemmas

Proof of Lemma 1. Observe that

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

Note that (I) is $o(n^{-1})$ by Lemma B.5 of Ing et al. (2010).

We next show that (II) is $o(n^{-1})$. Denote $a_i - a_i(k)$ by $\gamma_i(k)$. 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 it follows that

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

where

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

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

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

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

By Cauchy-Schwarz inequality and Eq. (B.17) of Ing et al. (2010), we have $|(IV)| = o(n^{-1})$. Next, since $z_{n+1-i,\infty} - z_{n+1-i} = \sum_{j=n-i}^{\infty} b_j \epsilon_{n+1-i-j}$, it follows that

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

Thus, by the fact that $|(IV)| = o(n^{-1})$, we have $|(V)| = o(n^{-1})$ and $|(VI)| = o(n^{-1})$.

We now show $|(III)| = o(n^{-1})$. By Eq. (3.2) of Ing and Wei (2003), we have

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

Then, it follows that

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

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

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

where $\chi_{i-j}=\mathrm{E}(z_{i,\infty}z_{j,\infty})$. Thus, we have $(IX)=o(n^{-2})$.

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

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

By (2.2), we have

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

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

and

$$\begin{aligned}&\gamma_1(k)\sum_{i=n+1}^{\infty}\gamma_i(k)\chi_{i-1}+\cdots+\gamma_{\rho n}(k)\sum_{i=n+1}^{\infty}\gamma_i(k)\chi_{i-\rho n}\\ &\le C(\rho n)(\sum_{i=n+1}^{\infty}|a_i|)(\sum_{j=(1-\rho)n+1}^{\infty}|b_j|)=o(n^{-1})\end{aligned}\tag{S.2}$$

By (S.1) and (S.2), we have $(VIII)=o(n^{-1})$. By similar arguments, we have $(VII)=o(n^{-1})$. Since $(I)-(IX)$ are $o(n^{-1})$, the statement of Lemma 1 holds. This completes the proof. □

Proof of Lemma 2. For (i), it suffices to show that

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d}\mathrm{E}\left|\frac{\sqrt{N}}{\sqrt{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}}\big[\sum_{k=\max\{1,d\}}^{K_n}w_k(f_{2,n}(k-d)-f_{2,n}^*(k-d))\big]\right|^2=0.\tag{S.3}$$

Observe that

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

where

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

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

For any $p\ge 2$, by Hölder inequality, we have

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

where

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

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

By Lemma B.3 of Ing et al. (2010), (2.3), and Assumption 3, for all $d<k\le K_n$, we have

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

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

Then, it follows that

$$\mathrm{E}(|A_1(k-d)|^p)\le C\big[(k-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty}b_j^2\big]^{p/2}\le C((K_n-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty}b_j^2)^{p/2}. \tag{S.4}$$

Similarly,

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

where $b_1(k-d)=\mathbf{z}_n^{*'}(k-d)$ and $b_2(k-d)=[N(k-d)]^{-1/2}\sum_{j=n-\sqrt{n}}^{n-1}\mathbf{z}_j(k-d)\epsilon_{j+1}$. By Lemma B.3 of Ing et al. (2010), we have

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

Then, it follows that

$$\mathrm{E}(|A_2(k-d)|^p) \le C(\frac{k-d}{\sqrt{N}})^{p/2} \le C(\frac{K_n-d}{\sqrt{N}})^{p/2}. \tag{S.5}$$

Therefore, by (S.4) and (S.5), we have

$$\mathrm{E}\Big|\Big(\sqrt{\frac{N}{i}}[f_{2,n}(i)-f^*_{2,n}(i)]\Big)\Big(\sqrt{\frac{N}{j}}[f_{2,n}(j)-f^*_{2,n}(j)]\Big)\Big|^p \le C\Big\{[(K_n-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j^2]^{p/2}+(\frac{K_n-d}{\sqrt{N}})^{p/2}\Big\}, \tag{S.6}$$

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

$$\frac{\sum_{\max(1,d)\le i,j\le K_n} w_i w_j \sqrt{i-d}\sqrt{j-d}}{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d} = \frac{\sum_{\max(1,d)\le i,j\le K_n} w_i w_j \sqrt{i-d}\sqrt{j-d}}{\sum_{\max(1,d)\le i,j\le K_n} w_i w_j \min\{i,j\}-d} \le \sqrt{K_n-d}. \tag{S.7}$$

Thus, by (S.6) and (S.7), it follows that

$$\begin{aligned}
&\mathrm{E}\Big|\frac{\sqrt{N}}{\sqrt{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_{2,n}(k-d)-f^*_{2,n}(k-d))\Big]\Big|^2 \\
&\le C\frac{\sum_{\max(1,d)\le i,j\le K_n} w_i w_j \sqrt{i-d}\sqrt{j-d}}{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}\Big\{[(K_n-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j^2]+(\frac{K_n-d}{\sqrt{N}})\Big\} \\
&\le C\sqrt{K_n-d}\Big\{[(K_n-d)\sum_{j=\sqrt{n}-K_n+1}^{\infty} b_j^2]+(\frac{K_n-d}{\sqrt{N}})\Big\} \\
&\le C\sum_{j=\sqrt{n}-K_n+1}^{\infty} |jb_j|^2+\sqrt{\frac{K_n^3}{N}}.
\end{aligned} \tag{S.8}$$

Therefore, (S.3) holds by (S.8), (2.3), and Assumption 4.

For (ii), it suffices to show that

$$\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathcal{H}_n^d} \mathrm{E}\Big|\frac{\sqrt{N}}{\sqrt{\mathbf{w}'\Pi_{\min}(K_n)\mathbf{w}-d}}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f^*_{2,n}(k-d)-f^*_{2,n,\infty}(k-d))\Big]\Big|^2 = 0. \tag{S.9}$$

where

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

and

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

By Hölder inequality, Lemma B.3 of Ing et al. (2010), and Lemma 2 of Wei (1987), we have

$$
\begin{aligned}
&\mathrm{E}\Big|\sqrt{N}\sum_{k=\max\{1,d\}}^{K_n} w_k[f^*_{2,n}(k-d)-f^*_{2,n,\infty}(k-d)]\Big|^p \\
&\le \sum_{k=\max\{1,d\}}^{K_n} w_k\mathrm{E}\Big|\sqrt{N}[f^*_{2,n}(k-d)-f^*_{2,n,\infty}(k-d)]\Big|^p \\
&\le \sum_{k=\max\{1,d\}}^{K_n} w_k(\mathrm{E}\|\mathbf{z}^{*'}_n(k-d)\|^{3p})^{1/3}(\mathrm{E}\|\Gamma^{-1}(k-d)\|^{3p})^{1/3}(\mathrm{E}\|\frac{1}{\sqrt{N}}\sum_{j=K_n}^{n-\sqrt{n}-1}\tilde{\mathbf{z}}_{j,\infty}(k-d)\epsilon_{j+1}\|^{3p})^{1/3} \\
&\le C\max_{1\le k\le K_n-d} k^{p/2}k^{p/2}(N^{-1}\sum_{t=K_n}^{n-\sqrt{n}-1}\mathrm{E}(\tilde{\mathbf{z}}_{t,\infty})^2)^{p/2} \le CK_n^p(N^{-1}\sum_{t=K_n}^{n-\sqrt{n}-1}\sum_{i=t}^{\infty} b_i^2)^{p/2} \\
&\le C(\sum_{i=K_n}^{\infty}|ib_i|^2)^{p/2} \qquad \text{(S.10)}
\end{aligned}
$$

Therefore, (S.9) holds by (S.10) and (2.3). This completes the proof. □

Proof of Lemma 3. The result (i) is a special case of the result (ii) and we omit the proof for brevity. Without loss of generality, we assume that $k < l$. Define

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

where $\Gamma^*(k) = \mathrm{E}(\mathbf{z}^*_t(k)\mathbf{z}^{*'}_t(k))$ is a $k\times k$ matrix, $\Gamma_{k,l}(0)$ is a $k\times l$ matrix, and $\Gamma_{l,k}(0)$ and $\Gamma^*_{l,k}(0)$ are $l\times k$ matrices. Observe that

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

Then, by the Woodbury matrix identity and partitioned matrix inversion formula, we have

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

Then, it follows that

$$|\mathrm{E}(Nf^*_{2,n,\infty}(k)f^*_{2,n,\infty}(l)) - \min(k,l)\sigma^2| \le C\big(\frac{\sum_{j=\sqrt{n}-K_n+1}^{\infty}|jb_j|^2}{\sqrt{n}-K_n} + \frac{\sqrt{n}K_n}{n-K_n}\big). \tag{S.11}$$

Therefore, (A.2) holds by (S.11), (2.3), and Assumption 4. This completes the proof. □

Proof of Lemma 4. Define

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

$$\mathbf{z}^\star_n(k) = \left(\sum_{j=0}^{\sqrt{n}/2-K_n} b_j\epsilon_{n-j}, ..., \sum_{j=0}^{\sqrt{n}/2-K_n} b_j\epsilon_{n-k+1-j}\right)', k\ge 1,$$

$$S^*_n(k-d) = \sum_{i=1}^{\sqrt{n}/2}(a_i - a_i(k-d))z^{**}_{n+1-i}, \quad z^{**}_{n+1-i} = \sum_{j=0}^{\sqrt{n}/2} b_j\epsilon_{n+1-i-j}.$$

Note that for all $1 \le u, v \le K_n - d$, $z^\star_n(u)$ is independent from $(S_n(v) - S^*_n(v))$ and $\sum_{j=K_n}^{n-\sqrt{n}-1} z_j(v)\epsilon_{j+1}$. Also, $\sum_{j=K_n}^{n-\sqrt{n}-1} z_j(u)\epsilon_{j+1}$ is independent from $(S^*_n(v), z^\star_n(v))$. Therefore, we have

$$\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f^\star_{2,n}(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k[S_n(k-d) - S^*_n(k-d)])\big] = 0,$$

and

$$\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f^\star_{2,n}(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k S^*_n(k-d))\big] = 0.$$

Then, it follows that

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

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

Therefore, we have

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d))(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))\big]}{L_n^d(\mathbf{w})}\right| \\
&\le \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}\big[(\sum_{k=\max\{1,d\}}^{K_n} w_k[f_{2,n}(k-d) - f_{2,n}^{\star}(k-d)])(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))\big]}{L_n^d(\mathbf{w})}\right| \\
&\le \sup_{\mathbf{w}\in\mathcal{H}_n^d}\frac{\mathrm{E}^{1/2}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_{2,n}(k-d) - f_{2,n}^{\star}(k-d))\big]^2}{\sqrt{L_n^d(\mathbf{w})}} \times \sup_{\mathbf{w}\in\mathcal{H}_n^d}\frac{\mathrm{E}^{1/2}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]^2}{\sqrt{L_n^d(\mathbf{w})}} \\
&\le C\Big\{\sum_{j=\sqrt{n}/2-K_n+1}^{\infty}|jb_j|^2 + \sqrt{\frac{K_n^3}{N}}\Big\}^{1/2}, \qquad \text{(S.12)}
\end{aligned}$$

where the last inequality is insured by Lemma 1 and with the same argument as Lemma 2 (i). Therefore, (A.3) holds by (S.12), (2.3), and Assumption 4. This completes the proof. □

Proof of Lemma 5. 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, it follows that

$$(I) \le (II) + (III) + (IV) + (V) + (VI) + (VII), \qquad \text{(S.13)}$$

where

$$\begin{aligned}
(II) &= \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}(f_{1,n}(d))^2 - \frac{d(d+1)\sigma^2}{N}}{L_n^d(\mathbf{w})}\right|, \\
(III) &= \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}(\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d))^2 - \sigma^2\sum_{\max\{1,d\}\le i,j\le K_n} w_i w_j(\min\{i,j\}-d)}{L_n^d(\mathbf{w})}\right|, \\
(IV) &= \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{\mathrm{E}(\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d))^2 - \sum_{\max\{1,d\}\le i,j\le K_n} w_i w_j\|a - a(\max\{i,j\}-d)\|_z^2}{L_n^d(\mathbf{w})}\right|, \\
(V) &= \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{2\mathrm{E}\big[f_{1,n}(d)\sum_{k=\max\{1,d\}}^{K_n} w_k f_{2,n}(k-d)\big]}{L_n^d(\mathbf{w})}\right|, \\
(VI) &= \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{2\mathrm{E}\big[f_{1,n}(d)\sum_{k=\max\{1,d\}}^{K_n} w_k S_n(k-d)\big]}{L_n^d(\mathbf{w})}\right|,
\end{aligned}$$

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

By Lemma 2 of Ing et al. (2010), we have $\lim_{n\to\infty}(II) = 0$. By Lemmas 2 and 3, we have $\lim_{n\to\infty}(III) = 0$. By Lemma 1, we have $\lim_{n\to\infty}(IV) = 0$. By Lemma 4, we have $\lim_{n\to\infty}(VII) = 0$. Following a similar argument to the proofs of (B.40) and (B.41) in Ing et al. (2010) and Hölder's inequality, we have

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

and

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

Then, by 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 have $\lim_{n\to\infty}(V) = 0$ and $\lim_{n\to\infty}(VI) = 0$. Therefore, (A.4) holds by (S.13) and the fact that $\lim_{n\to\infty}((II) + (III) + (IV) + (V) + (VI) + (VII)) = 0$. This completes the proof. □

Proof of Lemma 6. For any $\mathbf{w} \in \mathcal{H}_n^d$, $B_n(k-d) := B_{1n}(k,d) + B_{2n}(k-d)$, where $B_{1n}(k,d)$ and $B_{2n}(k-d)$ are defined after (A.1). Observe that

$$\begin{aligned}
&\mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_n(k) + S_n(k-d))\big]^2 \\
&= \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_n(k) - B_n(k-d))\big]^2 + \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(B_n(k-d) + S_n(k-d))\big]^2 \\
&+ \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(f_n(k) - B_n(k-d))\big]\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(B_n(k-d) + S_n(k-d))\big] \\
&= (I) + (II) + (III). \qquad \text{(S.14)}
\end{aligned}$$

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

$$\frac{(I)}{L_n^d(\mathbf{w})} \le \frac{\sum_{k=\max\{1,d\}}^{K_n} w_k \mathrm{E}\big[f_n(k) - B_n(k-d)\big]^2}{L_n^d(\mathbf{w})} \le \frac{C \sum_{k=\max\{1,d\}}^{K_n} w_k k^3}{N^2 L_n^d(\mathbf{w})} \le \frac{C}{N L_n^d(\mathbf{w})} \frac{K_n^3}{N}. \qquad \text{(S.15)}$$

For (II), we have

$$(II) = \mathrm{E}\big[\sum_{k=\max\{1,d\}}^{K_n} w_k(B_n(k-d) + S_n(k-d))\big]^2$$

$$
\begin{aligned}
&= \mathrm{E}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k(B_n(k-d) - F_n(k-d))\Big]^2 \\
&\quad + \mathrm{E}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k(B_n(k) - F_n(k-d))\Big]\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k(F_n(k-d) + S_n(k-d))\Big] \\
&\quad + \mathrm{E}\Big[\sum_{k=\max\{1,d\}}^{K_n} w_k(F_n(k,d) + S_n(k-d))\Big]^2 \\
&= (IV) + (V) + (VI). \qquad \text{(S.16)}
\end{aligned}
$$

Since

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

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

$$
\frac{(IV)}{L_n^d(\mathbf{w})} \le \frac{C}{NL_n^d(\mathbf{w})}. \qquad \text{(S.17)}
$$

Also, by the Cauchy-Schwarz inequality, with sufficiently large N, and Lemma 5, we have

$$
\frac{(V)}{L_n^d(\mathbf{w})} \le \frac{C}{(NL_n^d(\mathbf{w}))^{1/2}}. \qquad \text{(S.18)}
$$

Next, by the Cauchy-Schwarz inequality, the above decomposition of (II), and similar arguments for (V), we have

$$
\frac{(III)}{L_n^d(\mathbf{w})} \le \left(\frac{C}{NL_n^d(\mathbf{w})}\frac{K_n^3}{N}\right)^{1/2}. \qquad \text{(S.19)}
$$

Then, it follows that

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{E(f_n(k,d), S_n(k-d), \mathbf{w}) - E(F_n(k,d), S_n(k-d), \mathbf{w})}{L_n^d(\mathbf{w})}\right| \\
&= \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{E(f_n(k,d), S_n(k-d), \mathbf{w}) - (VI)}{L_n^d(\mathbf{w})}\right| \\
&\le \sup_{\mathbf{w}\in\mathcal{H}_n^d}\left|\frac{(I) + (II) + (III) + (IV) + (V)}{L_n^d(\mathbf{w})}\right|, \qquad \text{(S.20)}
\end{aligned}
$$

Therefore, (A.5) holds by (S.14)-(S.20). This completes the proof. □

Proof of Lemma 7. To simplify the notation, we define

$$
A := \sum_{1\le i,j\le d-1} w_i w_j \min(i,j), \quad A_d := \sum_{1\le i,j\le d-1} w_i w_j d, \qquad B := \sum_{d\le i,j\le K_n} w_i w_j \min(i,j),
$$

$$C := \sum_{1\le i,j\le d-1} w_i w_j \max(i,j), \quad D := \sum_{d\le i,j\le K_n} w_i w_j \max(i,j), \quad E := \sum_{1\le i,j\le d-1} w_i w_j \hat{\sigma}^2(\max(i,j)),$$
$$E_d := \sum_{1\le i,j\le d-1} w_i w_j \hat{\sigma}^2(d), \qquad F := \sum_{d\le i,j\le K_n} w_i w_j \hat{\sigma}^2(\max(i,j)),$$

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

For (i), by Lemma 4.1 and Eq. (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 $\check{\sigma}^2 = \hat{\sigma}^2(K_n)$. When $\hat{w}_{\text{MMA},k} > 0$, it means that there exists some $\mathbf{w} = (w_1, ..., w_k, ..., w_{K_n}) \in \mathcal{H}_n$ such that $\hat{\mathbf{w}}_{\text{MMA}} = \mathbf{w},\ w_k > 0,$ for any $1 \le k < d$. Then, it follows that

$$\begin{aligned}
&\Pr(\hat{w}_{\text{MMA},k} > 0, 1 \le k < d) \\
&= \Pr(w_k > 0, 1 \le k < d) \\
&= \Pr\big(N(E+F) + (A+B+C+D)\check{\sigma}^2 \le N(E_d+F) + (A_d+B+A_d+D)\check{\sigma}^2 \big) \\
&= \Pr\big(N[E-E_d] \le [(A_d-A)+(A_d-C)]\check{\sigma}^2 \big) \\
&\le \Pr\big(N[E-E_d] \le 2A_d\check{\sigma}^2 \big) \\
&\le \Pr\big(N[\sum_{1\le i,j\le d-1} w_i w_j(\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d))] \le 2\sum_{1\le i,j\le d-1} w_i w_j\, d\,\check{\sigma}^2 \big) \\
&\le \Pr\big(N[\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d)] \le 2d\,\check{\sigma}^2 \big) \\
&\le \Pr\big(N[\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d)] \le 2d\,(\sigma^2(K_n)+\epsilon) \big) + \Pr\big(|\check{\sigma}^2 - \sigma^2(K_n)| > \epsilon \big),
\end{aligned} \tag{S.21}$$

where the second and third inequalities hold by the fact that $\hat{\sigma}^2(k) \le \hat{\sigma}^2(l)$ for all $l < k$, $\sum_{k=1}^{K_n} w_k = 1$, and $\sum_{1\le i,j\le d-1} w_i w_j \hat{\sigma}^2(d-1) \le E$. Therefore, Lemma 7 (i) holds by (4.30) and Theorem 4.5 of Ing et al. (2012), and (S.21).

For (ii), when $\hat{w}_{\text{SMA},k} > 0$, it means that there exists some $\mathbf{w} = (w_1, ..., w_k, ..., w_{K_n}) \in \mathcal{H}_n$ such that $\hat{\mathbf{w}}_{\text{SMA}} = \mathbf{w},\ w_k > 0,$ for any $1 \le k < d$. Then, it follows that

$$\begin{aligned}
&\Pr(\hat{w}_{\text{SMA},k} > 0, 1 \le k < d) \\
&= \Pr(w_k > 0, 1 \le k < d) \\
&= \Pr\big([N+A+B+C+D] \times [E+F] \le [N+A_d+B+A_d+D] \times [E_d+F] \big) \\
&= \Pr\big(N[E-E_d] \le [(A_d-A)+(A_d-C)][E_d+F] + [A+B+C+D][E_d-E] \big) \\
&\le \Pr\big(N[E-E_d] \le [(A_d-A)+(A_d-C)][E_d+F] \big) \\
&\le \Pr\big(N[E-E_d] \le 2A_d\hat{\sigma}^2(d) \big) \\
&\le \Pr\big(N[\sum_{1\le i,j\le d-1} w_i w_j(\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d))] \le 2\sum_{1\le i,j\le d-1} w_i w_j\, d\,\hat{\sigma}^2(d) \big) \\
&\le \Pr\big(N[\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d)] \le 2d\,\hat{\sigma}^2(d) \big),
\end{aligned} \tag{S.22}$$

where the first to third inequalities hold by the fact that $\hat{\sigma}^2(k) \le \hat{\sigma}^2(l)$ for all $l < k$, $\sum_{k=1}^{K_n} w_k = 1$, and $\sum_{1\le i,j\le d-1} w_i w_j \hat{\sigma}^2(d-1) \le E$. Therefore, Lemma 7 (ii) holds by (4.30) and Theorem 4.5 of Ing et al. (2012), and (S.22).

For (iii), when $\hat{w}_{\text{AMA},k} > 0$, it means that there exists some $\mathbf{w} = (w_1, ..., w_k, ..., w_{K_n}) \in \mathcal{H}_n$ such that $\hat{\mathbf{w}}_{\text{AMA}} = \mathbf{w},\ w_k > 0,$ for any $1 \le k < d$. Then, it follows that

$$
\begin{aligned}
&\Pr(\hat{w}_{\text{AMA},k} > 0, 1 \le k < d) \\
&= \Pr(w_k > 0, 1 \le k < d) \\
&= \Pr\big(\ \log(E+F) + \frac{(A+B+C+D)}{N} \le \log(E_d+F) + \frac{(A_d+B+A_d+D)}{N}\ \big) \\
&= \Pr\big(\ \log(E+F) - \log(E_d+F) \le \frac{[(A_d-A)+(A_d-C)]}{N}\ \big) \\
&\le \Pr\big(\ \log(\frac{E+F}{E_d+F}) \le \frac{2d}{N}\ \big) \\
&= \Pr\big(\ \frac{E+F}{E_d+F} \le \exp(\frac{2d}{N})\ \big) \\
&= \Pr\big(\ \frac{E-E_d}{E_d+F} \le \exp(\frac{2d}{N}) - 1\ \big) \\
&\le \Pr\big(\ \frac{E-E_d}{E_d+F} \le \frac{2d}{N-2d}\ \big) \\
&= \Pr\big(\ (N-2d)(E-E_d) \le 2d(E_d+F)\ \big) \\
&\le \Pr\big(\ (N-2d)[\sum_{1\le i,j\le d-1} w_i w_j(\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d))] \le 2d\,\hat{\sigma}^2(d)\ \big) \\
&\le \Pr\big(\ (N-2d)[\hat{\sigma}^2(d-1) - \hat{\sigma}^2(d)] \le C\,\hat{\sigma}^2(d)\ \big),
\end{aligned}
\tag{S.23}
$$

where the second inequality holds by the fact that $\exp(x) - 1 \le \frac{x}{1-x}$ if $0 < x < 1$, and the last two inequalities hold by the fact that $\hat{\sigma}^2(k) \le \hat{\sigma}^2(l)$ for all $l < k$, $\sum_{1\le i,j\le d-1} w_i w_j > 0$, $E_d + F \le \hat{\sigma}^2(d)$, and $\sum_{1\le i,j\le d-1} w_i w_j \hat{\sigma}^2(d-1) \le E$. Therefore, Lemma 7 (iii) holds by (4.30) and Theorem 4.5 of Ing et al. (2012), and (S.23). This completes the proof. □

Proof of Lemma 8. Note that $\hat{\mathbf{w}}^d_{\tilde{S}_n} = \arg\min_{\mathbf{w}\in\mathcal{H}^d_n} \tilde{S}_n(\mathbf{w})$, $L^d_n(\mathbf{w}^*_n) = \inf_{\mathbf{w}\in\mathcal{H}^d_n} L^d_n(\mathbf{w})$, and $g(\cdot)$ is an increasing function. Then, it follows that

$$
\begin{aligned}
0 \ge g(\tilde{S}_n(\hat{\mathbf{w}}^d_{\tilde{S}_n})) - g(\tilde{S}_n(\mathbf{w}^*_n)) &= C_n(\hat{\mathbf{w}}^d_{\tilde{S}_n}) - G_n(\hat{\mathbf{w}}^d_{\tilde{S}_n}) - (C_n(\mathbf{w}^*_n) - G_n(\mathbf{w}^*_n)) \\
&= NL^d_n(\hat{\mathbf{w}}^d_{\tilde{S}_n}) - NL^d_n(\mathbf{w}^*_n) - V_n(\hat{\mathbf{w}}^d_{\tilde{S}_n}, \mathbf{w}^*_n) - (G_n(\hat{\mathbf{w}}^d_{\tilde{S}_n}) - G_n(\mathbf{w}^*_n))), \\
(G_n(\hat{\mathbf{w}}^d_{\tilde{S}_n}) - G_n(\mathbf{w}^*_n)) + V_n(\hat{\mathbf{w}}^d_{\tilde{S}_n}, \mathbf{w}^*_n) &\ge NL^d_n(\hat{\mathbf{w}}^d_{\tilde{S}_n}) - NL_n(\mathbf{w}^*_n) \ge 0,
\end{aligned}
$$

and

$$
\sup_{\mathbf{w}\in\mathcal{H}^d_n} \Big|\frac{G_n(\mathbf{w}) - G_n(\mathbf{w}^*_n))}{NL^d_n(\mathbf{w})}\Big| + \sup_{\mathbf{w}\in\mathcal{H}^d_n} \Big|\frac{V_n(\mathbf{w}, \mathbf{w}^*_n)}{NL^d_n(\mathbf{w})}\Big| \ge \frac{V_n(\hat{\mathbf{w}}^d_{\tilde{S}_n}, \mathbf{w}^*_n)}{NL^d_n(\hat{\mathbf{w}}^d_{\tilde{S}_n})} \ge 1 - \frac{L^d_n(\mathbf{w}^*_n)}{L^d_n(\hat{\mathbf{w}}^d_{\tilde{S}_n})} \ge 0,
$$

Therefore, by (A.6) and (B.6), we have $\lim_{n\to\infty} L_n^d(\mathbf{w}_n^*)/L_n^d(\hat{\mathbf{w}}_{\tilde{S}_n}^d) \xrightarrow{p} 1$. This completes the proof. $\square$

S.2 Additional Simulation Results

Figures S.1-4 present the relative MSPEs of the various estimates for $d = 0$ and $d = 2$ in both algebraic-decay and exponential-decay cases. The results show that the MMA, AMA, and SMA have similar MSPEs and perform quite well in both cases. Overall, the ranking of different estimators in $d = 0$ and $d = 2$ is quite similar to that in $d = 1$. Figures S.5-8 examine the effect of the sample size on the MSPE in both algebraic-decay and exponential-decay cases. Like the results in $d = 1$, the AMA, MMA, and SMA have much lower MSPEs than those of the AIC, Cp, and SIC in the algebraic-decay case, but the MSPEs of AIC, Cp, and SIC are approaching those of AMA, MMA, and SMA in the exponential-decay case.

References

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

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

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

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.

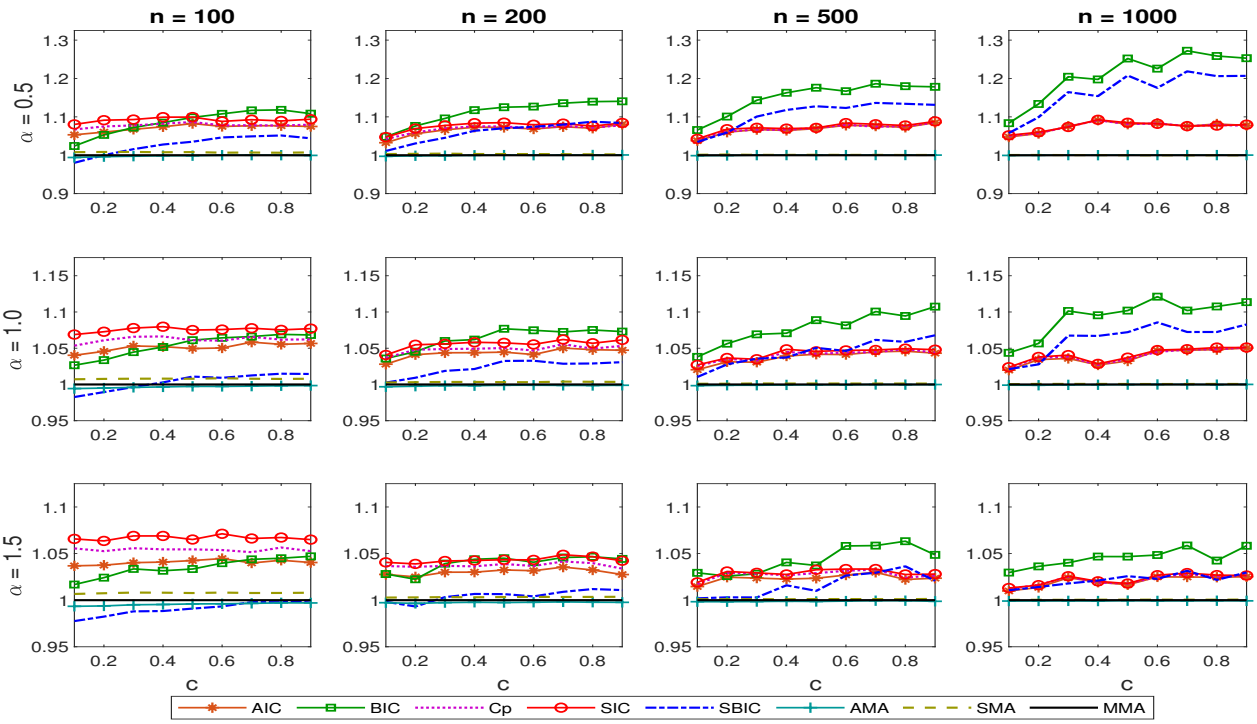

Figure S.1: Relative MSPEs for $d = 0$ in the algebraic-decay case

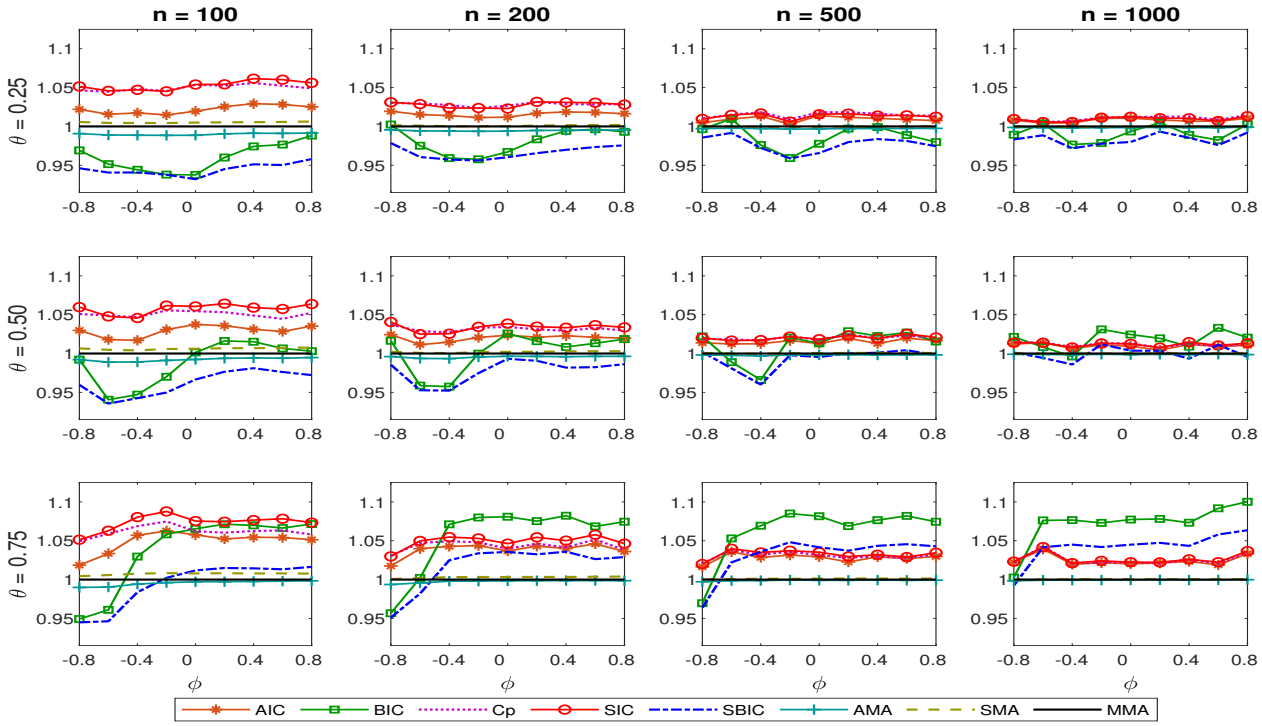

Figure S.2: Relative MSPEs for $d = 0$ in the exponential-decay case

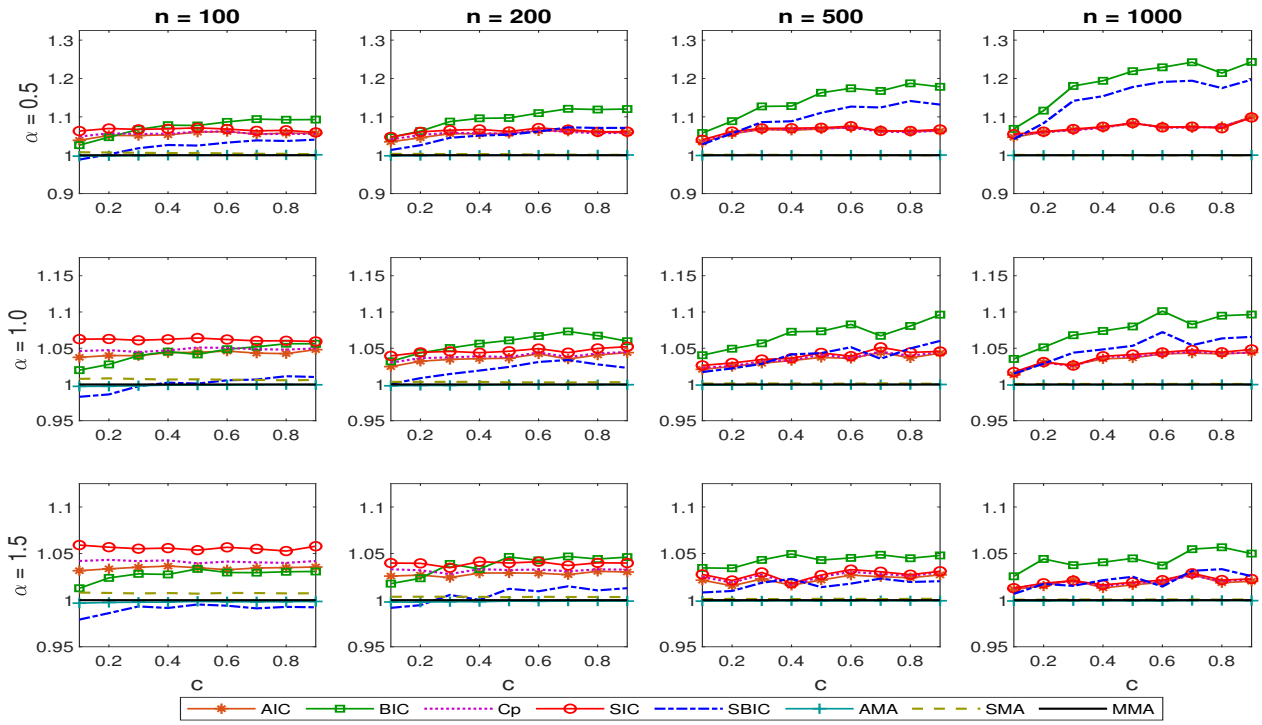

Figure S.3: Relative MSPEs for $d = 2$ in the algebraic-decay case

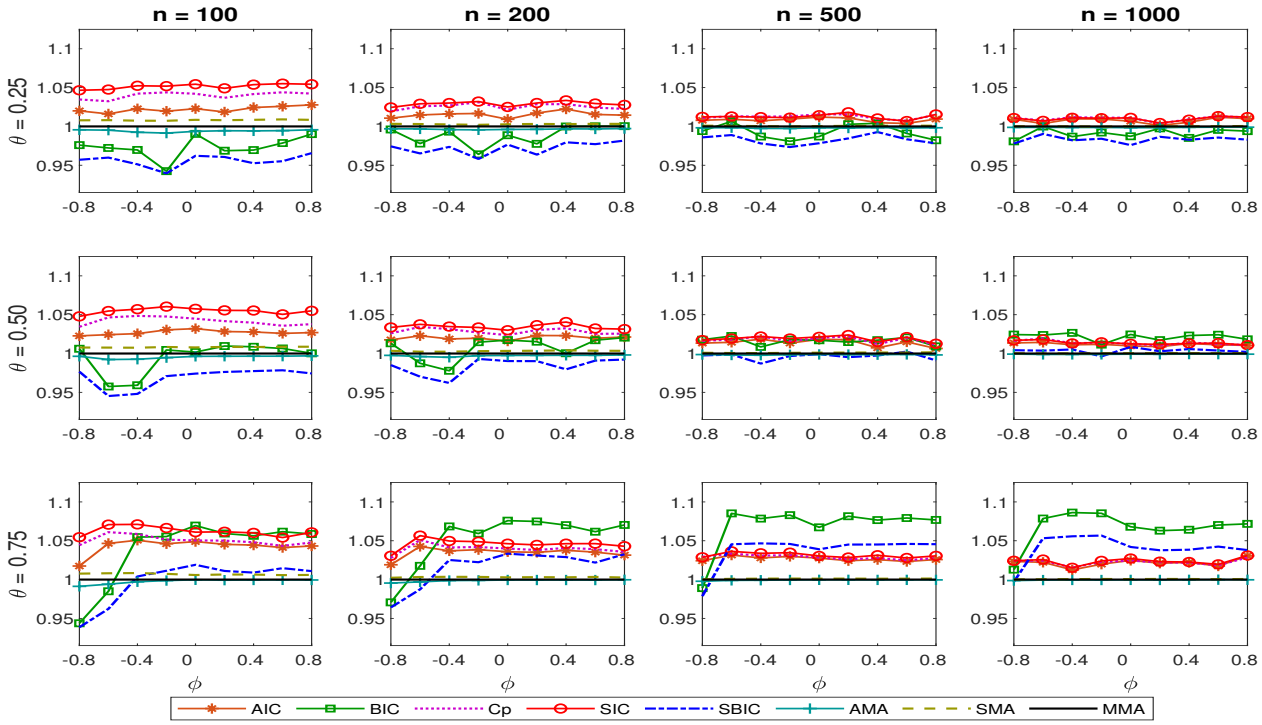

Figure S.4: Relative MSPEs for $d = 2$ in the exponential-decay case

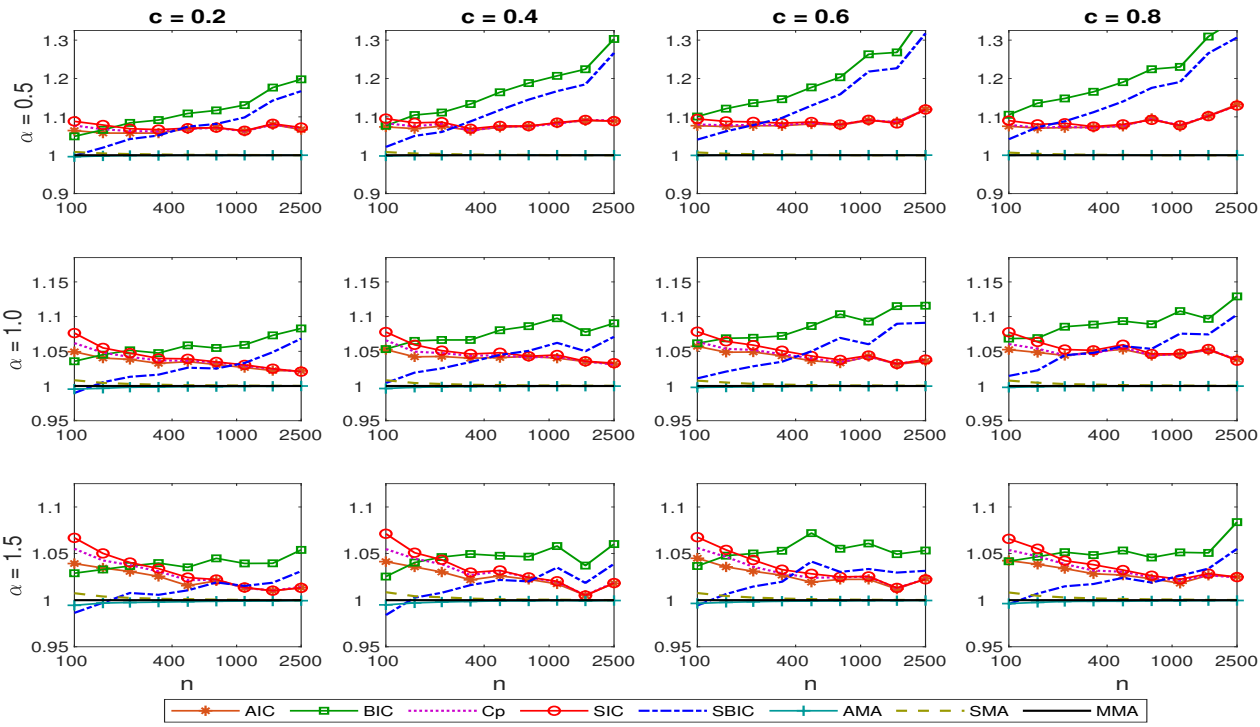

Figure S.5: Relative MSPEs for the algebraic-decay case, $d = 0$, various sample sizes

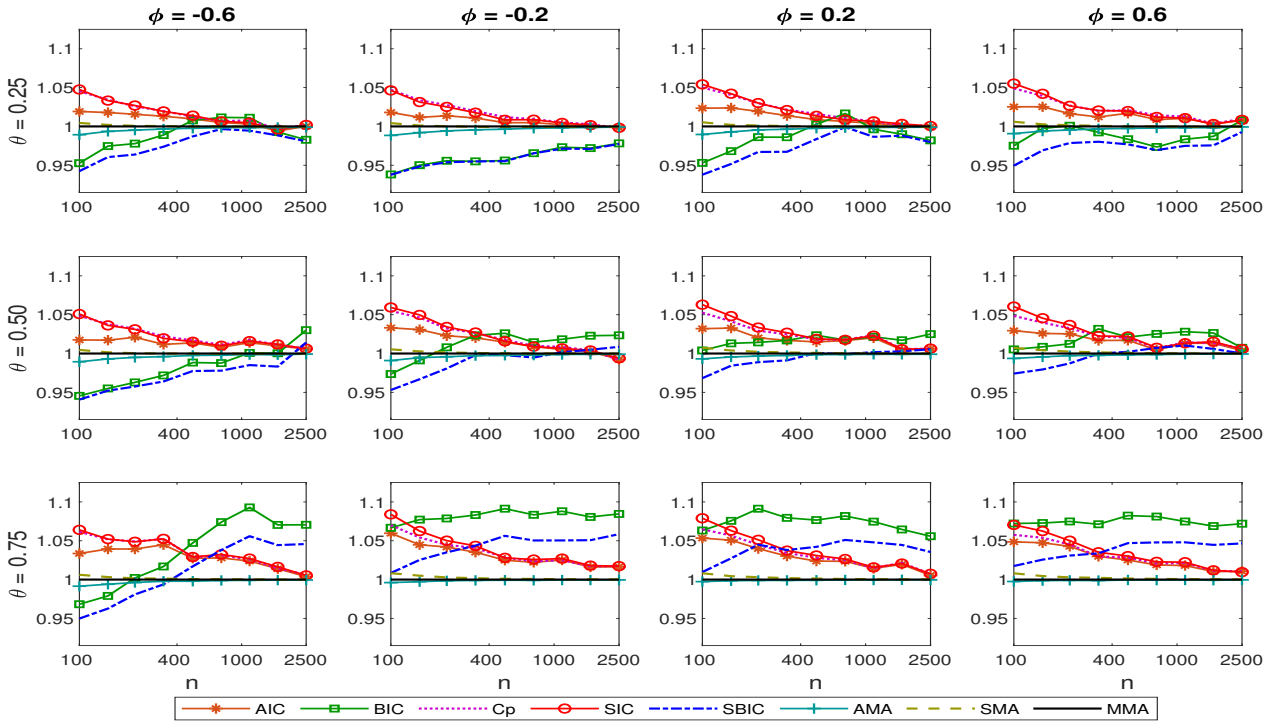

Figure S.6: Relative MSPEs for the exponential-decay case, $d = 0$, various sample sizes

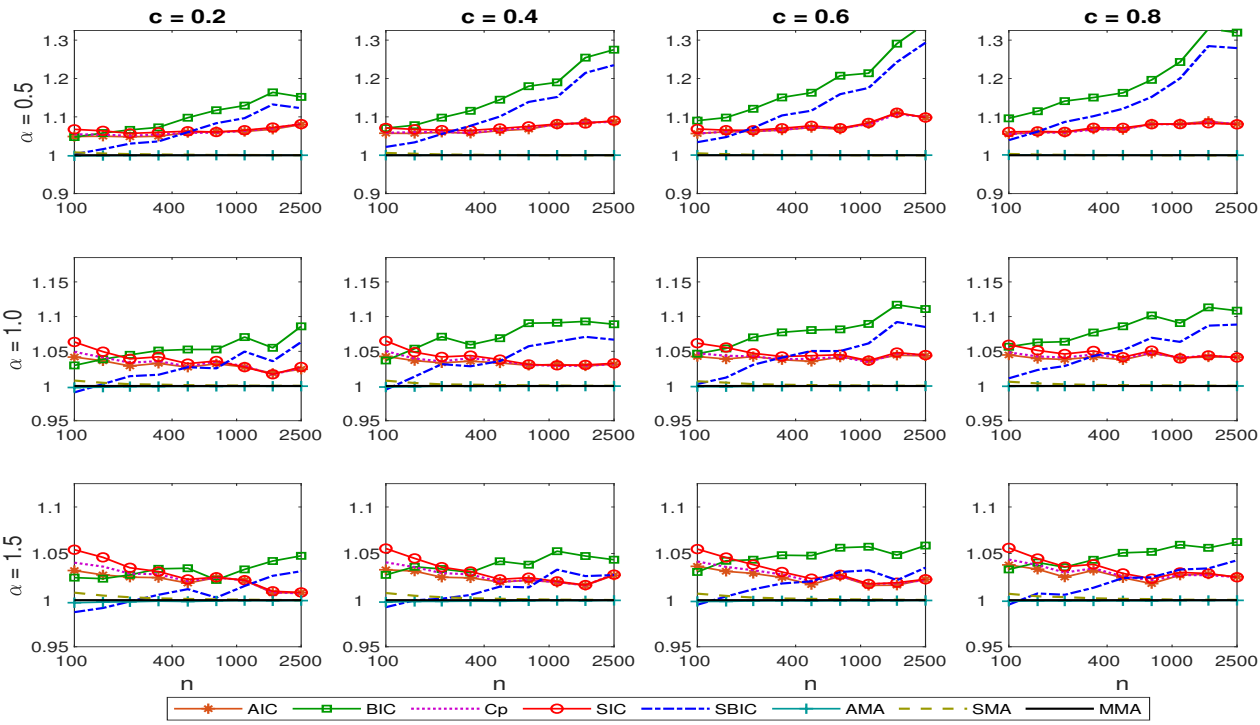

Figure S.7: Relative MSPEs for the algebraic-decay case, $d = 2$, various sample sizes

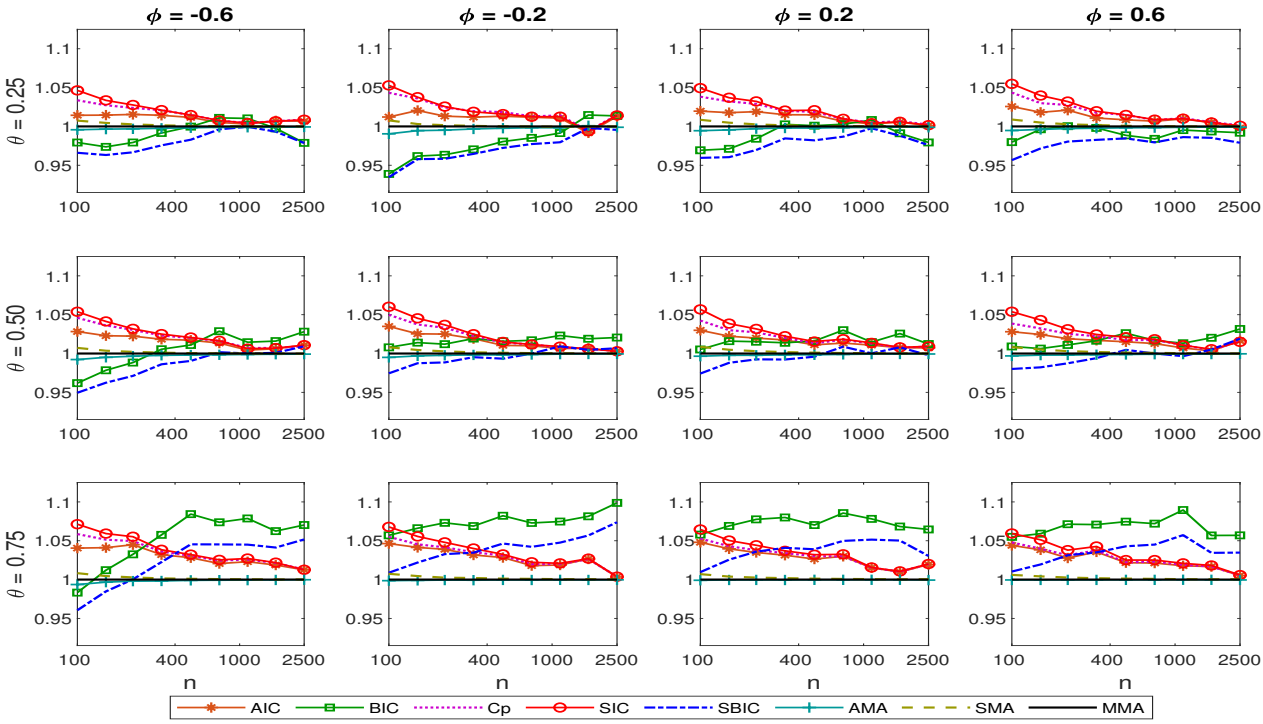

Figure S.8: Relative MSPEs for the exponential-decay case, $d = 2$, various sample sizes