Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

ON MODEL SELECTION FROM A FINITE FAMILY OF POSSIBLY MISSPECIFIED TIME SERIES MODELS

By Hsiang-Ling Hsu[*,§], Ching-Kang Ing[†,¶] and Howell Tong[‡,‖]

National University of Kaohsiung, Taiwan[§], National Tsing Hua University, Taiwan [¶], University of Electronic Science & Technology, China and London School of Economics, United Kingdom[‖]

Consider finite parametric time series models. 'I have n observations and k models, which model should I choose on the basis of the data alone?' is a frequently asked question in many practical situations. This poses the key problem of selecting a model from a collection of candidate models, none of which is necessarily the true data generating process (DGP). Although existing literature on model selection is vast, there is a serious lacuna in that the above problem does not seem to have received much attention. In fact, existing model selection criteria have avoided addressing the above problem directly, either by assuming that the true DGP is included among the candidate models and aiming at choosing this DGP, or by assuming that the true DGP can be asymptotically approximated by an increasing sequence of candidate models and aiming at choosing the candidate having the best predictive capability in some asymptotic sense. In this article, we propose a misspecification-resistant information criterion (MRIC) to address the key problem directly. We first prove the asymptotic efficiency of MRIC whether the true DGP is among the candidates or not, within the fixed-dimensional framework. We then extend this result to the high-dimensional case in which the number of candidate variables is much larger than the sample size. In particular, we show that MRIC can be used in conjunction with a high-dimensional model selection method to select the (asymptotically) best predictive model across several high-dimensional misspecified time series models.

1. Introduction. Let us consider finite parametric time series models. In the vast literature of model selection, problems tend to be classified into

[*]Research was supported in part by the Ministry of Science and Technology of Taiwan under grants MOST 103-2118-M-390-004-MY2.

[†]Research was supported in part by the Science Vanguard Research Program of the Ministry of Science and Technology, Taiwan

[‡]Research was supported in part by a special research grant, University of Electronic Science and Technology, Chengdu, China

AMS 2000 subject classifications: Primary 63M30; secondary 62F07, 62F12

Keywords and phrases: AIC, BIC, Misspecification-resistant information criterion, Multi-step prediction error, High-dimensional misspecified models, Orthogonal greedy algorithm

two categories according to whether the true data generating process (DGP) is included among the collection of candidate models. The first category (referred to as category I) assumes that the true DGP belongs to a stipulated collection of candidate models, and the objective of model selection is simply choosing the true DGP. A model selection criterion is said to be consistent if it can choose the (most parsimonious) true DGP with probability tending to 1. In time series models as well as in linear regression, Bayesian information criterion (BIC) (Schwarz, 1978) has been shown to have this property; see, e.g., Nishii (1984), Rao and Wu (1989) and Wei (1992). On the other hand, Akaike's information criterion (AIC) (Akaike, 1974) and Mallows' C_p (Mallows, 1973), which tend to choose overfitting models, are not consistent in category I (e.g., Shibata, 1976 and Shao, 1997). The second category (category II) assumes that the true DGP is not one of the candidate models. In this category, choosing the model having the best predictive capabilities becomes the objective. When the true DGP is a linear regression model with infinitely many parameters and the number of predictor (explanatory) variables in the candidate models increases to infinity with the sample size such that the corresponding approximation error vanishes asymptotically, Shibata (1981) and Li (1987) showed that AIC and Mallows' C_p possess asymptotic efficiency, in the sense that these criteria can choose the model whose finite-sample mean squared prediction error (MSPE) is asymptotically equivalent to the smallest one among those of the candidate models. In contrast, BIC fails to achieve asymptotic efficiency under category II; see Shibata (1980), Shao (1997) and Ing and Wei (2005). For a survey of the performance of various model selection criteria in both categories, see Shao (1997).

It is usually difficult for practitioners to perceive which category applies. Since, as mentioned in the previous paragraph, most existing criteria cannot *simultaneously* enjoy consistency in category I and asymptotic efficiency in category II, the choice of selection criteria has become a key point of contention over the past decades. For example, Ing (2007) and Yang (2007) have proposed similar adaptive procedures. They first compare two models selected by BIC, one for *partial* data points and another for *full* data points. They adopt AIC if the two selected models are different suggesting the plausibility of category II, and BIC otherwise. By suitably deciding the number of partial data points in the first step, they have shown that the proposed two-step procedure possesses consistency and asymptotic efficiency in categories I and II, respectively. More recently, Liu and Yang (2011) devised the so called "parametricness index" to determine between categories I and II, and Zhang and Yang (2015) proposed using cross-validation to select

between AIC and BIC in the absence of prior information on the underlying category. For a related result on solving the AIC-BIC dilemma from the point of view of cumulative risk, see van Erven et al. (2012).

While these recent efforts to resolve the controversy between AIC and BIC are novel, they mainly contribute to the increasing-dimensional (ID) framework, which requires that the number of candidate variables to grow to infinity with the sample size, n. However, in many realistic situations, we are often faced with the problem of selecting a model from a *finite and fixed* collection of candidate models, none of which is necessarily the true DGP. It was, in fact, this problem that Akaike was originally trying to solve. He said (1978, p.217), '... at some stage, we have at hand several models which are the candidates for our final choice.' Although existing literature on model selection is vast, the above problem does not seem to have received much attention. This motivates us to ask whether there exists a model selection procedure that can perform well in both categories and within the fixed-dimensional (FD) framework in which the number of candidate models does not change with n, thus filling a serious lacuna in the vast literature on model selection.

In this article, we propose a misspecification-resistant information criterion (MRIC). Specifically, we prove that MRIC, within the FD framework, possesses asymptotic efficiency in the sense of (3.6) whether the true DGP belongs to the candidate models or not. The MRIC has additional advantages. First, it is applicable to h-step prediction of time series data with $h \geq 1$. In particular, by changing the prediction lead times in the MRIC formula, the asymptotic efficiency of MRIC is guaranteed for each $h \geq 1$. Second, unlike the resolutions proposed for the ID case (e.g., Ing (2007), Yang (2007) and Zhang and Yang (2015)), MRIC can achieve asymptotic efficiency on its own without the help of additional/auxiliary criteria. Indeed, there are already several 'single-step' model selection procedures proposed to combat model misspecification, e.g., TIC (Takeuchi , 1976), GIC (Konishi and Kitagawa, 1996) and GBIC and GBIC$_p$ (Lv and Liu, 2014). However, it seems decidedly difficult to justify their asymptotic efficiency within the FD framework; see Section section S5 of the supplementary material for this paper (Hsu et al., 2018). We summarize the performance of major model selection procedures discussed above in the form of the two tables; Table 1 is for the ID framework and Table 2 for the FD framework.

When several high-dimensional and (possibly) misspecified time series models are entertained, MRIC can also be used in conjunction with high-dimensional model selection methods to choose good predictive models. Note that high-dimensional model selection problems have been extensively inves-

TABLE 1
Increasing-dimensional case (# of candidates increases with n)

Criteria	Case I: The true model is included as a candidate. Goal: Consistency	Case II: The true model is NOT included as a candidate. Goal: Asymp. efficiency for prediction.	Case III: No info. on whether the true model is included. Goal: Consistency when the true model is included + asymptotic efficiency when the true model is not included.
AIC	No	Yes	No
BIC	Yes	No	No
GAIC	No	Yes	No
GBIC	Yes	No	No
Two-stage IC	Yes	Yes	Yes

TABLE 2
Fixed-dimensional case (# of candidates is fixed with n)

Criteria	Case I: Consistency	Case II: Asymp. efficiency	Case III: Consistency + Asymp. efficiency
AIC	No	No	No
BIC	Yes	No	No
GAIC	No	No	No
GBIC	Yes	No	No
GBIC_p	Yes	No	No
MRIC	Yes	Yes	Yes

tigated over the past decade. However, most studies are devoted to the case where observations are independent over time. Recent papers of Basu and Michailidis (2015) and Wu and Wu (2016) are among the few dedicated to high-dimensional time series models. Although some desirable asymptotic properties of the Lasso estimates (Tibshirani , 1996) have been established by these authors under correct model specification, the question of how to choose the best predictive model across several different high-dimensional misspecified time series models still remains untouched. To fill this gap, we start by introducing a three-step model selection procedure, OGA+HDIC_h+Trim (Section 4.1), and apply the procedure to each high-dimensional model. We then suggest choosing the model that achieves the lowest MRIC value among those decided by OGA+HDIC_h+Trim. This approach is shown to have forecast optimality in the sense of (4.15).

The rest of the paper is organized as follows. In Section 2.1, we provide an asymptotic expression for the finite-sample MSPE of the least squares predictor, which is valid regardless of whether the model is correctly or incorrectly specified. In Section 2.2, we list the technical conditions needed in Section 2.1 and discuss their suitability. Based on a consistent estimate of the expression obtained in Section 2.1, we propose our MRIC and prove its asymptotic efficiency within the FD framework in Section 3. Applications of MRIC to misspecified ARX models are also given in the same section. In Section 4, the results in Sections 2 and 3 are extended to high-dimensional models. We show that MRIC can be used together with OGA+HDIC_h+Trim

to achieve asymptotic efficiency in the sense of (4.15) when several high-dimensional and (possibly) misspecified models are simultaneously taken into account. We conclude in Section 5. A detailed discussion of model misspecification is provided in Appendix A. All proofs and an extension of MRIC to a class of nonlinear models are relegated to Hsu et al. (2018). The finite-sample performance of the proposed methods in both low- and high-dimensional cases is also illustrated via simulated and real data in Hsu et al. (2018).

2. Mean Squared Prediction Error under Possible Misspecification.

2.1. *An Asymptotic Expression.* Let $\{y_t\}$ and $\{\mathbf{x}_t\} = \{(x_{t,1}, \ldots, x_{t,m})^\top\}$, $m \geq 1$, be weakly stationary processes on the probability space $(\Omega, \mathcal{F}, P)$. Given observations up to n, we are interested in forecasting $y_{n+h}, h \geq 1$, based on the following model,

$$y_{t+h} = \alpha_h + \boldsymbol{\beta}_h^\top \mathbf{x}_t + \varepsilon_{t,h}, \tag{2.1}$$

where $\boldsymbol{\beta}_h = (\beta_{1,h}, \ldots, \beta_{m,h})^\top = \arg\min_{\mathbf{c} \in R^m} \mathrm{E}\{y_{t+h} - \mathrm{E}(y_{t+h}) - \mathbf{c}^\top[\mathbf{x}_t - \mathrm{E}(\mathbf{x}_t)]\}^2$ and $\alpha_h = \mathrm{E}(y_{t+h}) - \boldsymbol{\beta}_h^\top \mathrm{E}(\mathbf{x}_t)$. Note that we allow that (i) $h \geq 1$, (ii) $\mathbf{x}_t$ contains both endogenous and exogenous variables, and (iii) $\varepsilon_{t,h}$ are serially correlated and correlated with $\mathbf{x}_k$ for $k \neq t$. Thus, model (2.1) actually represents very general situations beyond the special cases of multistep prediction in (possibly) misspecified AR models. In addition, we allow $\mathbf{x}_t$ to vary with h, but suppress its dependence on h in order to simplify the notation.

To gain further insights into the effect of model misspecification on the correlations between $\{\mathbf{x_t}\}$ and $\varepsilon_{t,h}$, we assume that the data are generated according to the following DGP,

$$y_{t+1} = a w_t + \varepsilon_{t+1}, \tag{2.2}$$

in which $a \neq 0$, $\{\varepsilon_t\}$ is a sequence of independent and identically distributed (i.i.d.) random errors obeying $\mathrm{E}(\varepsilon_1) = 0$ and $\mathrm{E}(\varepsilon_1^2) > 0$, and $w_t = \theta_1 w_{t-1} + \theta_2 w_{t-2} + \delta_t$ is a stationary AR(2) process, with $\theta_1 \theta_2 \neq 0$ and $\{\delta_t\}$ being a sequence of zero-mean i.i.d. random errors independent of $\{\varepsilon_t\}$. We also let

$$\mathrm{E}(\delta_1^2) = 1 - \theta_2^2 - \{\theta_1^2(1+\theta_2)/(1-\theta_2)\},$$

yielding $\gamma_w(0) = 1$, where $\gamma_w(j) = \mathrm{E}(w_t w_{t+j})$. If one is interested in predicting y_{n+2}, then, in view of (2.2), a correctly specified model for two-step

prediction is

$$y_{t+2} = a\theta_1 w_t + a\theta_2 w_{t-1} + \varepsilon_{t,2}^{(0)},$$

where $\varepsilon_{t,2}^{(0)} = \varepsilon_{t+2} + a\delta_{t+1}$. It is easy to see that $\mathrm{E}(\varepsilon_{t,2}^{(0)} w_{t-j}) = 0$ for $j \geq 0$. On the other hand, if a misspecified two-step prediction model,

$$y_{t+2} = \beta w_t + \varepsilon_{t,2},$$

is used, where $\beta = \mathrm{E}(y_{t+2}w_t) = a\theta_1 + a\theta_2\theta_1/(1-\theta_2)$ and $\varepsilon_{t,2} = \varepsilon_{t,2}^{(0)} - a\theta_2[\{\theta_1/(1-\theta_2)\}w_t - w_{t-1}]$, then $\mathrm{E}(\varepsilon_{t,2}w_{t-j}) = [-a\theta_2/(1-\theta_2)](\gamma_w(j+1) - \gamma_w(j-1)) \neq 0$ for $j \geq 1$ although $\mathrm{E}(\varepsilon_{t,2}w_t) = 0$ still follows. For a more detailed discussion on model misspecification, see Appendix A.

Model (2.1) can be rewritten as $y_{t+h} - \mathrm{E}(y_{t+h}) = \boldsymbol{\beta}_h^\top(\mathbf{x}_t - \mathrm{E}(\mathbf{x}_t)) + \varepsilon_{t,h}$. Having observed $y_1, \ldots, y_n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we may replace $\mathrm{E}(y_{t+h})$ by $\bar{y}$ and $\mathrm{E}(\mathbf{x}_t)$ by $\bar{\mathbf{x}}$, where $\bar{y} = n^{-1}\sum_{t=1}^n y_t$ and $\bar{\mathbf{x}} = n^{-1}\sum_{t=1}^n \mathbf{x}_t$. Although $y_{t+h} - \bar{y}$ and $\mathbf{x}_t - \bar{\mathbf{x}}$ constitute triangular arrays, the difference between $y_{t+h} - \mathrm{E}(y_{t+h})$ and $y_{t+h} - \bar{y}$ and that between $\mathbf{x}_t - \mathrm{E}(\mathbf{x}_t)$ and $\mathbf{x}_t - \bar{\mathbf{x}}$ vanish asymptotically. In order to simplify the exposition, we assume throughout the paper that $\mathrm{E}(y_t) = 0$ and $\mathrm{E}(\mathbf{x}_t) = \mathbf{0}$, and hence (2.1) becomes

$$y_{t+h} = \boldsymbol{\beta}_h^\top \mathbf{x}_t + \varepsilon_{t,h}. \tag{2.3}$$

Using the least squares estimator (LSE),

$$\hat{\boldsymbol{\beta}}_n(h) = \left(\sum_{t=1}^N \mathbf{x}_t\mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^N \mathbf{x}_t y_{t+h} = \widehat{\mathbf{R}}^{-1}\frac{1}{N}\sum_{t=1}^N \mathbf{x}_t y_{t+h},$$

of $\boldsymbol{\beta}_h$, one can predict y_{n+h} by

$$\hat{y}_{n+h} = \hat{\boldsymbol{\beta}}_n^\top(h)\mathbf{x}_n,$$

where $N = n - h$ and $\widehat{\mathbf{R}} = N^{-1}\sum_{t=1}^N \mathbf{x}_t\mathbf{x}_t^\top$.

In the next theorem, we provide an asymptotic expression for the finite-sample mean squared prediction error (MSPE) of $\hat{y}_{n+h}$, $\mathrm{E}(y_{n+h} - \hat{y}_{n+h})^2$, which is referred to as the MSPE in the sequel. One special feature of our expression is that it holds in both correctly and misspecified cases, thereby offering insight into pursuing asymptotically efficient model selection without knowing the category to which the underlying problem belongs.

THEOREM 2.1. *Assume* (2.3) *and conditions* (C1)–(C6) *in Section 2.2. Then, for any* $h \geq 1$,

$$\mathrm{E}(y_{n+h} - \hat{y}_{n+h})^2 = \mathrm{E}(\varepsilon_{n,h}^2) + n^{-1}(L_h + o(1)), \tag{2.4}$$

where $L_h = \mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{C}_{h,0}\right) + 2\sum_{s=1}^{h-1}\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{C}_{h,s}\right)$, *with* $\mathbf{R} = \mathrm{E}(\mathbf{x}_1\mathbf{x}_1^\top)$ *being nonsingular and* $\mathbf{C}_{h,s} = \mathrm{E}(\mathbf{x}_1\mathbf{x}_{1+s}^\top\varepsilon_{1,h}\varepsilon_{1+s,h})$.

The first term on the right-hand side of (2.4), referred to as the population MSPE, can be viewed as a measure of the goodness fit of model (2.3), whereas the second term on the right-hand side of (2.4) is related to the estimation error of $\hat{\boldsymbol{\beta}}_n(h)$. To appreciate the novelty of Theorem 2.1, assume that y_t is a stationary AR(m) model,

$$y_{t+1} = \sum_{i=1}^{m} a_i y_{t+1-i} + \epsilon_{t+1}, \tag{2.5}$$

where $1 - a_1 z - \cdots - a_l z^m \neq 0$ for all $|z| \leq 1$ and ϵ_t are independent random disturbances with $\mathrm{E}(\epsilon_t) = 0$ and $\mathrm{E}(\epsilon_t^2) = \sigma^2 > 0$ for all t. In view of (2.5), a correctly specified model for the h-step, $h \geq 1$, prediction is given by

$$y_{t+h} = \boldsymbol{\beta}_h^\top \mathbf{x}_t + \varepsilon_{t,h}, \tag{2.6}$$

where $\mathbf{x}_t = (y_t, \ldots, y_{t-m+1})^\top$, $\varepsilon_{t,h} = \sum_{j=0}^{h-1} b_j \epsilon_{t+h-j}$, with b_j satisfying $(1 - a_1 z - \cdots - a_m z^m)\sum_{j=0}^{\infty} b_j z^j = 1$, and $\boldsymbol{\beta}_h = \mathbf{A}^{h-1}(m)\mathbf{a}$ with $\mathbf{a} = (a_1, \ldots, a_m)^\top$ and

$$\mathbf{A}(m) = \left(\begin{array}{c|c} \mathbf{a} & \begin{array}{c} \mathbf{I}_{m-1} \\ \hline \mathbf{0}_{m-1}^\top \end{array} \end{array}\right),$$

noting that $\mathbf{I}_k$ and $\mathbf{0}_k$, respectively, denote the k-dimensional identity matrix and the k-dimensional vector of zeros. Under suitable conditions on ϵ_t (see Section 2.2), it can be shown that (C1)–(C6) hold, and hence by Theorem 2.1 and some algebraic manipulations,

$$\lim_{n\to\infty} n\{\mathrm{E}\left(y_{n+h} - \hat{y}_{n+h}\right)^2 - \mathrm{E}\left(\varepsilon_{n,h}^2\right)\} = L_h = \mathrm{tr}\left(\mathbf{R}^{-1}\mathrm{cov}\left(\sum_{j=0}^{h-1} b_j \mathbf{x}_{1+j}\right)\right)\sigma^2, \tag{2.7}$$

which is the key conclusion of Theorem 2 of Ing (2003). It is, however, important to note that when the model is misspecified, $\varepsilon_{t,h}$ and $\{\mathbf{x}_k, k < t\}$ are generally correlated, and hence the normalized MSPE,

$$\begin{aligned} &N\{\mathrm{E}\left(y_{n+h} - \hat{y}_{n+h}\right)^2 - \mathrm{E}(\varepsilon_{n,h}^2)\} \\ &= -2\mathrm{E}\{\varepsilon_{n,h}\mathbf{x}_n^\top \widehat{\mathbf{R}}^{-1}\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\} + \mathrm{E}(\mathbf{x}_n^\top \widehat{\mathbf{R}}^{-1} N^{-1/2}\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h})^2, \end{aligned} \tag{2.8}$$

may have a nonnegligible "cross-product" term, $-2\mathrm{E}\{\varepsilon_{n,h}\mathbf{x}_n^\top \widehat{\mathbf{R}}^{-1}\sum_{t=1}^N \mathbf{x}_t\varepsilon_{t,h}\}$, which vanishes in the correctly specified case due to the independence between $\varepsilon_{t,h}$ and $\{\mathbf{x}_k, k \le t\}$. In fact, it is shown in Ing (2003) that the rightmost term of (2.7) is solely attributed to the second term on the right-hand side of (2.8). At first sight, it would seem unrealistic to expect that L_h is still valid under model misspecification, without any correction or adjustment. To our amazement, we are able to reveal L_h's generality for both correct and misspecified cases after discovering some unexpected cancelation between some components in the first and the second terms on the right-hand side of (2.8); see (S1.4) and (S1.5) in Section section S1 of Hsu et al. (2018).

Before closing this section, we remark that in the case of *independent observations*, a term similar to $L_1 = \mathrm{tr}(\mathbf{R}^{-1}\mathrm{E}(\mathbf{x}_1\mathbf{x}_1^\top\varepsilon_{1,1}^2))$ has been used by Takeuchi (1976) as a bias correction for the log-likelihood in order to obtain an asymptotically unbiased estimate of the Kullback-Leibler divergence between the true model and a misspecified working model. For related discussion, see Stone (1977), Konishi and Kitagawa (1996), Burnham and Anderson (2002), Bozdogan (2000) and Lv and Liu (2014). All these authors, however, focus on independent observations, and hence time series data are regrettably precluded. Although Wei (1992) allowed for dependence among the data and showed that L_1 is the constant associated with the $\log n$ term in an asymptotic expression for the accumulated prediction error (APE) of the least squares predictor, his approach, focusing exclusively on the APE and the one-step prediction, is applicable to neither the MSPE nor the multistep prediction.

2.2. *Conditions (C1)–(C6).* In order to facilitate exposition, we impose the following regularity conditions.

(C1) There exist $q_1 > 5$ and $0 < C_1 < \infty$ such that for any $1 \le n_1 < n_2 \le n$ and any $1 \le i, j \le m$,

$$\mathrm{E}\left|(n_2 - n_1 + 1)^{-1/2}\sum_{t=n_1}^{n_2} x_{t,i}x_{t,j} - \mathrm{E}\,(x_{t,i}x_{t,j})\right|^{q_1} \le C_1. \tag{2.9}$$

(C2) $\mathbf{C}_{h,s} = \mathrm{E}(\mathbf{x}_t\mathbf{x}_{t+s}^\top\varepsilon_{t,h}\varepsilon_{t+s,h})$ is independent of t, and for any $1 \le i, j \le m$,

$$\mathrm{E}\,(x_{1,i}x_{n,j}\varepsilon_{1,h}\varepsilon_{n,h}) = o(n^{-1}). \tag{2.10}$$

(C3) $\sup_{-\infty<t<\infty}\mathrm{E}\|\mathbf{x}_t\|^{10} < \infty$ and $\sup_{-\infty<t<\infty}\mathrm{E}|\varepsilon_{t,h}|^6 < \infty$, where for vector $\mathbf{f} = (f_1, \cdots, f_m)^\top$, $\|\mathbf{f}\|^2 = \sum_{t=1}^m f_t^2$.

(C4) There exists $0 < C_2 < \infty$ such that for $1 \le n_1 < n_2 \le n$,

$$\mathrm{E}\left\|(n_2 - n_1 + 1)^{-1/2} \sum_{t=n_1}^{n_2} \mathbf{x}_t \varepsilon_{t,h}\right\|^5 < C_2. \tag{2.11}$$

(C5) For any $q > 0$,

$$\mathrm{E}\|\widehat{\mathbf{R}}^{-1}\|^q = O(1), \tag{2.12}$$

where for a square matrix $\mathbf{A}$, $\|\mathbf{A}\|^2 = \sup_{\|\mathbf{w}\|=1} \|\mathbf{A}\mathbf{w}\|^2$.

(C6) There exists an increasing sequence of σ-fields $\mathcal{F}_t \subseteq \mathcal{F}$ such that $\mathbf{x}_t$ is $\mathcal{F}_t$-measurable and

$$\sup_{-\infty<t<\infty} \mathrm{E}\left\|\mathrm{E}\left(\mathbf{x}_t\mathbf{x}_t^\top \middle| \mathcal{F}_{t-k}\right) - \mathbf{R}\right\|^3 = o(1), \tag{2.13}$$

$$\sup_{-\infty<t<\infty} \mathrm{E}\left\|\mathrm{E}\left(\mathbf{x}_t\varepsilon_{t,h} | \mathcal{F}_{t-k}\right)\right\|^3 = o(1), \tag{2.14}$$

as $k \to \infty$.

Some comments are in order. Suppose that $\{x_{t,i}\}$ and $\{\varepsilon_{t,h}\}$ admit linear representations,

$$x_{t,i} = \sum_{s=0}^{\infty} \mathbf{a}_{s,i}^\top \boldsymbol{\epsilon}_{t-s}, \tag{2.15}$$

and

$$\varepsilon_{t,h} = \sum_{s=0}^{\infty} \mathbf{b}_s^\top \boldsymbol{\epsilon}_{t+h-s}, \tag{2.16}$$

where $\boldsymbol{\epsilon}_t = (\epsilon_t, \epsilon_t^{(1)}, \ldots, \epsilon_t^{(m)})^\top$ is a martingale difference sequence with respect to an increasing sequence of σ-fields, say $\mathcal{G}_t$, and $\mathbf{a}_{s,i}$ and $\mathbf{b}_s$ are $(m+1)$-dimensional non-random vectors. Define $\gamma_i(k) = \mathrm{E}(x_{t,i}x_{t+k,i})$ and $\gamma(h,k) = \mathrm{E}(\varepsilon_{t,h}\varepsilon_{t+k,h})$. Then, (2.9) and (2.11) hold true, provided

$$\sum_{k=-\infty}^{\infty} (\gamma_1^2(k) + \cdots + \gamma_m^2(k)) < \infty, \sum_{k=-\infty}^{\infty} \gamma^2(h,k) < \infty, \tag{2.17}$$

(2.18)

$$\mathrm{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\top | \mathcal{G}_{t-1}) = \boldsymbol{\Sigma} \text{ and } \sup_{-\infty<t<\infty} \mathrm{E}(\|\boldsymbol{\epsilon}_t\|^{q^*} | \mathcal{G}_{t-1}) < C_{q^*} \text{ with probability 1,}$$

where $\mathbf{\Sigma}$ is a positive definite non-random matrix, $q^* > 10$ and C_{q^*} is a positive finite constant. To see this, note that by the First Moment Bound Theorem of Findley and Wei (1993) and an argument similar to that used in Lemma 2 of Ing and Wei (2003), it can be shown that (2.15)–(2.18) lead to (2.11) and (2.9), with $q_1 = q^*/2$ and C_1 and C_2 depending on q^*, C_{q^*} and $\mathbf{\Sigma}$. It may be worth pointing out that (2.15)–(2.17) are fulfilled by not only short-memory autoregressive moving average (ARMA) processes but also some long-memory processes; see Section 3 for more details. While it is possible to justify (2.9) and (2.11) under more general time series models, we leave this work for future exploration.

Condition (C2) leads to an unexpected cancelation associated with the right-hand side of (2.8) mentioned previously. The first requirement of (C2) holds when $(y_t, \mathbf{x}_t^\top)^\top$ is a fourth-order stationary process or a stationary Gaussian process, whereas the second one essentially says that the dependence between $\mathbf{x}_i\varepsilon_{i,h}$ and $\mathbf{x}_j\varepsilon_{j,h}$ vanishes sufficiently quickly as $|i-j|$ tends to infinity.

Condition (C6) requires that the conditional expectations of $\mathbf{x}_t\mathbf{x}_t^\top$ and $\mathbf{x}_t\varepsilon_{t,h}$ given $\mathcal{F}_{t-k}$ can be well approximated by their unconditional counterparts as long as k is large enough. Conditions (C5) and (C6) are used to show that the first and second terms on the right-hand side of (2.8) are asymptotically equivalent to

$$-2\mathrm{E}\{\varepsilon_{n,h}\mathbf{x}_n^\top\mathbf{R}^{-1}\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\} \text{ and } \mathrm{E}\big\{N^{-1}\sum_{t=1}^{N}(\mathbf{x}_t^\top\varepsilon_{t,h})\mathbf{R}^{-1}\sum_{t=1}^{N}(\mathbf{x}_t\varepsilon_{t,h})\big\},$$

respectively, which facilitate mathematical analysis. According to Theorem 2.1 of Chan and Ing (2011), (2.12) in (C5) is ensured by the following distributional assumption: there exist positive integer D and positive numbers δ, α and M such that for any $t > D$, any $0 < s_2 - s_1 \le \delta$ and any $\|\mathbf{v}\| = 1$,

$$P(s_1 < \mathbf{v}^\top\mathbf{x}_t \le s_2|\mathcal{F}_{t-D}) \le M\,(s_2 - s_1)^\alpha \text{ almost surely.} \tag{2.19}$$

Equation (2.19) is flexible enough to allow for a variety of time series applications. For example, Lemma S2.1 in Section S2 of Hsu et al. (2018) shows that (2.19) holds when $\mathbf{x}_t$ is the regressor of the ARX model described in Section 3. Hence (C5) is fulfilled by this type of model. In the special case of (2.5), (2.19) can be superseded by a simpler condition,

$$P(s_1 < \epsilon_t \le s_2) \le M\,(s_2 - s_1)^\alpha. \tag{2.20}$$

It is shown in Ing and Wei (2003) that (2.20) is satisfied when ϵ_t are i.i.d. with bounded density function. Finally, we mention that the moment restrictions imposed by (C1)–(C6) are by no means the weakest possible, but

they allow us to avoid unnecessary technicalities in the derivations of the key conclusions of this paper.

3. Misspecification-resistant Information Criterion. Being the population MSPE of model (2.3), the first term on the right-hand side of (2.4) is sometimes referred to as the *misspecification index* (MI) in the sequel. On the other hand, the dominant constant, L_h, associated with the second term on the right-hand side of (2.4) is refereed to as *variability index* (VI) because it is contributed by the sampling variability of $\hat{y}_{n+h} = \hat{\boldsymbol{\beta}}_n^\top(h)\mathbf{x}_n$. As revealed by (2.4), selecting the model with the smallest MSPE amounts to selecting the model with the smallest VI among those with the smallest MI.

More specifically, consider K candidate models for predicting y_{n+h}, having observations up to n,

$$y_{n+h} = \boldsymbol{\beta}_{h,l}^\top \mathbf{x}_n^{(l)} + \varepsilon_{n,h}^{(l)}, l = 1, \ldots, K, \tag{3.1}$$

where $\{\mathbf{x}_t^{(l)}\}$ is a weakly stationary processes with mean zero, $\boldsymbol{\beta}_{h,l}^\top \mathbf{x}_t^{(l)}$ is the best linear predictor of y_{t+h} based on $\mathbf{x}_t^{(l)}$, and

$$\varepsilon_{t,h}^{(l)} = y_{t+h} - \boldsymbol{\beta}_{h,l}^\top \mathbf{x}_t^{(l)}. \tag{3.2}$$

Let

$$\hat{y}_{n+h}(l) = \hat{\boldsymbol{\beta}}_{n,l}^\top(h)\mathbf{x}_n^{(l)} \tag{3.3}$$

be the least squares predictor of y_{n+h} corresponding to model l, where

$$\hat{\boldsymbol{\beta}}_{n,l}(h) = (\sum_{t=1}^{N} \mathbf{x}_t^{(l)} \mathbf{x}_t^{(l)^\top})^{-1} \sum_{t=1}^{N} \mathbf{x}_t^{(l)} y_{t+h}.$$

Throughout this section, we assume $\mathbf{R}(l) = \mathrm{E}(\mathbf{x}_1^{(l)} \mathbf{x}_1^{(l)^\top})$ is nonsingular for $l = 1, \ldots, K$. Let $\mathbf{C}_{h,s}(l) = \mathrm{E}(\mathbf{x}_1^{(l)} \mathbf{x}_{1+s}^{(l)^\top} \varepsilon_{1,h}^{(l)} \varepsilon_{1+s,h}^{(l)})$, and define

$$\mathrm{MI}_h(l) = \mathrm{E}(\varepsilon_{1,h}^{(l)})^2, \tag{3.4}$$

and

$$L_h(l) = \mathrm{tr}(\mathbf{R}^{-1}(l)\mathbf{C}_{h,0}(l)) + \sum_{s=0}^{h-1} \mathrm{tr}(\mathbf{R}^{-1}(l)\mathbf{C}_{h,s}(l)), \tag{3.5}$$

noting that (3.4) and (3.5), respectively, are the MI and the VI for model l. As mentioned, our goal is to find model $\hat{l}$ in a data-driven fashion such that

$$\lim_{n\to\infty} \mathrm{P}\left(\hat{l} \in M_2\right) = 1, \tag{3.6}$$

where

$$M_2 = \{k : k \in M_1, L_h(k) = \min_{l\in M_1} L_h(l)\}, \tag{3.7}$$

with

$$M_1 = \{k : 1 \le k \le K, \mathrm{MI}_h(k) = \min_{1\le l\le K} \mathrm{MI}_h(l)\}. \tag{3.8}$$

A model selection criterion is said to be asymptotically efficient if (3.6) is fulfilled. Section S5 of Hsu et al. (2018) provides several interesting examples showing that to achieve (3.6), one may face the challenging problem of choosing the best predictive model from those having the same MI (goodness of fit) and the same number of parameters. These examples also reveal that the best predictive model may vary with the prediction lead time h, raising another subtle issue.

Inspired by (2.4), our strategy to achieve (3.6) is to first construct the method of moments estimators of $\mathrm{MI}_h(l)$ and $L_h(l)$,

$$\hat{\sigma}_h^2(l) = N^{-1}\sum_{t=1}^{N}\left(y_{t+h} - \hat{\boldsymbol{\beta}}_{n,l}^{\top}(h)\mathbf{x}_t^{(l)}\right)^2 \equiv N^{-1}\sum_{t=1}^{N}(\hat{\varepsilon}_{t,h}^{(l)})^2,$$

and

$$\widehat{L}_h(l) = \mathrm{tr}\left(\widehat{\mathbf{R}}^{-1}(l)\widehat{\mathbf{C}}_{h,0}(l)\right) + 2\,\mathrm{tr}\left(\sum_{s=1}^{h-1}\widehat{\mathbf{R}}^{-1}(l)\widehat{\mathbf{C}}_{h,s}(l)\right),$$

respectively, where $\widehat{\mathbf{R}}(l) = N^{-1}\sum_{t=1}^{N}\mathbf{x}_t^{(l)}\mathbf{x}_t^{(l)^{\top}}$ and

$$\widehat{\mathbf{C}}_{h,s}(l) = (N-s)^{-1}\sum_{t=1}^{N-s}\mathbf{x}_t^{(l)}\mathbf{x}_{t+s}^{(l)^{\top}}\hat{\varepsilon}_{t,h}^{(l)}\hat{\varepsilon}_{t+s,h}^{(l)}.$$

We then use h-step MRIC, $\mathrm{MRIC}_h(l)$, to quantify the performance of model l, where

$$\mathrm{MRIC}_h(l) = \hat{\sigma}_h^2(l) + \frac{C_n}{n}\widehat{L}_h(l), \tag{3.9}$$

with

$$\frac{C_n}{n^{1/2}} \to \infty, \tag{3.10}$$

and

$$\frac{C_n}{n} \to 0. \tag{3.11}$$

Finally, we choose model $\hat{l}_h$, which satisfies

$$\mathrm{MRIC}_h(\hat{l}_h) = \min_{1 \le l \le K} \mathrm{MRIC}_h(l).$$

The major difference between $\mathrm{MRIC}_h(l)$ and the natural estimator $\hat{\sigma}_h^2(l) + n^{-1}\widehat{L}_h(l)$ of $\mathrm{E}\,(y_{n+h} - \hat{y}_{n+h}(l))^2$ (cf.(2.4)) is that $\mathrm{MRIC}_h(l)$ puts an additional penalty factor C_n on $\widehat{L}_h(l)$. This factor plays a crucial role in search of the best predictive model and is particularly relevant in situations where several competing models share the same MI. To see this, note first that under (3.17)–(3.21) (described below), we have

$$\hat{\sigma}_h^2(l) = \mathrm{MI}_h(l) + O_p(n^{-1/2}), \tag{3.12}$$

and

$$\widehat{L}_h(l) = L_h(l) + o_p(1), \tag{3.13}$$

yielding

$$\mathrm{MRIC}_h(l) = \mathrm{MI}_h(l) + O_p(n^{-1/2}) + \frac{C_n}{n} L_h(l) + o_p\left(\frac{C_n}{n}\right). \tag{3.14}$$

In view of (3.11), property (3.14) immediately implies

$$\lim_{n\to\infty} P(\hat{l}_h \in M_1) = 1.$$

Moreover, it follows from (3.14) and (3.10) that for $J_{l_1}, J_{l_2} \in M_1$ with $L_h(l_1) \neq L_h(l_2)$,

$$\lim_{n\to\infty} P(\mathrm{sign}(\mathrm{MRIC}_h(l_1) - \mathrm{MRIC}_h(l_2)) = \mathrm{sign}(L_h(l_1) - L_h(l_2))) = 1, \tag{3.15}$$

and hence

$$\lim_{n\to\infty} P(\hat{l}_h \in M_2) = 1. \tag{3.16}$$

The above discussion is summarized in the next theorem.

THEOREM 3.1. *Suppose for each* $1 \le l \le K$ *and* $0 \le s \le h-1$,

$$n^{-1}\sum_{t=1}^{n}(\varepsilon_{t,h}^{(l)})^2 = \mathrm{E}(\varepsilon_{1,h}^{(l)})^2 + O_p(n^{-1/2}), \tag{3.17}$$

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t^{(l)}\mathbf{x}_{t+s}^{(l)\top}\varepsilon_{t,h}^{(l)}\varepsilon_{t+s,h}^{(l)} = \mathbf{C}_{h,s}(l) + o_p(1), \tag{3.18}$$

$$n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t^{(l)}\varepsilon_{t,h}^{(l)} = O_p(1), \tag{3.19}$$

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t^{(l)}\mathbf{x}_t^{(l)\top} = \mathbf{R}(l) + o_p(1), \tag{3.20}$$

and

$$\sup_{-\infty<t<\infty}\mathrm{E}(\varepsilon_{t,h}^{(l)})^4 + \sup_{-\infty<t<\infty}\mathrm{E}\|\mathbf{x}_t^{(l)}\|^4 < \infty. \tag{3.21}$$

Then, (3.12) *and* (3.13) *hold. As a result,* (3.16) *follows.*

Remark 1. Note that (3.16) (or $\mathrm{MI}_h(l)$ and $L_h(l)$) is relevant only when the asymptotic expression (2.4) holds for each candidate model, which in turn is ensured by (C1)–(C6). If we assume that (C1)–(C6) hold for each $1 \le l \le K$, then conditions (3.19)–(3.21) can be dropped from Theorem 3.1 because they are weaker than (C4), (C1), and (C3), respectively. Another two conditions of Theorem 3.1, (3.17) and (3.18), are easily fulfilled when $\mathbf{x}_t^{(l)}$ and $\varepsilon_{t,h}^{(l)}$ are linear processes obeying (2.15) and (2.16); see Theorem 3.2 and Section S2 of Hsu et al. (2018). Moreover, if the elements in M_1 are nested, the restriction on C_n in (3.10) can be weakened to

$$C_n \to \infty, \tag{3.22}$$

and hence a weaker penalty on $\widehat{L}_h(l)$ is allowed. To see this, assume $J_{l_1}, J_{l_2} \in M_1$ with $J_{l_1} \subset J_{l_2}$ and $L_h(l_1) \ne L_h(l_2)$. Then, it can be shown that $\hat{\sigma}_h^2(l_1) - \hat{\sigma}_h^2(l_2) = O_p(1/n)$ and $\mathrm{MRIC}_h(l_1) - \mathrm{MRIC}_h(l_2) = (C_n/n)(L_h(l_1) - L_h(l_2)) + o_p(C_n/n) + O_p(1/n)$. This and (3.22) yield (3.15), and hence the desired conclusion.

Remark 2. It is shown in Sin and White (1996) and Inoue and Kilian (2006) that BIC has the so-call 'strong parsimony property' in the sense

that it will asymptotically choose the most parsimonious model among those candidates having the smallest MI. However, when two misspecified models have the same MI, the one with fewer parameters does not necessarily lead to a smaller VI; see Findley (1991) for a related discussion. Moreover, two non-nested misspecified models with the same MI may have different VIs even if they share the same number of parameters; see Section S5 of Hsu et al. (2018). In this latter case, both BIC and AIC tend to randomly choose between the two alternatives instead of selecting the one having the smaller VI. For more details on the comparison of the finite-sample performance of MRIC with AIC, BIC, GAIC, GBIC, and GAIC$_p$; see Sections S5 and S6 of Hsu et al. (2018).

Remark 3. Theorem 3.1 is readily extended to deal with multiple lead times. Assume that for each $1 \le h \le H$, there are K_h candidate models for forecasting y_{n+h}. Let $\hat{y}_{n+h}(1), \ldots, \hat{y}_{n+h}(K_h)$ denote the least squares predictors of y_{n+h} derived from these K_h models. To predict $\mathbf{y}_{n+H} = (y_{n+1}, \ldots, y_{n+H})^\top$, we use $(\hat{y}_{n+1}(l_1), \ldots, \hat{y}_{n+H}(l_H))^\top$, where $(l_1, \ldots, l_H)^\top \in \mathcal{A}_H = A_1 \times \cdots \times A_H$ with $A_h = \{1, \ldots, K_h\}$. Denote $(\hat{y}_{n+1}(l_1), \ldots, \hat{y}_{n+H}(l_H))^\top$ by $\hat{\mathbf{y}}_{n+H}(\mathbf{l})$, where $\mathbf{l} = (l_1, \ldots, l_H)^\top$. The performance of $\hat{\mathbf{y}}_{n+H}(\mathbf{l})$ is evaluated by $\mathrm{E}\|\mathbf{y}_{n+H} - \hat{\mathbf{y}}_{n+H}(\mathbf{l})\|^2$. Under the assumptions of Theorem 2.1, it holds that for each $1 \le h \le H$ and $1 \le l \le K_h$, $\lim_{n\to\infty} n\{\mathrm{E}(y_{n+h} - \hat{y}_{n+h}(l))^2 - \mathrm{MI}_h(l)\} = L_h(l)$, and hence

$$\lim_{n\to\infty} n\{\mathrm{E}\|\mathbf{y}_{n+H} - \hat{\mathbf{y}}_{n+H}(\mathbf{l})\|^2 - \mathcal{MI}_H(\mathbf{l})\} = \mathcal{L}_H(\mathbf{l}),$$

where $\mathcal{MI}_H(\mathbf{l}) = \sum_{h=1}^H \mathrm{MI}_h(l_h)$ and $\mathcal{L}_H(\mathbf{l}) = \sum_{h=1}^H L_h(l_h)$. Define

$$\mathcal{M}_1 = \{\mathbf{k} : \mathbf{k} \in \mathcal{A}_H, \mathcal{MI}_H(\mathbf{k}) = \min_{\mathbf{l}\in\mathcal{A}_H} \mathcal{MI}_H(\mathbf{l})\},$$
$$\mathcal{M}_2 = \{\mathbf{k} : \mathbf{k} \in \mathcal{M}_1, \mathcal{L}_H(\mathbf{k}) = \min_{\mathbf{l}\in\mathcal{M}_1} \mathcal{L}_H(\mathbf{l})\}.$$

By an argument similar to that used to prove Theorem 3.1, we obtain the extension

$$\lim_{n\to\infty} P(\hat{\mathbf{l}}_H \in \mathcal{M}_2) = 1,$$

where $\hat{\mathbf{l}}_H = (\hat{l}_1, \cdots, \hat{l}_H)^\top$ with $\hat{l}_h$ satisfying $\mathrm{MRIC}_h(\hat{l}_h) = \min_{1\le l\le K_h} \mathrm{MRIC}_h(l)$. In fact, based on a set of conditions similar to (C1)–(C6), extensions of Theorems 2.1 and 3.1 to a class of nonlinear models have also been obtained; see Section S4 of Hsu et al. (2018).

To further illustrate Theorems 2.1 and 3.1, we consider the following autoregressive exogenous (ARX) model,

$$\boldsymbol{\phi}(B)y_{t+1} = \sum_{v=1}^{p}\sum_{j=0}^{r_v}\eta_j^{(v)}s_{t-j}^{(v)} + \epsilon_{t+1}, \tag{3.23}$$

where B denotes the back shift operator such that $By_t = y_{t-1}$, p and r_v are positive integers, ϵ_t are independent random disturbances with $\mathrm{E}(\epsilon_t) = 0$ and $\mathrm{E}(\epsilon_t^2) = \sigma^2 > 0$, $\boldsymbol{\phi}(z) = \sum_{j=0}^{\infty}\phi_j z^j$ with $\phi_0 = 1$ and $\sum_{j=0}^{\infty}\phi_j^2 < \infty$, $\eta_j^{(v)}$ are real numbers, and $s_t^{(v)} = \sum_{j=0}^{\infty}\psi_j^{(v)}\delta_{t-j}^{(v)}$ with $\sum_{j=0}^{\infty}(\psi_j^{(v)})^2 < \infty$ and $\boldsymbol{\delta}_t(p) = (\delta_t^{(1)}, \ldots, \delta_t^{(p)})^\top$ being independent random vectors satisfying $\mathrm{E}(\boldsymbol{\delta}_t(p)) = \mathbf{0}$ and $\mathrm{E}\left(\boldsymbol{\delta}_t(p)\boldsymbol{\delta}_t^\top(p)\right) = \Sigma_p$, a p-dimensional positive definite matrix independent of t. Moreover, it is assumed that $\{\epsilon_t\}$ and $\{\boldsymbol{\delta}_t(p)\}$ are independent, for any $|z| < 1$,

$$\boldsymbol{\phi}^{-1}(z) = \boldsymbol{\theta}(z) = \sum_{j=0}^{\infty}\theta_j z^j, \text{ with } \sum_{j=0}^{\infty}\theta_j^2 < \infty, \tag{3.24}$$

and

$$\sum_{j=0}^{\infty}(c_j^{(v)})^2 < \infty, \text{ with } c_j^{(v)} = \sum_{k=0}^{j}\psi_k^{(v)}\theta_{j-k}, 1 \le v \le p. \tag{3.25}$$

We are interested in forecasting $y_{n+h}, h \ge 1$, using one of model $1, \ldots,$ model K, where the explanatory vector in model l at time t is given by

$$\mathbf{x}_t^{(l)} = (y_{t-j}, j \in J_0^{(l)}, s_{t-j}^{(v)}, j \in J_v^{(l)}, 1 \le v \le p)^\top, \tag{3.26}$$

with $J_v^{(l)}, 0 \le v \le p$, being given finite sets of non-negative integers. We illustrate that (3.26) can be misspecified via a special case of (3.23),

$$y_{t+1} = ay_t + s_t^{(1)} + \epsilon_{t+1},$$

where $0 < |a| < 1$ and $s_t^{(1)}$ is a stationary MA(1) model satisfying $\sum_{j=0}^{\infty}b^j s_{t-j}^{(1)} = \delta_t^{(1)}$ with $0 < |b| < 1$. Straightforward calculations show that the correctly specified ARX model for two-step prediction is

$$y_{t+2} = a^2 y_t + (a-b)s_t^{(1)} - \sum_{j=2}^{\infty}b^j s_{t+1-j}^{(1)} + v_{t+2},$$

where $v_{t+2} = \epsilon_{t+2} + a\epsilon_{t+1} + \delta^{(1)}_{t+1}$. Since the model involves the infinite past $s^{(1)}_t, s^{(1)}_{t-1}, \ldots$, any candidate model containing only a finite number of the lagged variables of $s^{(1)}_t$ is misspecified.

We aim at finding a data-driven method to choose among the candidate models such that (3.6) is satisfied. Let $\varepsilon^{(l)}_{t,h}$, $\hat{y}_{n+h}(l)$, $\mathrm{MI}_h(l)$, $L_h(l)$, M_2 and M_1 be defined as in (3.2)–(3.5), (3.7), and (3.8). The next theorem shows that MRIC, introduced in (3.9)–(3.11), attains the desired goal under suitable assumptions on the moments and distributions of $\mathbf{v}_t = (\boldsymbol{\delta}_t^\top(p), \epsilon_t)^\top$ as well as the decay rates of $\psi^{(v)}_j, \theta_j$ and $c^{(v)}_j$.

THEOREM 3.2. *Assume that (3.23)–(3.25) hold. Suppose that the fourth moments of $\mathbf{v}_t$ are independent of t,*

$$\sup_{-\infty<t<\infty} \mathrm{E}\|\mathbf{v}_t\|^\theta < \infty, \text{ for some } \theta > 10, \tag{3.27}$$

and there exist $K_1 > 0$, $\delta_1 > 0$ and $\nu > 0$ such that for all $-\infty < t < \infty$ and all $0 < w - u \le \delta_1$,

$$\sup_{\|\mathbf{a}\|=1} \mathrm{P}\left(u < \mathbf{a}^\top \mathbf{v}_t \le w\right) \le K_1 (w-u)^\nu. \tag{3.28}$$

Assume also that there exist $c_1 > 0$ and $s > 3/4$ for which

$$|\theta_j| \le c_1 (j+1)^{-s} \text{ and } |\psi^{(v)}_j| + |c^{(v)}_j| \le c_1 (j+1)^{-s}, 1 \le v \le p. \tag{3.29}$$

Then, (C1)–(C6) hold for $\mathbf{x}_t = \mathbf{x}^{(l)}_t$, $\varepsilon_{t,h} = \varepsilon^{(l)}_{t,h}$, and $\mathcal{F}_t = \sigma(\mathbf{v}_t, \mathbf{v}_{t-1}, \ldots)$, yielding

$$\lim_{n\to\infty} n\left\{\mathrm{E}(y_{n+h} - \hat{y}_{n+h}(l))^2 - \mathrm{MI}_h(l)\right\} = L_h(l).$$

Moreover, (3.17)–(3.21) follow, and hence (3.16) holds true.

Remark 4. Assumption (3.29) allows the component of $\mathbf{x}^{(l)}_t$ to not only be a short-memory ARMA process, but also belong to some important classes of long-memory processes, e.g., the fractionally integrated I(d) process with $-1/2 < d < 1/4$. As is clear from the proof of Theorem 3.2, (3.29) is crucial for verifying (3.18) and condition (C2), and can hardly be weakened.

Remark 5. Assumption (3.28) is used to prove (2.19), which in turn leads to condition (C5) according to Chan and Ing (2011). More details can be found

in Section S2 of Hsu et al. (2018). Note that (C5) has played an increasingly important role in deriving model selection criteria or MSPE formulas in a rigorous manner; see, for example, Findley and Wei (2002), Ing and Wei (2003, 2005), Schorfheide (2005), Chan and Ing (2011) and Greenaway-McGrevy (2013, 2015). However, most of these papers verify (C5) only in situations where regressors contain no exogenous variables.

4. An Extension to High-dimensional Misspecified Time Series Models.

4.1. *Consisiency of OGA+HDIC$_h$+Trim.* In this section, we consider the high-dimensional time series model,

$$y_{t+h} = \boldsymbol{\beta}_h^\top \mathbf{x}_t + \varepsilon_{t,h} = \sum_{j=1}^{p} \beta_{j,h} x_{t,j} + \varepsilon_{t,h}, \tag{4.1}$$

where $\{y_t\}$ and $\{\mathbf{x}_t\}$ are weakly stationary processes with mean zero, p is allowed to be larger than n, $\boldsymbol{\beta}_h$ is the unique minimizer of $\mathrm{E}(y_{t+h} - \mathbf{c}^\top \mathbf{x}_t)^2$ over $\mathbf{c} \in R^p$, and the dependence of $\varepsilon_{t,h}$ on p is suppressed in the notation. Like Section 2, this section also assumes that $\varepsilon_{t,h}$ can be serially correlated and correlated with $\mathbf{x}_k$ for $k \neq t$. In other words, model misspecification is allowed. It is worth mentioning that although high-dimensional regressions with independent observations have been extensively studied over the past decade, relatively less efforts have been devoted to the investigation of high-dimensional time series models. Aiming at bridging this gap, Basu and Michailidis (2015) and Wu and Wu (2016) have recently studied the asymptotic behavior of Lasso estimates under the following high-dimensional model,

$$y^t = \boldsymbol{\beta}^{*\top} \mathbf{x}^t + \varepsilon^t, \tag{4.2}$$

where $\{\varepsilon^t\}$ is a stationary time series, and $\{\mathbf{x}^t\}$ is a p-dimensional stationary time series independent of $\{\varepsilon^t\}$ (Basu and Michailidis, 2015) or a sequence of p-dimensional non-random vectors (Wu and Wu, 2016). However, when $\{\mathbf{x}^t\}$ is random, the assumption of independence between $\{\mathbf{x}^t\}$ and $\{\varepsilon^t\}$ not only precludes autoregressive time series, but is also often violated under model misspecification. On the other hand, (4.1) is flexible enough to accommodate these cases.

Define

$$N_h = \{j : 1 \le j \le p, \beta_{j,h} \neq 0\},$$

which is the index set corresponding to all relevant variables. In the sequel, we call the index set of a subset model of (4.1) a 'model' whenever no confusion is possible. Obviously, N_h is the smallest model among those having the lowest MI, and also the smallest true model when (4.1) is correctly specified. The goal of this subsection is to consistently estimate N_h.

Since p can be much larger than n, we introduce a recursive procedure, which we call an orthogonal greedy algorithm (OGA), to select variables one at a time. The procedure goes as follows. First, let $\hat{\mathbf{f}}^{(0)} = \mathbf{y}_h = (y_{1+h}, \ldots, y_n)^\top$ and $\hat{J}_0 = \emptyset$. For $1 \le m \le p$, $\hat{\mathbf{f}}^{(m)}$, $\hat{J}_m$, and $\hat{j}_m \in \{1, \ldots, p\}$ are given recursively by

$$\begin{aligned} &\hat{j}_m = \arg \max_{1 \le j \le p, j \notin \hat{J}_{m-1}} |\hat{\boldsymbol{\mu}}_{\hat{J}_{m-1}, j}|, \\ &\hat{J}_m = \hat{J}_{m-1} \bigcup \{\hat{j}_m\}, \\ &\hat{\mathbf{f}}^{(m)} = (\mathbf{I}_N - \mathbf{H}_{\hat{J}_m}) \mathbf{y}_h, \end{aligned} \tag{4.3}$$

where $\mathbf{H}_J, J \subset \{1, \ldots, p\}$, is the orthogonal projection matrix onto the linear span of $\{\mathbf{X}_i = (x_{1,i}, \ldots, x_{N,i})^\top, i \in J\}$, and $\hat{\boldsymbol{\mu}}_{J,i} = \mathbf{X}_i^\top (\mathbf{I}_N - \mathbf{H}_J)\mathbf{y}_h / (N^{1/2} \|\mathbf{X}_i\|)$. When the number of the OGA iterations achieves a prescribed upper bound $1 \le K_n \le p$, the algorithm outputs model $\hat{J}_{K_n}$. As shown in Theorem 4.1 below, $\hat{J}_{K_n}$ enjoys the so-called 'sure screening property' (meaning that the event $\{N_h \subseteq \hat{J}_{K_n}\}$ has a probability tending to 1 as $n \to \infty$), provided K_n is sufficiently large and conditions (F1)–(F6) below hold true.

- **(F1)** For some $q_1 \ge 2$, $\max_{1 \le i,j \le p} \mathrm{E}|n^{-1/2} \sum_{t=1}^n (z_{t,i} z_{t,j} - \rho_{i,j})|^{2q_1} = O(1)$, where $z_{t,i} = x_{t,i}/\sigma_i$, $\sigma_i^2 = \mathrm{E}(x_{t,i}^2) > 0$, and $\rho_{i,j} = \mathrm{E}(z_{t,i} z_{t,j})$.
- **(F2)** For some $q_2 \ge 2$, $\max_{1 \le i \le p} \mathrm{E}|n^{-1/2} \sum_{t=1}^n z_{t,i} \varepsilon_{t,h}|^{q_2} = O(1)$.
- **(F3)** p is a nondecreasing function of n and obeys $p^{2/q}/n = o(1)$, where $q \ge 2$ is a known lower bound for $\min\{q_1, q_2\}$.
- **(F4)** There exists some $0 < G_1 < \infty$ such that $\sum_{j=1}^p |\beta_{j,h}^*| \equiv \sum_{j=1}^p |\sigma_j \beta_{j,h}| < G_1$.
- **(F5)** For any $1 \le m \le p$, there are some $c_1, c_2 > 0$, $0 \le \theta_1 < 1$, and $\theta_2 \ge 0$ such that

$$\begin{aligned} &\min_{\sharp(J) \le m} \lambda_{\min}(\boldsymbol{\Gamma}(J)) \ge c_1 m^{-\theta_1}, \\ &\max_{\sharp(J) \le m, 1 \le i \le p, i \notin J} \|\boldsymbol{\Gamma}(J)^{-1} \mathbf{g}_i(J)\|_1 \le c_2 m^{\theta_2}, \end{aligned} \tag{4.4}$$

  where $\lambda_{\min}(\boldsymbol{A})$ denotes the minimum eigenvalue of $\boldsymbol{A}$, $\boldsymbol{\Gamma}(J) = \mathrm{E}(\mathbf{z}_t(J)\mathbf{z}_t^\top(J))$ with $\mathbf{z}_t(J) = (z_{t,j}, j \in J)^\top$, $\mathbf{g}_i(J) = \mathrm{E}(\mathbf{z}_t(J) z_{t,i})$, and $\|\cdot\|_1$ denotes the l_1 norm of a vector.

(F6) $N_h \neq \emptyset$ and for some small $\underline{\delta} > 0$,

$$\min_{j \in N_h} |\beta^*_{j,h}| \geq \underline{\delta}. \tag{4.5}$$

Remark 6. Some comments are in order. First, (F1) and (F2) are parallel to (C1) and (C4) in Section 2. As mentioned previously, these two assumptions are fulfilled when $z_{t,i}$ and $\varepsilon_{t,h}$ are linear processes with square summable autocovariance functions, and hence allow y_t and $x_{t,i}$ to be I(d) processes with $-1/2 < d < 1/4$. Note that stationary I(d) processes with $d \neq 0$ is precluded by Basu and Michailidis (2015). In addition, (F1) and (F2) are substantially weaker than sub-Gaussian and sub-exponential assumptions, which are commonly adopted in the high-dimensional statistics literature, but may seem restrictive in practice. On the other hand, since there is a tradeoff between the moment conditions and the conditions on p, the frequently used condition, $p = O(\exp(\xi n)), 0 < \xi \leq 1$, under sub-Gaussianity/sub-exponentiality is now strengthened to (F3). Condition (F5) imposes mild restrictions on the correlations among regressors. For example, it allows $\mathbf{x}_t$ to consist of a stationary I(d) variable and its lagged values. Conditions (F4) and (F6) together imply that $\sharp(N_h)$ is bounded above by a finite constant. While $\underline{\delta} > 0$ in (4.5) can be weakened to $\underline{\delta} \to 0$ at a sufficiently slow rate, such a generalization is not pursued here. Finally, we mention that our results do not rely on assumptions like $\lambda_{\max}(\mathbf{\Gamma}) < \infty$, where $\mathbf{\Gamma} = \mathrm{E}(\mathbf{z}_t \mathbf{z}_t^\top)$ with $\mathbf{z}_t = (z_{t,1}, \ldots, z_{t,p})^\top$ and $\lambda_{\max}(A)$ denotes the maximum eigenvalue of A, but this type of assumption is needed by Basu and Michailidis (2015) to derive asymptotic properties of the Lasso estimates under (4.2).

THEOREM 4.1. *Assume that* (F1)–(F6) *hold. Then, for* $K_n = \min\{p, \overline{m}_n\}$,

$$\lim_{n \to \infty} P(N_h \subset \hat{J}_{K_n}) = 1, \tag{4.6}$$

where $\{\overline{m}_n\}$ *is any nondecreasing sequence of positive integers tending to* ∞ *as* n *does.*

Remark 7. Under correctly specified high-dimensional regression models with independent observations, the sure screening property has been established for the Lasso by Bickel et al. (2009), for the OGA by Ing and Lai (2011), for the Sure Independence Screening (SIS) by Fan and Lv (2008), and a forward regression procedure by Wang (2009). Wu and Wu (2016) focused instead on high-dimensional time series models and developed the sure screening property of the Clime estimate under (4.2) with $\{\mathbf{x}^t\}$ and $\{\varepsilon^t\}$

being stationary, but not necessarily independent. However, they required that both $\{\mathbf{x}^t\}$ and $\{\varepsilon^t\}$ are short-memory time series.

While $\hat{J}_{K_n}$ possesses the sure screening property, it may contain many irrelevant indices j whose corresponding coefficients $\beta_{j,h}$ are zero. In the following, we shall choose a subset from $\hat{J}_{K_n}$ that is equivalent to N_h asymptotically. To this end, we start by introducing a high-dimensional information criterion (HDIC), which assigns a real number to a model J as follows:

$$\mathrm{HDIC}_h(J) = \big(1 + \frac{\sharp(J)p^{2/q}\omega_n}{n}\big)\hat{\sigma}_h^2(J), \tag{4.7}$$

where $\hat{\sigma}_h^2(J) = N^{-1}\mathbf{y}_h^\top(\mathbf{I}_N - \mathbf{H}_J)\mathbf{y}_h$ and $\omega_n \to \infty$ at a rate to be specified later. We then choose a subset $\hat{J}_{\hat{k}_n}$ of $\hat{J}_{K_n}$ that minimizes $\mathrm{HDIC}_h(J)$ along the OGA path. More precisely, $\hat{k}_n$ is defined to be the smallest integer k satisfying

$$\mathrm{HDIC}_h(\hat{J}_k) = \min_{1\le m\le K_n} \mathrm{HDIC}_h(\hat{J}_m). \tag{4.8}$$

Since $\hat{J}_{\hat{k}_n}$ may still contain redundant indices, we further trim $\hat{J}_{\hat{k}_n}$ by making use of HDIC_h to come up with

$$\hat{N}_h = \begin{cases} \{\hat{j}_k : 1 \le k \le \hat{k}_n, \mathrm{HDIC}_h(\hat{J}_{\hat{k}_n}) < \mathrm{HDIC}_h(\hat{J}_{\hat{k}_n} - \{\hat{j}_k\})\}, & \hat{k}_n > 1, \\ \{\hat{j}_1\}, & \hat{k}_n = 1. \end{cases} \tag{4.9}$$

The above model selection procedure is referred to as "OGA+HDIC_h+Trim". The main result of this section is reported in the next theorem.

THEOREM 4.2. *Assume* (F1)–(F6), *and*

$$n^{-1}\sum_{t=1}^{n} \varepsilon_{t,h}^2 = \mathrm{E}(\varepsilon_{1,h}^2) + o_p(1). \tag{4.10}$$

Suppose that K_n *and* ω_n *satisfy*

$$K_n = \min\{p, \overline{m}_n\},\ \omega_n \to \infty,\ \omega_n = O(n^{1/2}/p^{1/q}), \tag{4.11}$$

where $\{\overline{m}_n\}$ *is a sequence of positive integers obeying*

$$\overline{m}_n \to \infty,\ \overline{m}_n^{\theta_1+2\theta_2} = o(\omega_n),\ \overline{m}_n^{1+\max\{\theta_1,\theta_2\}} = o(n^{1/2}/p^{1/q}). \tag{4.12}$$

Then,

$$\lim_{n\to\infty} P(\hat{N}_h = N_h) = 1. \tag{4.13}$$

To the best of our knowledge, Theorem 4.2 is the first result showing that selection consistency is still achievable under high-dimensional misspecified time series models. It is further shown in Sections S5 and S6 of Hsu et al. (2018) that OGA+HDIC$_h$+Trim has satisfactory finite-sample performance.

4.2. *Asymptotically Efficient Model Selection across Several High-dimensional Time Series Models.* In real world situations, prediction is often conducted by several different forecasters. Some forecasters may live in a variable-rich environment, where hundreds and thousands of variables are readily accessible, whereas others may rely more on rich domain-specific knowledge, and hence only require a relatively small set of candidate variables. Specifically, assume that there are K (high-dimensional) models,

$$y_{t+h} = \boldsymbol{\beta}_{h,l}^{\top}\mathbf{x}_t^{(l)} + \varepsilon_{t,h}^{(l)} = \sum_{j=1}^{p_l} \beta_{j,h}^{(l)} x_{t,j}^{(l)} + \varepsilon_{t,h}^{(l)},\ l = 1, \ldots, K, \tag{4.14}$$

proposed by K different forecasters, where $\{x_{t,j}^{(l)}, 1 \le j \le p_l\}$ are the candidate variables employed by the lth forecaster at time t, with p_l, the number of candidate variables, varying from one model to another. Define

$$N_h^{(l)} = \{j : 1 \le j \le p_l, \beta_{j,h}^{(l)} \ne 0\}.$$

In addition to identifying $N_h^{(l)}, 1 \le l \le K$, the goal of this section is to find the best pair among $(l, N_h^{(l)}), l = 1, \ldots, K$, in terms of their prediction capabilities.

Let $J \subset \{1, \ldots, p_l\} \equiv \mathcal{P}_l$ be a model in the lth candidate family, i.e., the family of all possible subsets of the lth model in (4.14). In view of (3.4) and (3.5), the MI and VI of J are given by $\mathrm{MI}_{h,l}(J) = \mathrm{E}(\varepsilon_{n,h}^{(l)}(J))^2$ and

$$L_{h,l}(J) = \operatorname{tr}\left(\mathbf{R}^{(l)^{-1}}(J)\mathbf{C}_{h,0}^{(l)}(J)\right) + 2\operatorname{tr}\left(\sum_{s=1}^{h-1} \mathbf{R}^{(l)^{-1}}(J)\mathbf{C}_{h,s}^{(l)}(J)\right),$$

respectively, where

$$\begin{aligned}
&\varepsilon_{t,h}^{(l)}(J) = y_{t+h} - \boldsymbol{\beta}_{h,l}^{\top}(J)\mathbf{x}_t^{(l)}(J),\\
&\mathbf{R}^{(l)}(J) = \mathrm{E}(\mathbf{x}_t^{(l)}(J)\mathbf{x}_t^{(l)^{\top}}(J)),\\
&\mathbf{C}_{h,s}^{(l)}(J) = \mathrm{E}(\mathbf{x}_t^{(l)}(J)\mathbf{x}_{t+s}^{(l)^{\top}}(J)\varepsilon_{t,h}^{(l)}(J)\varepsilon_{t+s,h}^{(l)}(J)),
\end{aligned}$$

with $\mathbf{x}_t^{(l)}(J) = (x_{t,j}^{(l)}, j \in J)^\top$ and $\boldsymbol{\beta}_{h,l}(J) = \arg\min_{\mathbf{c}\in R^{\sharp(J)}} \mathrm{E}(y_{t+h} - \mathbf{c}^\top \mathbf{x}_t^{(l)}(J))^2$. It is clear that $\mathrm{MI}_{h,l}(\mathcal{P}_l) = \mathrm{MI}_{h,l}(N_h^{(l)})$. In addition, (3.8) and (3.7) motivate us to define

$$M_{A,h} = \left\{ l : 1 \le l \le K, \mathrm{MI}_{h,l}(N_h^{(l)}) = \min_{1\le j\le K} \mathrm{MI}_{h,j}(N_h^{(j)}) \right\},$$

$$M_{B,h} = \left\{ l : L_{h,l}(N_h^{(l)}) = \min_{j\in M_{A,h}} L_{h,j}(N_h^{(j)}) \right\},$$

and

$$M_{C,h} = \left\{ (l, N_h^{(l)}) : l \in M_{B,h} \right\},$$

noting that $M_{C,h}$ is the collection of the (asymptotically) best forecaster-model pairs for h-step prediction. We aim at proposing a data-driven $(\hat{l}, \hat{J})$, where $1 \le \hat{l} \le K$ and $\hat{J} \subseteq \{1, \ldots, p_{\hat{l}}\}$, such that

$$\lim_{n\to\infty} P((\hat{l}, \hat{J}) \in M_{C,h}) = 1. \tag{4.15}$$

Define $(\sigma_i^{(l)})^2 = \mathrm{E}(x_{t,i}^{(l)})^2$, $z_{t,i}^{(l)} = x_{t,i}^{(l)}/\sigma_i^{(l)}$, $\beta_{j,h}^{(l)*} = \beta_{j,h}^{(l)}\sigma_j^{(l)}$, $\boldsymbol{\beta}_h^{(l)*} = (\beta_{j,h}^{(l)*}, 1 \le j \le p_l)^\top$, $\rho_{i,j}^{(l)} = \mathrm{E}(z_{t,i}^{(l)} z_{t,j}^{(l)})$, $\mathbf{z}_t^{(l)} = (z_{t,i}^{(l)}, 1 \le i \le p_l)^\top$, $\mathbf{z}_t^{(l)}(J) = (z_{t,i}^{(l)}, i \in J \subseteq \mathcal{P}_l)^\top$, and $\mathbf{g}_i^{(l)}(J) = \mathrm{E}(\mathbf{z}_t^{(l)}(J) z_{t,i}^{(l)})$. We assume that for each $1 \le l \le K$, there exist $0 \le \theta_{1,l} < 1$, $\theta_{2,l} \ge 0$, and positive numbers $q_{1,l}, q_{2,l}, G_{1,l}, c_{1,l}, c_{2,l}$ and $\underline{\delta}_l$ such that (F1(l))–(F6(l)) hold, where (F1(l))–(F6(l)) are (F1)–(F6) with $\mathbf{z}_t$, $\rho_{i,j}$, $\varepsilon_{t,h}$, p, $\boldsymbol{\beta}_h^*$, N_h, $\boldsymbol{\Gamma}(J)$, and $\boldsymbol{\Gamma}(J)^{-1}\mathbf{g}_i(J)$ therein replaced by $\mathbf{z}_t^{(l)}$, $\rho_{i,j}^{(l)}$, $\varepsilon_{t,h}^{(l)}$, p_l, $\boldsymbol{\beta}_h^{(l)*}$, $N_h^{(l)}$, $\boldsymbol{\Gamma}_l(J) = \mathrm{E}(\mathbf{z}_t^{(l)}(J)\mathbf{z}_t^{(l)^\top}(J))$, and $\boldsymbol{\Gamma}_l^{-1}(J)\mathbf{g}_i^{(l)}(J)$, and with θ_1 θ_2, q_1, q_2, q, G_1, c_1, c_2, and $\underline{\delta}$ replaced by $\theta_{1,l}$ $\theta_{2,l}$, $q_{1,l}, q_{2,l}, q_l = \min\{q_{1,l}, q_{2,l}\}, G_{1,l}, c_{1,l}$, $c_{2,l}$, and $\underline{\delta}_l$. Moreover, define

$$\mathrm{HDIC}_{h,l}(J) = \big(1 + \frac{\sharp(J) p_l^{2/q_l} \omega_n^{(l)}}{n}\big)\hat{\sigma}_{h,l}^2(J), \tag{4.16}$$

where $\hat{\sigma}_{h,l}^2(J) = N^{-1}\mathbf{y}_h^\top(\mathbf{I}_N - \mathbf{H}_J^{(l)})\mathbf{y}_h$, with $\mathbf{H}_J^{(l)}$ denoting the orthogonal projection matrix onto the linear span of the set of vectors $\{\mathbf{X}_j^{(l)} = (x_{1,j}^{(l)}, \ldots, x_{n,j}^{(l)})^\top, j \in J\}$, and $\omega_n^{(l)} \to \infty$ at a suitable rate.

Our strategy is to use OGA+HDIC$_{h,l}$+Trim to determine a model, $\hat{N}_h^{(l)}$, from the lth candidate family, and then employ MRIC to choose among $\hat{N}_h^{(l)}, l = 1, \ldots, K$. This procedure starts with applying the OGA to each model in (4.14), yielding

$$\hat{J}_{K_n^{(l)}}^{(l)} = \{\hat{j}_1^{(l)}, \ldots, \hat{j}_{K_n^{(l)}}^{(l)}\},\ l = 1, \ldots, K,$$

where $K_n^{(l)}$ is a prescribed upper bound for the number of iterations when the OGA is applied to the lth model in (4.14). Let $\hat{k}_n^{(l)}$ be the smallest integer k such that

$$\text{HDIC}_{h,l}(\hat{J}_k^{(l)}) = \min_{1\le m\le K_n^{(l)}} \text{HDIC}_{h,l}(\hat{J}_m^{(l)}), \tag{4.17}$$

where $\hat{J}_m^{(l)} = \{\hat{j}_1^{(l)}, \ldots, \hat{j}_m^{(l)}\}$. Then, $\hat{N}_h^{(l)}$ is given by

(4.18)
$$\hat{N}_h^{(l)} = \begin{cases} \{\hat{j}_k^{(l)} : 1 \le k \le \hat{k}_n^{(l)}, \text{HDIC}_{h,l}(\hat{J}_{\hat{k}_n^{(l)}}) < \text{HDIC}_{h,l}(\hat{J}_{\hat{k}_n^{(l)}} - \{\hat{j}_k^{(l)}\})\}, & \hat{k}_n^{(l)} > 1, \\ \{\hat{j}_1^{(l)}\}, & \hat{k}_n^{(l)} = 1. \end{cases}$$

The last step of this procedure is to choose $\hat{N}_h^{(\hat{l}_h)}$ from $\{\hat{N}_h^{(j)}, 1 \le j \le K\}$, where $\hat{l}_h$ satisfies

$$\text{MRIC}_{h,l}(\hat{N}_h^{(\hat{l}_h)}) = \min_{1\le j\le K} \text{MRIC}_{h,j}(\hat{N}_h^{(j)}).$$

Here for $J \subset \mathcal{P}_l$,

$$\text{MRIC}_{h,l}(J) = \hat{\sigma}_{h,l}^2(J) + \frac{C_n}{n}\widehat{L}_{h,l}(J), \tag{4.19}$$

in which C_n obeys (3.10) and (3.11), and

$$\widehat{L}_{h,l}(J) = \text{tr}\left(\widehat{\mathbf{R}}^{(l)^{-1}}(J)\widehat{\mathbf{C}}_{h,0}^{(l)}(J)\right) + 2\,\text{tr}\left(\sum_{s=1}^{h-1}\widehat{\mathbf{R}}^{(l)^{-1}}(J)\widehat{\mathbf{C}}_{h,s}^{(l)}(J)\right),$$

with

$$\begin{aligned}
&\widehat{\mathbf{R}}^{(l)} = N^{-1}\sum_{t=1}^{N}\mathbf{x}_t^{(l)}(J)\mathbf{x}_t^{(l)^\top}(J),\\
&\widehat{\mathbf{C}}_{h,s}^{(l)}(J) = (N-s)^{-1}\sum_{t=1}^{N-s}\mathbf{x}_t^{(l)}(J)\mathbf{x}_{t+s}^{(l)^\top}(J)\hat{\varepsilon}_{t,h}^{(l)}(J)\hat{\varepsilon}_{t+s,h}^{(l)}(J),\\
&\hat{\varepsilon}_{t,h}^{(l)}(J) = y_{t+h} - \hat{\boldsymbol{\beta}}_{h,l}^\top(J)\mathbf{x}_t^{(l)}(J),\\
&\hat{\boldsymbol{\beta}}_{h,l}(J) = (\sum_{t=1}^{N}\mathbf{x}_t^{(l)}(J)\mathbf{x}_t^{(l)^\top}(J))^{-1}\sum_{t=1}^{N}\mathbf{x}_t^{(l)}(J)y_{t+h}.
\end{aligned}$$

The above model selection procedure is referred to as 'OGA+HDIC$_{h,l}$+Trim+MRIC'. The next theorem shows that $(\hat{l}_h, \hat{N}_h^{(\hat{l}_h)})$ satisfies (4.15).

THEOREM 4.3. *Assume that for* $l = 1, \ldots, K$, (F1(l))–(F6(l)), (3.17),

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t^{(l)}(N_h^{(l)})\mathbf{x}_{t+s}^{(l)\top}(N_h^{(l)})\varepsilon_{t,h}^{(l)}\varepsilon_{t+s,h}^{(l)} = C_{h,s}^{(l)}(N_h^{(l)}) + o_p(1), \tag{4.20}$$

and

$$\sup_{-\infty<t<\infty}\mathrm{E}(\varepsilon_{t,h}^{(l)})^4 + \sup_{-\infty<t<\infty}\mathrm{E}\|\mathbf{x}_t^{(l)}(N_h^{(l)})\|^4 < \infty \tag{4.21}$$

hold true. Moreover, suppose for $l = 1, \ldots, K$,

$$K_n^{(l)} = \min\{p_l, \overline{m}_n^{(l)}\},\ \omega_n^{(l)} \to \infty,\ \omega_n^{(l)} = O(n^{1/2}/p_l^{1/q_l}), \tag{4.22}$$

where $\overline{m}_n^{(l)}$ *obeys*

$$\overline{m}_n^{(l)} \to \infty,\ (\overline{m}_n^{(l)})^{\theta_{1,l}+2\theta_{2,l}} = o(\omega_n^{(l)}),\ (\overline{m}_n^{(l)})^{1+\max\{\theta_{1,l},\theta_{2,l}\}} = o(n^{1/2}/p_l^{1/q_l}). \tag{4.23}$$

Then,

$$\lim_{n\to\infty} P\left((\hat{l}_h, \hat{N}_h^{(\hat{l}_h)}) \in M_{C,h}\right) = 1. \tag{4.24}$$

Remark 8. Because $N_h^{(l)}, l = 1, \ldots, K$, are not necessarily nested, the condition on $n^{-1}\sum_{t=1}^{n}(\varepsilon_{t,h}^{(l)})^2$ in Theorem 4.3 is the same as the one in Theorem 3.1, but is more stringent than conditions like (4.10). We also note that (4.20) and (4.21) are analogous to (3.18) and (3.21) of Theorem 3.1, respectively.

When compared to existing high-dimensional model selection methods, the most appealing feature of OGA+HDIC$_{h,l}$+Trim+MRIC is that it can select the (asymptotically) best forecaster-model combination in situations where predictions are made by several forecasters, using different (possibly misspecified) high-dimensional time series models. The advantage of OGA+HDIC$_{h,l}$+Trim+MRIC is also demonstrated via simulations in Section S5 of Hsu et al. (2018).

5. Conclusions. This paper has addressed a serious lacuna that has attracted little attention in the vast literature on model selection. We argue that in many realistic applications, we are faced with the problem of selecting a model from a *finite* and *fixed* collection of models, without knowing whether the true DGP is included in it or not, and without recourse to the

mathematical device of allowing the true DGP to be well approximated by an increasing sequence of candidate models. If we accept the partially tautological proposition that 'all models are wrong, but some are useful', then we are often faced with precisely the above fundamental issue.

The MRIC gives an explicit expression, namely equation (3.9), which addresses not only the one-step ahead prediction but also the multi-step case. We have shown how we can compute the explicit expressions and given detailed theoretical underpinnings. Moreover, with the help of OGA+HDIC$_h$+Trim, MRIC can even be used to identify the best subset across several high-dimensional misspecified time series models. It is hoped that filling the serious lacuna paves the way for the beginning of the final phase of the model selection enterprise started by Akaike, Mallows and others more than forty years ago.

Finally, in all the model selection criteria discussed in this paper, estimation of unknown parameters is rooted in the likelihood function or its equivalents. For misspecified models, attempts to justify the likelihood-based approach to estimation are often made by reference to the Kullbeck-Leibler information, which is well known to be *not* a distance measure. However, alternative (i.e., non-likelihood-based) approaches are available and beginning to attract attention; see, e.g., Davies (2008), Xia and Tong (2011) and others. Therefore, it remains a future challenge to develop a model selection criterion via a non-likelihood-based approach.

Appendix A: On Model Misspecification. This appendix provides a definition of 'model misspecification' with respect to (w.r.t.) an increasing sequence of σ-fields, $\{\mathcal{G}_t\}$, satisfying $\sigma(\mathbf{x}_j, j \le t) \subseteq \mathcal{G}_t \subseteq \mathcal{F}$, where $\mathbf{x}_j$ and $\mathcal{F}$ are defined at the beginning of Section 2 and $\sigma(\mathbf{x}_j, j \le t)$ denotes the σ-field generated by $\{\mathbf{x}_j, j \le t\}$. Model (2.3) is said to be correctly specified w.r.t. $\{\mathcal{G}_t\}$ if for any $-\infty < t < \infty$,

$$\mathrm{E}(y_{t+h}|\mathcal{G}_t) = \boldsymbol{\beta}_h^\top \mathbf{x}_t \text{ almost surely,} \tag{A.1}$$

otherwise it is called misspecified w.r.t. $\{\mathcal{G}_t\}$.

If (A.1) holds true, then it is easy to see that $\mathrm{E}(\mathbf{x}_{t-j}\varepsilon_{t,h}) = \mathbf{0}$ for any $j \ge 0$, where $\varepsilon_{t,h} = y_{t+h} - \boldsymbol{\beta}_h^\top \mathbf{x}_t$. To gain a better understanding of the concept of model misspecification, we assume that the data are generated by the following model,

$$y_{t+1} = ax_t + bw_t + \varepsilon_{t+1}, \tag{A.2}$$

where $ab \neq 0$, $\{\varepsilon_t\}$ is a sequence of i.i.d. random errors with $\mathrm{E}(\varepsilon_1) = 0$ and $0 < \mathrm{E}(\varepsilon_1^2) < \infty$, and $\{(x_t, w_t)^\top\}$ is a sequence of i.i.d. bivariate normal random vectors with $\mathrm{E}(x_1) = \mathrm{E}(w_1) = 0$, $\mathrm{E}(x_1^2) = \mathrm{E}(w_1^2) = 1$, and $0 < |\sigma_{1,2}| = |\mathrm{E}(x_1 w_1)| < 1$. We also assume that $\{\varepsilon_t\}$ and $\{(x_t, w_t)^\top\}$ are independent. Let $\mathcal{G}_t = \sigma(x_j, j \leq t)$. Then,

$$\mathrm{E}(y_{t+1}|\mathcal{G}_t) = \mathrm{E}(y_{t+1}|x_t) = (a + b\sigma_{1,2})x_t \text{ almost surely,}$$

Therefore, the simple regression model $(a+b\sigma_{1,2})x_t$ is correctly specified w.r.t. $\{\mathcal{G}_t\}$.

Alternatively, assume in (A.2), $x_t = \xi x_{t-1} + \delta_t$ and $w_t = \theta w_{t-2} + \eta_t$, where $0 < |\xi|, |\theta| < 1$ and $\{(\delta_t, \eta_t)^\top\}$ is a sequence of i.i.d. bivariate normal random vectors independent of $\{\varepsilon_t\}$, and satisfies $\mathrm{E}(\delta_1) = \mathrm{E}(\eta_1) = 0$, $\mathrm{E}(\delta_1^2) = 1 - \xi^2$, $\mathrm{E}(\eta_1^2) = 1 - \theta^2$, and $0 < \nu_{1,2}^2 = (\mathrm{E}(\delta_1\eta_1))^2 < (1 - \xi^2)(1 - \theta^2)$. Then, it can be shown that

(A.3)
$$\mathrm{E}(y_{t+1}|\mathcal{G}_t) = ax_t + \frac{b\nu_{1,2}}{1 - \xi^2}\sum_{j=0}^{\infty}\theta^j(x_{t-2j} - \xi x_{t-2j-1}) \text{ almost surely,}$$

and hence the simple regression model $\beta_{1,1}x_t$, where

$$\beta_{1,1} = a + \frac{b\nu_{1,2}}{1 - \theta\xi^2} = \arg\min_{c \in R}\mathrm{E}(y_{t+1} - cx_t)^2,$$

is no longer correctly specified w.r.t. $\{\mathcal{G}_t\}$. Moreover, since

$$\mathrm{E}(y_{t+1}|\mathcal{G}_t^{'}) = ax_t + bw_t \text{ almost surely,}$$

where $\mathcal{G}_t^{'} = \sigma(x_j, w_j, j \leq t)$, the model on the right-hand side of (A.3) is correctly specified w.r.t. $\{\mathcal{G}_t\}$, but misspecified w.r.t. $\{\mathcal{G}_t^{'}\}$.

Acknowledgments. We would like to thank an Associate Editor and two anonymous referees for their insightful and constructive comments, which greatly improve the presentation of this paper.

SUPPLEMENTARY MATERIAL

Supplement to "On model selection from a finite family of possibly misspecified time series models". The supplementary material contains the proofs of all theorems, an extension of MRIC to a class of nonlinear models, and simulation studies and real data analysis to illustrate the performance of the proposed methods in both low- and high-dimensional cases.

References.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control.*, **19**, 716–723.

Akaike, H. (1978). On the likelihood of a time series model. *J. Roy. Statist. Soc. Ser. D*, 217–235.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, **43**, 1535–1567.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *J. Math. Psych.*, **44**, 62–91.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer-Verlag.

Chan, N. H., and Ing, C.-K. (2011). Uniform moment bounds of Fisher's information with applications to time series. *Ann. Statist.*, **39**, 1526–1550.

Davies, P. L. (2008). Approximating data (with discussion). *J. Korean Statist. Soc.*, **37**, 191–240.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **70**, 849–911.

Findley, D. F. (1991). Counterexamples to parsimony and BIC. *Ann. Inst. Statist. Math.*, **43**, 505–514.

Findley, D. F., and Wei, C. Z. (1993). Moment bounds for deriving time series CLT's and model selection procedures. *Statist. Sinica*, **3**, 453–480.

Findley, D. F., and Wei, C. Z. (2002). AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. *J. Multivariate Anal.*, **83**, 415–450.

Greenaway-McGrevy, R. (2013). Multistep prediction of panel vector autoregressive processes. *Economet. Theory*, **29**, 699–734.

Greenaway-McGrevy, R. (2015). Evaluating panel data forecasts under independent realization. *J. Multivariate Anal.*, **136**, 108–125.

Hsu, H.-L., Ing, C.-K., and Tong, H. (2018). Supplement to "On model selection from a finite family of possibly misspecified time series models."

Ing, C.-K. (2003). Multistep prediction in autoregressive processes. *Economet. Theory*, **19**, 254–279.

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

Ing, C.-K., and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica*, **21**, 1473–1513.

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

Ing, C.-K., and Wei, C. Z. (2005). Order selection for same-realization predictions in autoregressive processes. *Ann. Statist.*, **33**, 2423–2474.

Inoue, A. and Kilian, L. (2006). On the selection of forecasting models. *J. Econometrics*, **130**, 273–306.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, **15**, 958–975.

Liu, W. and Yang, Y. (2011). Parametric or nonparametric? a parametricness index for

model selection. *Ann. Statist.*, **39**, 2074–2102.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **76**, 141–167.

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

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369–374.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, **7**, 221–264.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.*, **8**, 147–164.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

Schorfheide, F. (2005). VAR forecasting under misspecification. *J. Econometrics*, **128**, 99–136.

Sin, C. Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *J. Econometrics*, **71**, 207–225.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–620.

Takeuchi, K. (1976). The distribution of information statistic and the criterion of the adequacy of a model. *Suri-Kagaku (Mathematical Sciences)*, **3**, 12–18, (in Japanese).

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **58**, 267–288.

van Erven, T., Grünwald, P. D., and De Rooij, S. (2012). Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC - BIC dilemma. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **74**, 361–417.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.*, **104**, 1512–1524.

Wei, C. Z. (1992). On predictive least squares principles. *Ann. Statist.*, **20**, 1–42.

Wu, W. B. and Wu, Y. N. (2016). Performance Bounds for Parameter Estimates of High-dimensional Linear Models with Correlated Errors. *Electron. J. Statist.*, **10**, 352–379.

Xia, Y. and Tong, H. (2011). Feature matching (with discussion). *Statist. Science*, **26**, 21–46.

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

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

Supplement to "On model selection from a finite family of possibly misspecified time series models"

By Hsiang-Ling Hsu, Ching-Kang Ing, and Howell Tong

National University of Kaohsiung, National Tsing Hua University, and University of Electronic Science & Technology and London School of Economics

The supplementary material consists of 6 sections. The proofs of the theorems in Sections 2–4 of Hsu et al. (2018) are given in Sections S1–S3, respectively. Section S4 provides an extension of MRIC to a class of nonlinear models. Section S5 compares the performance of MRIC with AIC, BIC, GAIC, GBIC, and GBIC$_p$ in both low- and high-dimensional cases based on simulated data. The performance of these criteria is also compared via two real datasets in Section S6.

S1. Proof of Theorem 2.1. In view of (2.8), we have

$$N\left\{\mathrm{E}\left(y_{n+h}-\hat{\boldsymbol{\beta}}_n^{\top}(h)\mathbf{x}_n\right)^2-\mathrm{E}\left(\varepsilon_{n,h}^2\right)\right\}:=(\mathrm{I})+(\mathrm{II}), \tag{S1.1}$$

where $(\mathrm{I})=-2\mathrm{E}(\varepsilon_{n,h}\mathbf{x}_n^{\top}\widehat{\mathbf{R}}_N^{-1}\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h})$ and $(\mathrm{II})=\mathrm{E}(\mathbf{x}_n^{\top}\widehat{\mathbf{R}}_N^{-1}N^{-1/2}\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h})^2$. It is shown in Lemma S1.1 below that

$$(\mathrm{I})=-2\mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^{\top}\mathbf{R}^{-1}\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right)+o(1):=(\mathrm{III})+o(1), \tag{S1.2}$$

and

$$(\mathrm{II})=\mathrm{E}\left\{\left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right)^{\top}\mathbf{R}^{-1}\left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right)\right\}+o(1):=(\mathrm{IV})+o(1). \tag{S1.3}$$

By (C2), it follows that

(S1.4)

$$
\begin{aligned}
\text{(IV)} &= \frac{1}{N}\mathrm{E}\left(\sum_{t=1}^{N}\mathbf{x}_t^\top\mathbf{R}^{-1}\mathbf{x}_t\varepsilon_{t,h}^2\right) + \frac{2}{N}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\mathrm{E}\left(\mathbf{x}_j^\top\mathbf{R}^{-1}\mathbf{x}_k\varepsilon_{j,h}\varepsilon_{k,h}\right) \\
&= \mathrm{tr}\left\{\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_1^\top\varepsilon_{1,h}^2\right)\right\} + \frac{2}{N}\left\{\sum_{j=1}^{N-1}(N-j)\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right)\right\} \\
&= \mathrm{tr}\left\{\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_1^\top\varepsilon_{1,h}^2\right)\right\} + 2\sum_{s=1}^{h-1}\mathrm{tr}\left\{\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_{1+s}^\top\varepsilon_{1,h}\varepsilon_{1+s,h}\right)\right\} \\
&+ 2\sum_{j=h}^{N-1}\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right) - \frac{2}{N}\sum_{j=1}^{N-1}j\,\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right) \\
&= \mathrm{tr}\left(\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_1^\top\varepsilon_{1,h}^2\right)\right) + 2\sum_{s=1}^{h-1}\mathrm{tr}\left(\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_{1+s}^\top\varepsilon_{1,h}\varepsilon_{1+s,h}\right)\right) \\
&+ 2\sum_{j=h}^{N-1}\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right) + o(1),
\end{aligned}
$$

and

(S1.5)

$$
\begin{aligned}
\text{(III)} &= -2\sum_{j=h}^{n-1}\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right) \\
&= -2\sum_{j=h}^{N-1}\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right) + o(1).
\end{aligned}
$$

Since the third term on the right-hand side of (S1.4) and the first term on the right-hand side of (S1.5) cancel out each other, it follows from (S1.1)–(S1.5) that

$$
\begin{aligned}
&N\left\{\mathrm{E}\left(y_{n+h} - \hat{\boldsymbol{\beta}}_n^\top(h)\mathbf{x}_n\right)^2 - \mathrm{E}\left(\varepsilon_{n,h}^2\right)\right\} \\
&= \mathrm{tr}\left(\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_1^\top\varepsilon_{1,h}^2\right)\right) + 2\sum_{s=1}^{h-1}\mathrm{tr}\left(\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_{1+s}^\top\varepsilon_{1,h}\varepsilon_{1+s,h}\right)\right) + o(1),
\end{aligned}
$$

yielding the desired conclusion.

LEMMA S1.1. *Under the assumptions of Theorem 2.1, (S1.2) and (S1.3) follow.*

PROOF. Let $\{l_n\}$ be a sequence of positive integers satisfying $l_n \to \infty$ and $l_n/n^{1/2} = o(1)$. By (2.14) in (C6), we have

(S1.6)
$$\mathrm{E}\left(\|\mathrm{E}\left(\mathbf{x}_n\varepsilon_{n,h}|\mathcal{F}_{n-l_n}\right)\|^3\right) \le \sup_{-\infty<t<\infty} \mathrm{E}\left(\|\mathrm{E}\left(\mathbf{x}_t\varepsilon_{t,h}|\mathcal{F}_{t-l_n}\right)\|^3\right) = o(1).$$

Similarly, (2.13) in (C6) implies

$$\mathrm{E}\left\|\mathrm{E}\left(\mathbf{x}_n\mathbf{x}_n^\top|\mathcal{F}_{n-l_n}\right) - \mathbf{R}\right\|^3 = o(1). \tag{S1.7}$$

It suffices for (S1.2) to prove that

$$\mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^\top\left(\widehat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\right)\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right) = o(1). \tag{S1.8}$$

To show (S1.8), first observe that

(S1.9)
$$\begin{aligned}
&\mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^\top\left(\widehat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\right)\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right)\\
&= \mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^\top\left(\widehat{\mathbf{R}}^{-1} - \widetilde{\mathbf{R}}^{-1}\right)\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right) + \mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^\top\left(\widetilde{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\right)\sum_{t=n-l_n-h+1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right)\\
&+ \mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^\top\left(\widetilde{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\right)\sum_{t=1}^{n-l_n-h}\mathbf{x}_t\varepsilon_{t,h}\right) := \mathrm{(I)} + \mathrm{(II)} + \mathrm{(III)},
\end{aligned}$$

where $\widetilde{\mathbf{R}} = (n-l_n)^{-1}\sum_{t=1}^{n-l_n}\mathbf{x}_t\mathbf{x}_t^\top$. Since for large n,

$$\|\widehat{\mathbf{R}}^{-1} - \widetilde{\mathbf{R}}^{-1}\| \le \|\widehat{\mathbf{R}}^{-1}\|\|\widetilde{\mathbf{R}}^{-1}\|\left(\left\|\frac{1}{N}\sum_{t=n-l_n+1}^{N}\mathbf{x}_t\mathbf{x}_t^\top\right\| + \left\|\left(\frac{1}{N} - \frac{1}{n-l_n}\right)\sum_{t=1}^{n-l_n}\mathbf{x}_t\mathbf{x}_t^\top\right\|\right).$$

This, together with (2.12), (2.9), and Hölder's inequality, yields for any $0 < \gamma \le 5$,

$$\mathrm{E}\|\widehat{\mathbf{R}}^{-1} - \widetilde{\mathbf{R}}^{-1}\|^\gamma = O\left((l_n/n)^\gamma\right). \tag{S1.10}$$

Applying Hölder's inequality again, we have

(S1.11)

$$\text{(I)} \leq \left(\mathrm{E}|\varepsilon_{n,h}|^6\right)^{1/6} \left(\|\mathbf{x}_n\|^6\right)^{1/6} \left(\mathrm{E}\|\widehat{\mathbf{R}}^{-1} - \widetilde{\mathbf{R}}^{-1}\|^3\right)^{1/3} \left(\mathrm{E}\left\|\sum_{t=1}^{N} \mathbf{x}_t \varepsilon_{t,h}\right\|^3\right)^{1/3}.$$

By (S1.10), (S1.11), (C3) and (C4), it holds that

$$\text{(I)} = o(l_n/n^{1/2}) = o(1). \tag{S1.12}$$

An argument similar to that used to prove (S1.10) yields for any $0 < \gamma \leq 5$,

$$\mathrm{E}\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\|^\gamma = O(n^{-\gamma/2}), \quad \mathrm{E}\|\widetilde{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\|^\gamma = O(n^{-\gamma/2}). \tag{S1.13}$$

By making use of (S1.13), Hölder's inequality, (C3), (C4) and (S1.6), we obtain

$$\text{(II)} = O\left((l_n/n)^{1/2}\right) = o(1), \tag{S1.14}$$

and

$$\begin{aligned} \text{(III)} \leq{}& \left(\mathrm{E}\|\widetilde{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\|^3\right)^{1/3} \left(\mathrm{E}\left\|\sum_{t=1}^{n-l_n-h} \mathbf{x}_t \varepsilon_{n,h}\right\|^3\right)^{1/3} \\ &\times \left(\mathrm{E}\left(\|\mathrm{E}\left(\mathbf{x}_n \varepsilon_{n,h} | \mathcal{F}_{n-l_n}\right)\|^3\right)\right)^{1/3} = o(1). \end{aligned} \tag{S1.15}$$

Consequently, (S1.8) follows from (S1.9), (S1.12), (S1.14) and (S1.15).

To show (S1.3), let

$$\mathbf{M}_1 = \mathbf{x}_n^\top \left(\widehat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\right) N^{-1/2} \sum_{t=1}^{N} \mathbf{x}_t \varepsilon_{t,h}, \ \mathbf{M}_2 = \mathbf{x}_n^\top \mathbf{R}^{-1} N^{-1/2} \sum_{t=1}^{N} \mathbf{x}_t \varepsilon_{t,h},$$

$$\mathbf{M}_3 = \left(N^{-1/2} \sum_{t=1}^{N} \mathbf{x}_t \varepsilon_{t,h}\right)^\top \mathbf{R}^{-1} \left(N^{-1/2} \sum_{t=1}^{N} \mathbf{x}_t \varepsilon_{n,h}\right), \ \mathbf{u}_n = N^{-1/2} \sum_{t=1}^{n-l_n-h} \mathbf{x}_t \varepsilon_{t,h}.$$

It follows that

$$\mathrm{E}(\mathbf{x}_n^\top \widehat{\mathbf{R}}^{-1} N^{-1/2} \sum_{t=1}^{N} \mathbf{x}_n \varepsilon_{t,h})^2 = \mathrm{E}(\mathbf{M}_1^2) + \mathrm{E}(\mathbf{M}_2^2) + 2\mathrm{E}(\mathbf{M}_1 \mathbf{M}_2). \tag{S1.16}$$

Moreover, by (C3), (C4), (S1.13), (S1.7) and Hölder's inequality, we have

(S1.17)
$$\mathrm{E}(\mathbf{M}_1^2) \leq (\mathrm{E}\|\mathbf{x}_n\|^{10})^{1/5}(\mathrm{E}\|\widehat{\mathbf{R}}^{-1}-\mathbf{R}^{-1}\|^5)^{2/5}(\mathrm{E}\|N^{-1/2}\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\|^5)^{2/5} = O(n^{-1}),$$

and

(S1.18)
$$\begin{aligned}\mathrm{E}(\mathbf{M}_2^2) &= \mathrm{E}\left(\mathbf{M}_3\right) + \mathrm{E}\left\{\mathbf{u}_n^\top\mathbf{R}^{-1}\left(\mathrm{E}\left(\mathbf{x}_n\mathbf{x}_n^\top|\mathcal{F}_{n-l_n}\right)-\mathbf{R}\right)\mathbf{R}^{-1}\mathbf{u}_n\right\} + O((l_n/n)^{1/2})\\ &= \mathrm{E}(\mathbf{M}_3) + O\left(\left\{\mathrm{E}\|\mathbf{u}_n\|^3\right\}^{2/3}\left\{\mathrm{E}\left\|\mathrm{E}\left(\mathbf{x}_n\mathbf{x}_n^\top|\mathcal{F}_{n-l_n}\right)-\mathbf{R}\right\|^3\right\}^{1/3}\right) + O((l_n/n)^{1/2})\\ &= \mathrm{E}(\mathbf{M}_3) + o(1).\end{aligned}$$

Consequently, (S1.3) follows from (S1.16)–(S1.18).

S2. Proof of Theorems 3.1 and 3.2. Proof of Theorem 3.1. It suffices to show (3.12) and (3.13). To show (3.12), first note that

(S2.1)
$$\begin{aligned}\hat{\sigma}_h^2(l) &= N^{-1}\sum_{t=1}^{N}(\varepsilon_{t,h}^{(l)})^2 - N^{-2}\left(\sum_{t=1}^{N}\mathbf{x}_t^{(l)}\varepsilon_{t,h}^{(l)}\right)\widehat{\mathbf{R}}^{-1}(l)\left(\sum_{t=1}^{N}\mathbf{x}_t^{(l)}\varepsilon_{t,h}^{(l)}\right)\\ &= \mathrm{E}\left((\varepsilon_{1,h}^{(l)})^2\right) + N^{-1}\sum_{t=1}^{N}\left\{(\varepsilon_{t,h}^{(l)})^2 - \mathrm{E}(\varepsilon_{1,h}^{(l)})^2\right\}\\ &\quad - N^{-2}\left(\sum_{t=1}^{N}\mathbf{x}_t^{(l)^\top}\varepsilon_{t,h}^{(l)}\right)\widehat{\mathbf{R}}^{-1}(l)\left(\sum_{t=1}^{N}\mathbf{x}_t^{(l)}\varepsilon_{t,h}^{(l)}\right).\end{aligned}$$

In addition, by (3.20) and the non-singularity of $\mathbf{R}(l)$,

$$\|\widehat{\mathbf{R}}^{-1}(l)\| = O_p(1), \tag{S2.2}$$

which, together with (3.19), yields $N^{-2}(\sum_{t=1}^{N}\mathbf{x}_t^{(l)^\top}\varepsilon_{t,h}^{(l)})\widehat{\mathbf{R}}^{-1}(l)(\sum_{t=1}^{N}\mathbf{x}_t^{(l)}\varepsilon_{t,h}^{(l)}) = O_p(n^{-1})$. Thus, (S2.1) and (3.17) in turn imply (3.12).

To show (3.13), we first dissect $\widehat{\mathbf{C}}_{h,s}(l)$ as

(S2.3)
$$\begin{aligned}\widehat{\mathbf{C}}_{h,s}(l) &= N^{-1}\sum_{t=1}^{N}\mathbf{x}_t^{(l)}\mathbf{x}_{t+s}^{(l)\top}\varepsilon_{t,h}^{(l)}\varepsilon_{t+s,h}^{(l)}\\ &- N^{-1}\sum_{t=1}^{N}\mathbf{x}_t^{(l)}\mathbf{x}_{t+s}^{(l)\top}\left(\hat{\boldsymbol{\beta}}_{n,l}(h)-\boldsymbol{\beta}_{h,l}\right)^{\top}\mathbf{x}_t^{(l)}\varepsilon_{t+s,h}^{(l)}\\ &- N^{-1}\sum_{t=1}^{N}\mathbf{x}_t^{(l)}\mathbf{x}_{t+s}^{(l)\top}\left(\hat{\boldsymbol{\beta}}_{n,l}(h)-\boldsymbol{\beta}_{h,l}\right)^{\top}\mathbf{x}_{t+s}^{(l)}\varepsilon_{t,h}^{(l)}\\ &+ N^{-1}\sum_{t=1}^{N}\mathbf{x}_t^{(l)}\mathbf{x}_{t+s}^{(l)\top}\left(\hat{\boldsymbol{\beta}}_{n,l}(h)-\boldsymbol{\beta}_{h,l}\right)^{\top}\mathbf{x}_t^{(l)}\mathbf{x}_{t+s}^{(l)\top}\left(\hat{\boldsymbol{\beta}}_{n,l}(h)-\boldsymbol{\beta}_{h,l}\right).\end{aligned}$$

By (S2.2) and (3.19)–(3.21), it can be shown that the last three terms on the right-hand side of (S2.3) are of order $o_p(1)$. Combining this with (3.18) leads to the desired conclusion (3.13). □

Proof of Theorem 3.2. For notational simplicity, we shall suppress (l) in $\mathbf{x}_t^{(l)}$, $\varepsilon_{t,h}^{(l)}$, $J_v^{(l)}$ and $\mathbf{C}_{h,s}(l)$. It follows from (3.23)–(3.25) and (3.29) that

$$y_t=\sum_{j=0}^{\infty}\mathbf{w}_{j,y}^{\top}\mathbf{v}_{t-j},\ \ s_t^{(v)}=\sum_{j=0}^{\infty}\mathbf{w}_{j,v}^{\top}\mathbf{v}_{t-j},\ \ \varepsilon_{t,h}=\sum_{j=0}^{\infty}\mathbf{w}_{j,0}^{\top}\mathbf{v}_{t+h-j},\tag{S2.4}$$

where $\mathbf{w}_{j,y}$ and $\mathbf{w}_{j,v}$ are some non-random vectors satisfying

$$\|\mathbf{w}_{j,y}\|\le c^*(j+1)^{-s},\ \ \|\mathbf{w}_{j,v}\|\le c^*(j+1)^{-s},\tag{S2.5}$$

for some $0<c^*<\infty$ and all $j\ge 0$ and $1\le v\le p$. Moreover, since

$$\mathrm{E}(\varepsilon_{t,h}y_{t-j})=0, j\in J_0 \text{ and } \mathrm{E}(\varepsilon_{t,h}s_{t-j}^{(v)})=0, j\in J_v, 1\le v\le p,\tag{S2.6}$$

it holds that

$$\begin{aligned}&\sum_{k=0}^{\infty}\mathbf{w}_{k,y}^{\top}\Lambda\mathbf{w}_{k+h+j,0}=0,\ j\in J_0,\\ &\sum_{k=0}^{\infty}\mathbf{w}_{k,v}^{\top}\Lambda\mathbf{w}_{k+h+j,0}=0,\ j\in J_v, 1\le v\le p,\end{aligned}\tag{S2.7}$$

where $\Lambda = \mathrm{E}(\mathbf{v}_t\mathbf{v}_t^\top)$.

By (3.27) and (S2.4)–(S2.6), it is not difficult to show that conditions (C3) and (C6) follow. Moreover, since (S2.5) ensures that the autocovariance functions of y_t, $s_t^{(v)}$ and $\varepsilon_{t,h}$ are square summable, by (3.27), (S2.4), (S2.6) and the First Moment Bound Theorem of Findley and Wei (1993), it can be shown that conditions (C1) and (C4) also hold true. The proof of condition (C5) is complicated and deferred to Lemma S2.1 below. The first statement of condition (C2) is obviously guaranteed by (S2.4) and the hypothesis that the fourth moments of $\{\mathbf{v}_t\}$ are independent of t, whereas the second one holds if

$$\text{(S2.8)} \qquad \text{each component of } n\mathrm{E}(\mathbf{x}_1\mathbf{x}_n^\top\varepsilon_{1,h}\varepsilon_{n,h}) \text{ converges to } 0,$$

noting that $\mathbf{x}_t = (y_{t-j}, j \in J_0, s_{t-j}^{(v)}, j \in J_v, 1 \le v \le p)^\top$.

In the following, we shall prove

$$\text{(S2.9)} \qquad \mathrm{E}(\varepsilon_{1,h}y_1\varepsilon_{n,h}y_n) = o(n^{-1})$$

instead of (S2.8) because their proofs are exactly the same. Dissect y_n and $\varepsilon_{n,h}$ as $y_n = y_n^* + \tilde{y}_n$ and $\varepsilon_{n,h} = \varepsilon_{n,h}^* + \tilde{\varepsilon}_{n,h}$, where $y_n^* = \sum_{j=0}^{n-2-h}\mathbf{w}_{j,y}^\top\mathbf{v}_{n-j}$ and $\varepsilon_{n,h}^* = \sum_{j=0}^{n-2}\mathbf{w}_{j,0}^\top\mathbf{v}_{n+h-j}$. Since $(y_n^*, \varepsilon_{n,h}^*)$ is independent of $(y_1, \varepsilon_{1,h})$, we have

(S2.10)

$$\begin{aligned}
&\mathrm{E}(\varepsilon_{1,h}y_1\varepsilon_{n,h}y_n) = \mathrm{E}(\varepsilon_{1,h}y_1\tilde{\varepsilon}_{n,h}\tilde{y}_n)\\
&= \mathrm{E}\Big\{\sum_{j_1=-\infty}^{1+h}\sum_{j_2=-\infty}^{1+h}\sum_{j_3=-\infty}^{1+h}\sum_{j_4=-\infty}^{1+h}\mathbf{w}_{1-j_1,y}^\top\mathbf{v}_{j_1}\mathbf{w}_{1+h-j_2,0}^\top\mathbf{v}_{j_2}\mathbf{w}_{n-j_3,y}^\top\mathbf{v}_{j_3}\mathbf{w}_{n+h-j_4,0}^\top\mathbf{v}_{j_4}\Big\}\\
&= \mathrm{E}\Big\{\sum_{j=-\infty}^{1+h}\mathbf{w}_{1-j,y}^\top\mathbf{v}_j\mathbf{w}_{1+h-j,0}^\top\mathbf{v}_j\mathbf{w}_{n-j,y}^\top\mathbf{v}_j\mathbf{w}_{n+h-j,0}^\top\mathbf{v}_j\Big\}\\
&+ \mathrm{E}\Big\{\sum_{\substack{-\infty<m,k\le 1+h\\ m\ne k}}\mathbf{w}_{1-m,y}^\top\mathbf{v}_m\mathbf{v}_m^\top\mathbf{w}_{1+h-m,0}\mathbf{w}_{n-k,y}^\top\mathbf{v}_k\mathbf{v}_k^\top\mathbf{w}_{n+h-k,0}\Big\}\\
&+ \mathrm{E}\Big\{\sum_{\substack{-\infty<m,k\le 1+h\\ m\ne k}}\mathbf{w}_{1-m,y}^\top\mathbf{v}_m\mathbf{v}_m^\top\mathbf{w}_{n-m,y}\mathbf{w}_{1+h-k,0}^\top\mathbf{v}_k\mathbf{v}_k^\top\mathbf{w}_{n+h-k,0}\Big\}\\
&+ \mathrm{E}\Big\{\sum_{\substack{-\infty<m,k\le 1+h\\ m\ne k}}\mathbf{w}_{1-m,y}^\top\mathbf{v}_m\mathbf{v}_m^\top\mathbf{w}_{n+h-m,0}\mathbf{w}_{1+h-k,0}^\top\mathbf{v}_k\mathbf{v}_k^\top\mathbf{w}_{n-k,y}\Big\}\\
&= \mathrm{(I)} + \mathrm{(II)} + \mathrm{(III)} + \mathrm{(IV)},
\end{aligned}$$

where $\mathbf{w}_{s,y}$ is set to $\mathbf{0}$ when $s < 0$. By (3.27) and (S2.5), it holds that

(S2.11)
$$|(\mathrm{I})| \leq \sum_{j=-\infty}^{1+h} \mathrm{E}\|\mathbf{v}_j\|^4(\|\mathbf{w}_{1-j,y}\|\|\mathbf{w}_{1+h-j,0}\|\|\mathbf{w}_{n-j,y}\|\|\mathbf{w}_{n+h-j,0}\|) = O(n^{-2s}) = o(n^{-1}).$$

Using the first relation of (S2.7) with $j = 0$, we obtain

$$\begin{aligned}(\mathrm{II}) &= \sum_{k=-\infty}^{1+h} \mathbf{w}_{n-k,y}^\top \Lambda \mathbf{w}_{n+h-k,0}\Big(\sum_{m=-\infty,m\neq k}^{1+h} \mathbf{w}_{1-m,y}^\top \Lambda \mathbf{w}_{1+h-m,0}\Big) \\ &= -\sum_{k=-\infty}^{1+h} \mathbf{w}_{n-k,y}^\top \Lambda \mathbf{w}_{n+h-k,0}\mathbf{w}_{1-k,y}^\top \Lambda \mathbf{w}_{1+h-k,0}.\end{aligned}$$

Therefore,

(S2.12)
$$|(\mathrm{II})| \leq \|\Lambda\|^2 \sum_{k=-\infty}^{1+h} \|\mathbf{w}_{n-k,y}\|\|\mathbf{w}_{n+h-k,0}\|\|\mathbf{w}_{1-k,y}\|\|\mathbf{w}_{1+h-k,0}\| = O(n^{-2s}) = o(n^{-1}).$$

Straightforward calculations and (S2.5) yield

(S2.13)
$$\begin{aligned}|(\mathrm{III})| &= O\left(\sum_{k=-\infty}^{1+h} \|\mathbf{w}_{1+h-k,0}\|\|\mathbf{w}_{n+h-k,0}\|\Big(\sum_{m=-\infty,m\neq k}^{1+h} \|\mathbf{w}_{1-m,y}\|\|\mathbf{w}_{n-m,y}\|\Big)\right) \\ &= O(n^{-4s+2}) = o(n^{-1}).\end{aligned}$$

Similarly,

(S2.14)
$$|(\mathrm{IV})| = O(n^{-4s+2}) = o(n^{-1}).$$

The desired conclusion (S2.9) (and hence (S2.8)) now follows from (S2.10)–(S2.14).

It remains to show that (3.17)–(3.21) hold true. Note first that (3.17) is an immediate consequence of (S2.4), (S2.5), and the First Moment Bound Theorem of Findley and Wei (1993). To prove (3.18), define $\mathbf{C}_{h,s}^{(t)} = [c_{h,s}^{(t)}(i,j)] = \mathbf{x}_t\mathbf{x}_{t+s}^\top \varepsilon_{t,h}\varepsilon_{t+s,h}$. Express $\mathbf{C}_{h,s}$ as $[c_{h,s}(i,j)]$ and let

$D_{h,s}^{(t)}(i,j) = c_{h,s}^{(t)}(i,j) - c_{h,s}(i,j)$. By (3.27), (S2.4), (S2.5) and tedious algebraic manipulations, we obtain for each $1 \le i,j \le S_p \equiv \sum_{v=0}^{p} \sharp(J_v)$,

$$\sup_{|m-k|=r} |\mathrm{E}(D_{h,s}^{(m)}(i,j) D_{h,s}^{(k)}(i,j))| \to 0, \text{ as } r \to \infty, \tag{S2.15}$$

yielding $n^{-1}\sum_{t=1}^{n} D_{h,s}^{(t)}(i,j) = o_p(1)$, which in turn implies (3.18). Conditions (3.19), (3.20) and (3.21) are ensured by (C4), (C1) and (C3), which have been proved previously.

LEMMA S2.1. *Assume* (3.23)–(3.25) *and* (3.28). *Then condition* (C5) *follows.*

PROOF. In view of Theorem 2.1 of Chan and Ing (2011), it suffices for (C5) to show that (2.19) holds true. By (S2.4), there exist $S_p \times (p+1)$ matrices H_j, $j \ge 0$, with $\|H_j\| \le \bar{c}(j+1)^{-s}$ for some $0 < \bar{c} < \infty$, such that $\mathbf{x}_t = \sum_{j=0}^{\infty} H_j \mathbf{v}_{t-j}$. Moreover, by (S2.4), the independence between $\{\boldsymbol{\delta}_t(p)\}$ and $\{\epsilon_t\}$, and the positive definiteness of Σ_p, it can be shown that $\lambda_{\min}(\mathrm{E}(\mathbf{x}_t\mathbf{x}_t^\top)) > \delta_0$ for some positive constant δ_0. These facts yield that for a given $\delta_1^* < \delta_0$, there exists a positive integer D such that for all $t \ge D$,

$$\lambda_{\min}\left(\mathrm{E}(\mathbf{x}_{t,D}\mathbf{x}_{t,D}^\top)\right) > \delta_1^*, \tag{S2.16}$$

where $\mathbf{x}_{t,D} = \sum_{j=0}^{D-1} H_j \mathbf{v}_{t-j}$. Since (S2.16) ensures that $\mathrm{E}(\boldsymbol{s}^\top \mathbf{x}_{t,D})^2 = \sum_{j=0}^{D-1} \boldsymbol{s}^\top H_j \Lambda H_j^\top \boldsymbol{s} \ge \delta_1^*$ for any $\|\boldsymbol{s}\| = 1$, there is an integer $0 \le j(\boldsymbol{s}) \le D-1$ such that

$$\boldsymbol{s}^\top H_{j(\boldsymbol{s})} \Lambda H_{j(\boldsymbol{s})}^\top \boldsymbol{s} \ge \delta_1^*/D. \tag{S2.17}$$

Define $\mathcal{F}_{t,j(\boldsymbol{s})} = \{\mathbf{v}_t, \ldots, \mathbf{v}_{t-j(\boldsymbol{s})+1}, \mathbf{v}_{t-j(\boldsymbol{s})-1}, \ldots\}$ and $\eta_{j(\boldsymbol{s})} = \boldsymbol{s}^\top H_{j(\boldsymbol{s})} H_{j(\boldsymbol{s})}^\top \boldsymbol{s}$. Then, by (3.28) and (S2.17), it follows that for any $\|\boldsymbol{s}\| = 1$, any $t \ge D$,

and any $0 < s_2 - s_1 \le \delta_1 \sqrt{\delta_1^*/D}\lambda_{\max}^{-1/2}(\Lambda)$, where δ_1 is defined in (3.28),

$$
\begin{aligned}
&P\big(s_1 < \boldsymbol{s}^\top \mathbf{x}_t \le s_2 | \mathcal{F}_{t-D}\big) \\
=&\mathrm{E}\left\{P\left(s_1 < \boldsymbol{s}^\top \mathbf{x}_t \le s_2 \big| \mathcal{F}_{t,j(\boldsymbol{s})}\right) \big| \mathcal{F}_{t-D}\right\} \\
=&\mathrm{E}\left\{ P\left(\frac{s_1 - \sum_{\substack{j=0 \\ j\neq j(\boldsymbol{s})}}^{\infty} \boldsymbol{s}^\top H_j \mathbf{v}_{t-j}}{\sqrt{\eta_{j(\boldsymbol{s})}}} < \frac{\boldsymbol{s}^\top H_{j(\boldsymbol{s})} \mathbf{v}_{t-j(\boldsymbol{s})}}{\sqrt{\eta_{j(\boldsymbol{s})}}} \le \frac{s_2 - \sum_{\substack{j=0 \\ j\neq j(\boldsymbol{s})}}^{\infty} \boldsymbol{s}^\top H_j \mathbf{v}_{t-j}}{\sqrt{\eta_{j(\boldsymbol{s})}}} \Big| \mathcal{F}_{t,j(\boldsymbol{s})} \right) \Big| \mathcal{F}_{t-D} \right\} \\
\le& K_1 \left(\sqrt{\frac{D\lambda_{\max}(\Lambda)}{\delta_1^*}} (s_2 - s_1) \right)^v \quad \text{almost surely,}
\end{aligned}
$$

recalling that $\mathcal{F}_t = \sigma(\mathbf{v}_t, \mathbf{v}_{t-1}, \ldots)$. Consequently, (2.19) holds with $\mathcal{F}_t = \sigma(\mathbf{v}_t, \mathbf{v}_{t-1}, \ldots)$, $\alpha = v$, $\delta = \delta_1 \sqrt{\delta_1^*/D}\lambda_{\max}^{-1/2}(\Lambda)$, $M = K_1 (D\lambda_{\max}(\Lambda)/\delta_1^*)^{v/2}$, and D given above.

S3. Proofs of Theorems 4.1–4.3. We first define a 'noiseless' OGA with parameter $0 < \xi \le 1$. The algorithm is initialized by setting $\tilde{J}_{0,\xi} = \emptyset$ and $\mathbf{f}_\xi^{(0)} \equiv \boldsymbol{u}_h = (\boldsymbol{\beta}_h^\top \mathbf{x}_t, t = 1, \ldots, N)^\top$. For $1 \le m \le p$, the algorithm updates $\tilde{J}_{m,\xi}$ and $\mathbf{f}_\xi^{(m)}$ recursively as follows:

$$
\begin{aligned}
\tilde{J}_{m,\xi} &= \tilde{J}_{m-1,\xi} \bigcup \{\tilde{j}_{m,\xi}\}, \\
\mathbf{f}_\xi^{(m)} &= (\mathbf{I}_N - \mathbf{H}_{\tilde{J}_{m,\xi}}) \boldsymbol{u}_h,
\end{aligned}
$$

where $\tilde{j}_{m,\xi}$ is any $l \in \{1, \ldots, p\} - \tilde{J}_{m-1,\xi}$ such that

$$
|\boldsymbol{\nu}_{\tilde{J}_{m-1,\xi},l}| \ge \xi \max_{1\le j\le p, j\notin \tilde{J}_{m-1,\xi}} |\boldsymbol{\nu}_{\tilde{J}_{m-1,\xi},j}|, \tag{S3.1}
$$

with $\boldsymbol{\nu}_{J,i} = \mathbf{X}_i^\top (\mathbf{I}_N - \mathbf{H}_J)\boldsymbol{u}_h / (N^{1/2}\|\mathbf{X}_i\|)$. Although the noiseless OGA cannot be implemented in practice, the rate of convergence of $N^{-1}\|\mathbf{f}_\xi^{(m)}\|^2$ plays an important role in our theoretical analysis. The next lemma provides a uniform bound for $N^{-1}\|\mathbf{f}_\xi^{(m)}\|^2$.

LEMMA S3.1. *Assume* (F4). *Then,*

$$
\max_{1\le m\le p} \frac{N^{-1}\|\mathbf{f}_\xi^{(m)}\|^2}{\frac{1}{1+m\xi^2}} = O_p(1). \tag{S3.2}
$$

PROOF. Note that

$$
\begin{aligned}
&N^{-1}\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m,\xi}})\boldsymbol{u}_h\|^2\\
&\le N^{-1}\Big\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{u}_h-\frac{\mathbf{X}_{\tilde{J}_{m,\xi}}^{\top}(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{u}_h}{\|\mathbf{X}_{\tilde{J}_{m,\xi}}\|^2}\mathbf{X}_{\tilde{J}_{m,\xi}}\Big\|^2\\
&= N^{-1}\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{u}_h\|^2-\boldsymbol{\nu}^2_{\tilde{J}_{m-1,\xi},\tilde{j}_{m,\xi}}\\
&\le N^{-1}\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{u}_h\|^2-\xi^2\max_{1\le j\le p, j\notin\tilde{J}_{m-1,\xi}}\boldsymbol{\nu}^2_{\tilde{J}_{m-1,\xi},j}.
\end{aligned}
\tag{S3.3}
$$

We also have

$$
\begin{aligned}
N^{-1}\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{u}_h\|^2&=\sum_{j=1,j\notin\tilde{J}_{m-1,\xi}}^{p}\beta^*_{j,h}\hat{\rho}^{1/2}_{j,j}\boldsymbol{\nu}_{\tilde{J}_{m-1,\xi},j}\\
&\le\max_{1\le j\le p,j\notin\tilde{J}_{m-1,\xi}}|\boldsymbol{\nu}_{\tilde{J}_{m-1,\xi},j}|\sum_{j=1}^{p}|\beta^*_{j,h}|\hat{\rho}^{1/2}_{j,j},
\end{aligned}
\tag{S3.4}
$$

where $\hat{\rho}_{i,j}=N^{-1}\sum_{t=1}^{N}z_{t,i}z_{t,j}$. Let $S_n=(\sum_{j=1}^{p}|\beta^*_{j,h}|\hat{\rho}^{1/2}_{j,j})^2$. By (S3.3), (S3.4), and some algebraic manipulations, one obtains

$$
\begin{aligned}
&N^{-1}\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{0,\xi}})\boldsymbol{u}_h\|^2=N^{-1}\|\boldsymbol{u}_h\|^2\le S_n,\\
&N^{-1}\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m,\xi}})\boldsymbol{u}_h\|^2\\
&\le N^{-1}\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{u}_h\|^2\Big\{1-\frac{\xi^2N^{-1}\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m-1,\xi}})\boldsymbol{u}_h\|^2}{S_n}\Big\},
\end{aligned}
\tag{S3.5}
$$

for each $1\le m\le p$. By (S3.5) and Lemma 3.1 of Temlyakov (2000), it holds that

$$
\max_{1\le m\le p}\frac{N^{-1}\|(\mathbf{I}_N-\mathbf{H}_{\tilde{J}_{m,\xi}})\boldsymbol{u}_h\|^2}{\frac{1}{1+m\xi^2}}\le S_n.
\tag{S3.6}
$$

Moreover, since (F4) and Minkowski's inequality yield

$$
\mathrm{E}(S_n)\le[\sum_{j=1}^{p}|\beta^*_{j,h}|\{\mathrm{E}(\hat{\rho}_{j,j})\}^{1/2}]^2\le G_1^2,
\tag{S3.7}
$$

the desired conclusion (S3.2) follows from (S3.7) and (S3.6).

The following lemma shows that when the $\mathbf{H}_{\tilde{J}_{m,\xi}}$ in $\mathbf{f}_{\xi}^{(m)}$ is replaced by $\mathbf{H}_{\hat{J}_m}$, a similar uniform bound can be obtained over a narrower range of m.

LEMMA S3.2. *Assume* (F1)–(F5). *Suppose* $K_n = \min\{p, \bar{l}_n\}$, *where*

$$\bar{l}_n \to \infty \text{ and } \bar{l}_n^{1+\max\{\theta_1,\theta_2\}} = o(n^{1/2}/p^{1/q}). \tag{S3.8}$$

Then,

$$\max_{1\le m\le K_n} \frac{N^{-1}\|(\mathbf{I}_N - \mathbf{H}_{\hat{J}_m})\boldsymbol{u}_h\|^2}{m^{-1}} = O_p(1). \tag{S3.9}$$

PROOF. By (F1) and (F2), it is not difficult to see that

$$\max_{1\le i\le p} |N^{-1}\sum_{t=1}^{N} z_{t,i}\varepsilon_{t,h}| = O_p(p^{1/q_2}/n^{1/2}), \tag{S3.10}$$

$$\max_{\sharp(J)\le K_n} \|N^{-1}\sum_{t=1}^{N} \mathbf{z}_t(J)\varepsilon_{t,h}\| = O_p(K_n^{1/2}p^{1/q_2}/n^{1/2}), \tag{S3.11}$$

$$\max_{1\le i,j\le p} |\hat{\rho}_{i,j} - \rho_{i,j}| = O_p(p^{1/q_1}/n^{1/2}), \tag{S3.12}$$

and

$$\max_{1\le i\le p}(N^{-1}\sum_{t=1}^{N} z_{t,i}^2)^{-1} = O_p(1). \tag{S3.13}$$

Equation (S3.12), the first relation of (F5), and

$$K_n^{1+\max\{\theta_1,\theta_2\}} = o\big(\frac{n^{1/2}}{p^{1/q}}\big) \text{ (which is ensured by (S3.8))} \tag{S3.14}$$

further imply

$$\max_{\sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}^{-1}(J) - \mathbf{\Gamma}^{-1}(J)\| = o_p(K_n^{\theta_1}), \ \max_{\sharp(J)\le K_n} \|\hat{\mathbf{\Gamma}}^{-1}(J)\| = O_p(K_n^{\theta_1}), \tag{S3.15}$$

where $\hat{\mathbf{\Gamma}}(J) = (\hat{\rho}_{i,j})_{i,j\in J}$. Recall the definition of $\hat{\boldsymbol{\mu}}_{J,i}$ in Section 4.1. In the following, we shall prove

$$\max_{\sharp(J)\le K_n-1, i\notin J} |\hat{\boldsymbol{\mu}}_{J,i} - \boldsymbol{\nu}_{J,i}| = O_p\big(\frac{K_n^{\theta_2} p^{1/q_2}}{n^{1/2}}\big). \tag{S3.16}$$

First observe that

(S3.17)

$$\begin{aligned}
\max_{\sharp(J)\le K_n-1, i\notin J} |\hat{\boldsymbol{\mu}}_{J,i} - \boldsymbol{\nu}_{J,i}| &\le \max_{1\le i\le p} (N^{-1}\sum_{t=1}^N z_{t,i}^2)^{-1/2} \Big\{ \max_{1\le i\le p} |N^{-1}\sum_{t=1}^N z_{t,i}\varepsilon_{t,h}| \\
&+ \max_{\sharp(J)\le K_n-1, i\notin J} \|N^{-1}\sum_{t=1}^N \mathbf{z}_t(J) z_{t,i;J}^{\perp}\| \|\hat{\mathbf{\Gamma}}^{-1}(J)\| \|N^{-1}\sum_{t=1}^N \mathbf{z}_t(J)\varepsilon_{t,h}\| \\
&+ \max_{\sharp(J)\le K_n-1, i\notin J} |N^{-1}\sum_{t=1}^N \mathbf{g}_i^{\top}(J)\mathbf{\Gamma}^{-1}(J)\mathbf{z}_t(J)\varepsilon_{t,h}| \Big\},
\end{aligned}$$

where $z_{t,i;J}^{\perp} = z_{t,i} - \mathbf{g}_i^{\top}(J)\mathbf{\Gamma}^{-1}(J)\mathbf{z}_t(J)$. It follows from (S3.10) and the second relation of (F5) that

(S3.18)

$$\begin{aligned}
&\max_{\sharp(J)\le K_n-1, i\notin J} |N^{-1}\sum_{t=1}^N \mathbf{g}_i^{\top}(J)\mathbf{\Gamma}^{-1}(J)\mathbf{z}_t(J)\varepsilon_{t,h}| \\
&\le \max_{1\le i\le p} |N^{-1}\sum_{t=1}^N z_{t,i}\varepsilon_{t,h}| \max_{\sharp(J)\le K_n-1, i\notin J} \|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1 = O_p\left(\frac{K_n^{\theta_2} p^{1/q_2}}{n^{1/2}}\right).
\end{aligned}$$

In addition, (S3.12) and the second relation of (F5) imply

$$\begin{aligned}
&\max_{\sharp(J)\le K_n-1, i\notin J} \|N^{-1}\sum_{t=1}^N \mathbf{z}_t(J) z_{t,i;J}^{\perp}\| \\
&\le K_n^{1/2} \max_{1\le i,j\le p} |\hat{\rho}_{i,j} - \rho_{i,j}| \big(1 + \max_{\sharp(J)\le K_n-1, i\notin J} \|\mathbf{\Gamma}^{-1}(J)\mathbf{g}_i(J)\|_1\big) \\
&= O_p\left(\frac{K_n^{\theta_2+(1/2)} p^{1/q_1}}{n^{1/2}}\right).
\end{aligned} \tag{S3.19}$$

Combining (S3.10), (S3.11), (S3.13)–(S3.15), and (S3.17)–(S3.19) yields

$$\max_{\sharp(J)\le K_n-1, i\notin J} |\hat{\boldsymbol{\mu}}_{J,i} - \boldsymbol{\nu}_{J,i}| = O_p\left(\frac{p^{1/q_2}}{n^{1/2}} + \frac{K_n^{1+\theta_1+\theta_2} p^{1/q_1+1/q_2}}{n} + \frac{K_n^{\theta_2} p^{1/q_2}}{n^{1/2}}\right)$$
$$= O_p\left(\frac{K_n^{\theta_2} p^{1/q_2}}{n^{1/2}}\right)$$

and hence (S3.16) follows.

Equation (S3.16) ensures that for any small $\epsilon > 0$, there exists a large constant V_ϵ for which

$$P\big(\max_{\sharp(J)\le K_n-1, i\notin J} |\hat{\boldsymbol{\mu}}_{J,i} - \boldsymbol{\nu}_{J,i}| > V_\epsilon K_n^{\theta_2} p^{1/q_2}/n^{1/2}\big) \le \epsilon. \tag{S3.20}$$

For $1 \le m \le K_n$, define

$$\mathcal{A}_m = \{\max_{\sharp(J)\le m-1, i\notin J} |\hat{\boldsymbol{\mu}}_{J,i} - \boldsymbol{\nu}_{J,i}| \le V_\epsilon K_n^{\theta_2} p^{1/q_2}/n^{1/2}\},$$
$$\mathcal{B}_m = \{\min_{0\le l\le m-1} \max_{1\le i\le p, i\notin \hat{J}_l} |\boldsymbol{\nu}_{\hat{J}_l,i}| > \{2/(1-\xi)\} V_\epsilon K_n^{\theta_2} p^{1/q_2}/n^{1/2}\},$$

where $0 < \xi < 1$. Then, on the set $\mathcal{A}_m \bigcap \mathcal{B}_m$, we have for any $1 \le l \le m$,

(S3.21)
$$|\boldsymbol{\nu}_{\hat{J}_{l-1},\hat{j}_l}| \ge -\max_{\sharp(J)\le m-1, i\notin J} |\hat{\boldsymbol{\mu}}_{J,i} - \boldsymbol{\nu}_{J,i}| + \max_{1\le i\le p, i\notin \hat{J}_{l-1}} |\hat{\boldsymbol{\mu}}_{\hat{J}_{l-1},i}|$$
$$\ge -2\max_{\sharp(J)\le m-1, i\notin J} |\hat{\boldsymbol{\mu}}_{J,i} - \boldsymbol{\nu}_{J,i}| + \max_{1\le i\le p, i\notin \hat{J}_{l-1}} |\boldsymbol{\nu}_{\hat{J}_{l-1},i}| \ge \xi \max_{1\le i\le p, i\notin \hat{J}_{l-1}} |\boldsymbol{\nu}_{\hat{J}_{l-1},i}|.$$

Equation (S3.21) indicates that on the set $\mathcal{A}_m \bigcap \mathcal{B}_m$, the update rule of OGA obeys (S3.1), and hence by (S3.6),

$$\frac{N^{-1}\|(\mathbf{I}_N - \mathbf{H}_{\hat{J}_m})\boldsymbol{u}_h\|^2}{\frac{1}{1+m\xi^2}} \le S_n \text{ on } \mathcal{A}_m \bigcap \mathcal{B}_m. \tag{S3.22}$$

Moreover, we have

$$N^{-1}\|(\mathbf{I}_N - \mathbf{H}_{\hat{J}_m})\boldsymbol{u}_h\|^2 \le \min_{0\le i\le m-1} N^{-1}\|(\mathbf{I}_N - \mathbf{H}_{\hat{J}_i})\boldsymbol{u}_h\|^2$$
$$\le \min_{0\le i\le m-1} \sum_{j=1, j\notin \hat{J}_i}^{p} |\beta^*_{j,h}||\boldsymbol{\nu}_{\hat{J}_i,j}|\hat{\rho}_{j,j}^{1/2} \tag{S3.23}$$
$$\le \min_{0\le i\le m-1} \max_{1\le j\le p, j\notin \hat{J}_i} |\boldsymbol{\nu}_{\hat{J}_i,j}| S_n^{1/2} \le \frac{2V_\epsilon K_n^{\theta_2} p^{1/q_2}}{(1-\xi)n^{1/2}} S_n^{1/2} \text{ on } \mathcal{B}_m^c.$$

It follows from (S3.14), (S3.22), (S3.23), and $\mathcal{A}_{K_n} \subseteq \mathcal{A}_m$ for $1 \le m \le K_n$ that

$$\max_{1\le m\le K_n} \frac{N^{-1}\|(\mathbf{I}_N - \mathbf{H}_{\hat{J}_m})\boldsymbol{u}_h\|^2}{m^{-1}} \le V^*(S_n + S_n^{1/2}) \text{ on } \mathcal{A}_{K_n}, \tag{S3.24}$$

where V^* is some positive constant depending only on V_ϵ and ξ. Now (S3.9) is an immediate consequence of (S3.24), (S3.20) and (S3.7).

We are now in a position to prove Theorems 4.1–4.3.

PROOF OF THEOREM 4.1. When p is bounded above by a finite constant, the proof of Theorem 4.1 is obvious because $K_n = p$ for all large n. Therefore, we only focus on the case where $p \to \infty$. Let $\underline{K}_n = \min\{p, \bar{m}_n, \bar{l}_n\}$, where $\bar{l}_n$ is given in (S3.8). It follows from (S3.8), (4.5), (F4), $0 \le \theta_1 < 1$, and $\underline{K}_n \to \infty$ that

$$(\underline{K}_n + \sharp(N_h))p^{1/q_1}/n^{1/2} = o\big((\underline{K}_n + \sharp(N_h))^{-\theta_1}\big) \text{ and } \underline{K}_n^{-1} = o\big((\underline{K}_n + \sharp(N_h))^{-\theta_1}\big). \tag{S3.25}$$

By Lemma S3.2,

$$N^{-1}\|(\mathbf{I}_N - \mathbf{H}_{\hat{J}_{\underline{K}_n}})\boldsymbol{u}_h\|^2 = O_p(\underline{K}_n^{-1}). \tag{S3.26}$$

Straightforward calculations yield that on $Z_n = \{N_h - \hat{J}_{\underline{K}_n} \ne \emptyset\}$,

$$\begin{aligned} &N^{-1}\|(\mathbf{I}_N - \mathbf{H}_{\hat{J}_{\underline{K}_n}})\boldsymbol{u}_h\|^2 \\ &\ge \{\min_{\sharp(J)\le \underline{K}_n+\sharp(N_h)} \lambda_{\min}(\boldsymbol{\Gamma}(J)) - \max_{\sharp(J)\le \underline{K}_n+\sharp(N_h)} \|\hat{\boldsymbol{\Gamma}}(J) - \boldsymbol{\Gamma}(J)\|\}\underline{\delta}^2. \end{aligned} \tag{S3.27}$$

Moreover, (F1) implies

$$\max_{\sharp(J)\le \underline{K}_n+\sharp(N_h)} \|\hat{\boldsymbol{\Gamma}}(J) - \boldsymbol{\Gamma}(J)\| = O_p\big((\underline{K}_n + \sharp(N_h))p^{1/q_1}/n^{1/2}\big). \tag{S3.28}$$

Set $S_{1n} = \{\max_{\sharp(J)\le \underline{K}_n+\sharp(N_h)} \|\hat{\boldsymbol{\Gamma}}(J) - \boldsymbol{\Gamma}(J)\| \ge c_1(\underline{K}_n + \sharp(N_h))^{-\theta_1}/2\}$. Then, it follows from (S3.25)–(S3.28) and the first relation of (F5) that

$$P(Z_n) \le P\big(N^{-1}\|(\mathbf{I}_N - \mathbf{H}_{\hat{J}_{\underline{K}_n}})\boldsymbol{u}_h\|^2 \ge c_1(\underline{K}_n + \sharp(N_h))^{-\theta_1}\underline{\delta}^2/2\big) + P(S_{1n}) = o(1),$$

yielding $1 + o(1) = P(Z_n^c) \le P(N_h \subseteq \hat{J}_{\underline{K}_n}) \le P(N_h \subseteq \hat{J}_{K_n})$.

PROOF OF THEOREM 4.2. Define $\tilde{k}_n = \min\{1 \le k \le K_n : N_h \subseteq \hat{J}_k\}$ and $K_n + 1$ if $N_h - \hat{J}_{K_n} \neq \emptyset$. In the following, we shall show that

$$\lim_{n\to\infty} P(\hat{k}_n = \tilde{k}_n) = 1. \tag{S3.29}$$

Let $E_n = \{N_h \subseteq \hat{J}_{K_n}\}$. In view of Theorem 4.1,

$$\lim_{n\to\infty} P(E_n) = 1, \tag{S3.30}$$

and hence (S3.29) is guaranteed by

$$\begin{aligned} P(\hat{k}_n < \tilde{k}_n, E_n) &= o(1), \\ P(\hat{k}_n > \tilde{k}_n, E_n) &= o(1). \end{aligned} \tag{S3.31}$$

To show the first identity of (S3.31), we note that

(S3.32)

$$\hat{z}_n(\hat{J}_{\tilde{k}_n}) \le \hat{\sigma}_h^2(\hat{J}_{\tilde{k}_n-1}) - \hat{\sigma}_h^2(\hat{J}_{\tilde{k}_n}) \le \frac{(\tilde{k}_n - \hat{k}_n)p^{2/q}\omega_n}{n}\hat{\sigma}_h^2(\hat{J}_{\tilde{k}_n}) \text{ on } \{\hat{k}_n < \tilde{k}_n\} \bigcap E_n,$$

where

$$\hat{z}_n(\hat{J}_{\tilde{k}_n}) = \underline{\delta}^2\lambda_{\min}(\hat{\mathbf{\Gamma}}(\hat{J}_{\tilde{k}_n})) - 2G_1 \left\{ |N^{-1}\sum_{t=1}^{N} z_{t,\hat{j}_{\tilde{k}_n}}\varepsilon_{t,h}| + \|\hat{\mathbf{\Gamma}}(\hat{J}_{\tilde{k}_n-1})\| \|N^{-1}\sum_{t=1}^{N}\mathbf{z}_t(\hat{J}_{\tilde{k}_n-1})\varepsilon_{t,h}\| \right.$$
$$\left. \|N^{-1}\sum_{t=1}^{N}\mathbf{z}_t(\hat{J}_{\tilde{k}_n-1})z^{\perp}_{t,\hat{j}_{\tilde{k}_n};\hat{J}_{\tilde{k}_n-1}}\| + |N^{-1}\sum_{t=1}^{N}\mathbf{g}^{\top}_{\hat{j}_{\tilde{k}_n}}(\hat{J}_{\tilde{k}_n-1})\mathbf{\Gamma}^{-1}(\hat{J}_{\tilde{k}_n-1})\mathbf{z}_t(\hat{J}_{\tilde{k}_n-1})\varepsilon_{t,h}| \right\}.$$

By an argument similar to that used to prove Lemma S3.2, it can be shown that there exists $c_3 > 0$ such that

(S3.33)

$$\lim_{n\to\infty} P(\hat{z}_n(\hat{J}_{\tilde{k}_n}) > c_3 K_n^{-\theta_1}) = 1 \text{ and } \frac{(\tilde{k}_n - \hat{k}_n)p^{2/q}\omega_n}{n}\hat{\sigma}^2_{\hat{J}_{\tilde{k}_n}} = O_p\left(\frac{K_n p^{2/q}\omega_n}{n}\right).$$

It follows from (4.11), (4.12), (S3.32), and (S3.33) that

$$P(\hat{k}_n < \tilde{k}_n, E_n) \le P\big(\hat{z}_n(\hat{J}_{\tilde{k}_n}) \le (\tilde{k}_n - \hat{k}_n)p^{2/q}\omega_n\hat{\sigma}^2_{\hat{J}_{\tilde{k}_n}}/n\big) = o(1),$$

and hence the first identity of (S3.31) holds true.

Next we prove the second identity of (S3.31). To this end, note that

(S3.34)
$$(1+\frac{\hat{k}_n\omega_n p^{2/q}}{n})(\hat{\sigma}_h^2(\hat{J}_{\tilde{k}_n})-\hat{\sigma}_h^2(\hat{J}_{\hat{k}_n}))\geq\frac{(\hat{k}_n-\tilde{k}_n)\omega_n p^{2/q}}{n}\hat{\sigma}_h^2(\hat{J}_{\tilde{k}_n}) \text{ on } \{\tilde{k}_n<\hat{k}_n\}\bigcap E_n.$$

Since $\hat{k}_n\omega_n p^{2/q}/n\leq K_n\omega_n p^{2/q}/n=o(1)$,

$$P(E_n^*)\equiv P(\hat{k}_n\omega_n p^{2/q}/n<1/2)=1+o(1), \tag{S3.35}$$

which, together with (S3.34), yields

(S3.36)
$$P(\hat{k}_n>\tilde{k}_n,E_n,E_n^*)\leq P\left(\frac{3}{2}(\hat{\sigma}_h^2(\hat{J}_{\tilde{k}_n})-\hat{\sigma}_h^2(\hat{J}_{\hat{k}_n}))\geq\frac{(\hat{k}_n-\tilde{k}_n)\omega_n p^{2/q}}{n}\hat{\sigma}^2_{\hat{J}_{\tilde{k}_n}}\right).$$

Straightforward calculations imply

$$\hat{\sigma}_h^2(\hat{J}_{\tilde{k}_n})-\hat{\sigma}_h^2(\hat{J}_{\hat{k}_n})\leq 2(\hat{k}_n-\tilde{k}_n)(\hat{\gamma}_{1n}+\hat{\gamma}_{2n}), \tag{S3.37}$$

where $\hat{\gamma}_{1n}=\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n})\|\max_{1\leq j\leq p}(N^{-1}\sum_{t=1}^N z_{t,j}\varepsilon_{t,h})^2$ and

$$\hat{\gamma}_{2n}=\|\hat{\mathbf{\Gamma}}^{-1}(\hat{J}_{K_n})\|\left\{\max_{\sharp(J)\leq\tilde{k}_n,i\notin J}\|\hat{\mathbf{\Gamma}}^{-1}(J)\|\|N^{-1}\sum_{t=1}^N\mathbf{z}_t(J)\varepsilon_{t,h}\|\right.$$
$$\left.\|N^{-1}\sum_{t=1}^N\mathbf{z}_t(J)z_{t,i;J}^{\perp}\|+\max_{\sharp(J)\leq\tilde{k}_n,i\notin J}|N^{-1}\sum_{t=1}^N\mathbf{g}_i^{\top}(J)\mathbf{\Gamma}^{-1}(J)\mathbf{z}_t(J)\varepsilon_{t,h}|\right\}^2.$$

By (4.10) and an argument similar to that used to prove Lemma S3.2, we obtain

(S3.38)
$$\hat{\gamma}_{1n}=O_p\left(\frac{K_n^{\theta_1}p^{2/q}}{n}\right),$$
$$\hat{\gamma}_{2n}=O_p\left(K_n^{\theta_1}(\frac{K_n^{1+\theta_1+\theta_2}p^{1/q_1+1/q_2}}{n}+\frac{K_n^{\theta_2}p^{1/q_2}}{n^{1/2}})^2\right)=O_p\left(\frac{K_n^{\theta_1+2\theta_2}p^{2/q}}{n}\right),$$
$$\frac{(\hat{k}_n-\tilde{k}_n)\omega_n p^{2/q}}{n}\hat{\sigma}^2_{\hat{J}_{\tilde{k}_n}}=\frac{(\hat{k}_n-\tilde{k}_n)\omega_n p^{2/q}}{n}(\mathrm{E}(\varepsilon_{1,h}^2)+o_p(1)).$$

Now, the second identity of (S3.31) follows from (4.12) and (S3.34)–(S3.38). This completes the proof of (S3.29).

With the help of (S3.29) and (S3.30), we have

(S3.39)
$$
\begin{aligned}
&P(\hat{N}_h \neq N_h) \leq P(\hat{N}_h \neq N_h, \hat{k}_n > 1, N_h \subseteq \hat{J}_{\hat{k}_n}) + P(\hat{N}_h \neq N_h, \hat{k}_n = 1) + P(N_h \nsubseteq \hat{J}_{\hat{k}_n}) \\
&\leq P(\mathrm{HDIC}(\hat{J}_{\tilde{k}_n} - \{\hat{j}_l\}) > \mathrm{HDIC}(\hat{J}_{\tilde{k}_n}) \text{ and } \beta_{\hat{j}_l,h} = 0 \text{ for some } 1 \leq l \leq \tilde{k}_n, \tilde{k}_n > 1, N_h \subseteq \hat{J}_{\tilde{k}_n}) \\
&+ P(\mathrm{HDIC}(\hat{J}_{\tilde{k}_n} - \{\hat{j}_l\}) < \mathrm{HDIC}(\hat{J}_{\tilde{k}_n}) \text{ and } \beta_{\hat{j}_l,h} \neq 0 \text{ for some } 1 \leq l \leq \tilde{k}_n, \tilde{k}_n > 1, N_h \subseteq \hat{J}_{\tilde{k}_n}) \\
&+ P(\hat{k}_n \neq \tilde{k}_n) + P(\hat{N}_h \neq N_h, \hat{k}_n = 1) + P(N_h \nsubseteq \hat{J}_{\hat{k}_n}) \\
&= \mathrm{(I)} + \mathrm{(II)} + \mathrm{(III)} + \mathrm{(IV)} + \mathrm{(V)} = \mathrm{(I)} + \mathrm{(II)} + o(1).
\end{aligned}
$$

In the same way as in the proof of (S3.29), we obtain

$$
\mathrm{(I)} = o(1) \text{ and } \mathrm{(II)} = o(1).
$$

Combining this with (S3.39) gives the desired conclusion (4.13).

PROOF OF THEOREM 4.3. By an argument similar to that used to prove (3.14), it holds that

(S3.40)
$$
\mathrm{MRIC}_{h,l}(N_h^{(l)}) = \mathrm{MI}_{h,l}(N_h^{(l)}) + O_p(n^{-1/2}) + \frac{C_n}{n} L_{h,l}(N_h^{(l)}) + o_p\left(\frac{C_n}{n}\right).
$$

It follows from Theorem 4.2 that for each $l = 1, \ldots, K$,

$$
\lim_{n\to\infty} P(\hat{N}_h^{(l)} = N_h^{(l)}) = 1. \tag{S3.41}
$$

Let $l_1 \in M_{A,h}$, $l_2 \in \{1, \ldots, K\} - M_{A,h} \neq \emptyset$, $l_3 \in M_{B,h}$, and $l_4 \in M_{A,h} - M_{B,h} \neq \emptyset$. Then (S3.40) and (S3.41) imply

(S3.42)
$$
\begin{aligned}
&P(\mathrm{MRIC}_{h,l}(\hat{N}_h^{(l_1)}) \geq \mathrm{MRIC}_{h,l}(\hat{N}_h^{(l_2)})) \leq P(\mathrm{MRIC}_{h,l}(N_h^{(l_1)}) \geq \mathrm{MRIC}_{h,l}(N_h^{(l_2)})) \\
&+ P(\hat{N}_h^{(l_1)} \neq N_h^{(l_1)} \text{ or } \hat{N}_h^{(l_2)} \neq N_h^{(l_2)}) = o(1),
\end{aligned}
$$

and

(S3.43)
$$
\begin{aligned}
&P(\mathrm{MRIC}_{h,l}(\hat{N}_h^{(l_3)}) \geq \mathrm{MRIC}_{h,l}(\hat{N}_h^{(l_4)})) \leq P(\mathrm{MRIC}_{h,l}(N_h^{(l_3)}) \geq \mathrm{MRIC}_{h,l}(N_h^{(l_4)})) \\
&+ P(\hat{N}_h^{(l_3)} \neq N_h^{(l_3)} \text{ or } \hat{N}_h^{(l_4)} \neq N_h^{(l_4)}) = o(1).
\end{aligned}
$$

Equations (S3.42) and (S3.43) yield

$$\lim_{n\to\infty} P(\hat{l}_h \in M_{B,h}) = 1. \tag{S3.44}$$

Now, by (S3.41) and (S3.44), one obtains

$$\begin{aligned} & P((\hat{l}_h, \hat{N}_h^{(\hat{l}_h)}) \notin M_{C,h}) \le P((\hat{l}_h, N_h^{(\hat{l}_h)}) \notin M_{C,h}) + o(1) \\ \le\ & P((\hat{l}_h, N_h^{(\hat{l}_h)}) \notin M_{C,h}, \hat{l}_h \in M_{B,h}) + o(1) = o(1), \end{aligned}$$

yielding the desired conclusion (4.24).

S4. A Nonlinear Extension.

S4.1. *A Nonlinear Extension of MRIC and Its Asymptotic Efficiency.* In this section, we generalize the results obtained in Sections 2 and 3 of Hsu et al. (2018) to nonlinear cases. Let $\{\mathcal{F}_t\}$ be an increasing sequence of sub-σ-fields of $\mathcal{F}$. We consider an h-step predictive model, $g_{t,h}(\boldsymbol{\theta})$, of y_{t+h}, where $g_{t,h}(\boldsymbol{\theta})$ is specified up to the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$ and is $\mathcal{F}_t$-measurable for each $\boldsymbol{\theta} \in \Theta$, with Θ denoting a compact parameter space in R^m. Assume that $V(\boldsymbol{\theta}) \equiv \mathrm{E}(y_{t+h} - g_{t,h}(\boldsymbol{\theta}))^2$ is independent of t and continuous on Θ. Let $\boldsymbol{\theta}^*$ denote the unique minimizer of $V(\boldsymbol{\theta})$ over Θ. On estimating $\boldsymbol{\theta}^*$ by $\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} S_n(\boldsymbol{\theta})$, where $S_n(\boldsymbol{\theta}) = \sum_{t=1}^{n-h}(y_{t+h} - g_{t,h}(\boldsymbol{\theta}))^2 \equiv \sum_{t=1}^{N} \varepsilon_{t,h}^2(\boldsymbol{\theta})$, the following theorem provides an asymptotic expression for $\mathrm{E}(y_{n+h} - g_{n,h}(\hat{\boldsymbol{\theta}}_n))^2$, taking a form similar to the right-hand side of (2.4). Define $D_1 g_{t,h}(\boldsymbol{\theta}) = (\partial g_{t,h}(\boldsymbol{\theta})/\partial\theta_1 \ldots \partial g_{t,h}(\boldsymbol{\theta})/\partial\theta_m)^\top$ and $D_2 g_{t,h}(\boldsymbol{\theta}) = (\partial^2 g_{t,h}(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j)_{1\le i,j\le m}$.

THEOREM S4.1. *Suppose that $g_{t,h}(\boldsymbol{\theta})$ is continuous on Θ and there is a $\delta > 0$ such that $D_1 g_{t,h}(\boldsymbol{\theta})$ is continuously differentiable on $B_\delta(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta\} \subset \Theta$, and each component of $D_2 g_{t,h}(\boldsymbol{\theta})$ is differentiable on $B_\delta(\boldsymbol{\theta}^*)$. Assume that conditions* (E1)–(E7) *in Section S4.2 hold. Then, for $h \ge 1$,*

$$\mathrm{E}(y_{n+h} - g_{n,h}(\hat{\boldsymbol{\theta}}_n))^2 = V(\boldsymbol{\theta}^*) + n^{-1}(L_h^* + o(1)), \tag{S4.1}$$

where $L_h^* = \mathrm{tr}((\mathbf{R}^* - \mathbf{A}^*)^{-1}\mathbf{C}_{h,0}^*) + 2\sum_{s=1}^{h-1}\mathrm{tr}((\mathbf{R}^* - \mathbf{A}^*)^{-1}\mathbf{C}_{h,s}^*)$, with $\mathbf{R}^* = \mathrm{E}\{D_1 g_{1,h}(\boldsymbol{\theta}^*) D_1^\top g_{1,h}(\boldsymbol{\theta}^*)\}$, $\mathbf{C}_{h,s}^* = \mathrm{E}\{D_1 g_{1,h}(\boldsymbol{\theta}^*) D_1^\top g_{1+s,h}(\boldsymbol{\theta}^*)\varepsilon_{1,h}(\boldsymbol{\theta}^*)\varepsilon_{1+s,h}(\boldsymbol{\theta}^*)\}$, $\mathbf{A}^* = \mathrm{E}\{D_2 g_{1,h}(\boldsymbol{\theta}^*)\varepsilon_{1,h}(\boldsymbol{\theta}^*)\}$, and $\mathbf{R}^*$ and $\mathbf{R}^* - \mathbf{A}^*$ being nonsingular.

Remark S1. There is a striking resemblance between (S4.1) and (2.4). In particular, (S4.1) reduces to (2.4) when $g_{t,h}(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$. Compared to the L_h in (2.4), L_h^* contains an additional matrix $\mathbf{A}^*$ reflecting the joint effect of nonlinearity and model misspecification. This matrix vanishes either when $g_{t,h}(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$ or is correct up to an independent error. See White (1981) for the definition of the latter property. In addition to its indispensable role in model selection, Theorem S4.1 is also of independent interest because it provides the first result revealing that the simple MSPE formula (2.7) obtained in correctly specified AR models carries over (after a mild modification) to misspecified nonlinear regressions with dependent observations in which previous research effort has mainly focused on the asymptotic properties of nonlinear least squares estimates; see, e.g., White (1984). The similarities and dissimilarities between (E1)–(E7) and (C1)–(C6) will be discussed in Section S4.2.

Consider K candidate models $g_{t,h}^{(l)}(\boldsymbol{\theta})$, $l = 1, \ldots, K$, for predicting y_{t+h}, where $g_{t,h}^{(l)}(\boldsymbol{\theta})$ is $\mathcal{F}_t$-measurable for each $\boldsymbol{\theta} \in \Theta_l$, with Θ_l denoting a compact parameter space whose dimension may vary with l. Assume that for each $1 \le l \le K$, $V_l(\boldsymbol{\theta}) \equiv \mathrm{E}(y_{t+h} - g_{t,h}^{(l)}(\boldsymbol{\theta}))^2$ is independent of t and continuous on Θ_l. Let $\boldsymbol{\theta}_l^*$ denote the unique minimizer of $V_l(\boldsymbol{\theta})$ over Θ_l. To estimate $\boldsymbol{\theta}_l^*$, we use $\hat{\boldsymbol{\theta}}_{nl} = \arg\min_{\boldsymbol{\theta}\in\Theta_l} S_n^{(l)}(\boldsymbol{\theta})$, where $S_n^{(l)}(\boldsymbol{\theta}) = \sum_{t=1}^{n-h}(y_{t,h} - g_{t,h}^{(l)}(\boldsymbol{\theta}))^2 \equiv \sum_{t=1}^{N}(\varepsilon_{t,h}^{(l)}(\boldsymbol{\theta}))^2$.

Define

$$\mathbf{R}^*(l) = \mathrm{E}\left(D_1 g_{1,h}^{(l)}(\boldsymbol{\theta}_l^*) D_1^\top g_{1,h}^{(l)}(\boldsymbol{\theta}_l^*)\right),$$

$$\mathbf{A}^*(l) = \mathrm{E}\left(D_2 g_{1,h}^{(l)}(\boldsymbol{\theta}_l^*) \varepsilon_{1,h}^{(l)}(\boldsymbol{\theta}_l^*)\right),$$

$$\mathbf{C}_{h,s}^*(l) = \mathrm{E}\left(D_1 g_{1,h}^{(l)}(\boldsymbol{\theta}_l^*) D_1^\top g_{1+s,h}^{(l)}(\boldsymbol{\theta}_l^*) \varepsilon_{1,h}^{(l)}(\boldsymbol{\theta}_l^*) \varepsilon_{1+s,h}^{(l)}(\boldsymbol{\theta}_l^*)\right),$$

and assume $\mathbf{R}^*(l)$ and $\mathbf{R}^*(l) - \mathbf{A}^*(l)$ are nonsingular. In view of Theorem S4.1, the nonlinear counterparts of M_1 and M_2 (defined in Section 3) are given by

$$\mathcal{D}_1 = \{k : 1 \le k \le K,\, V_l(\boldsymbol{\theta}_k^*) = \min_{1\le l\le K} V_l(\boldsymbol{\theta}_l^*)\} \text{ and } \mathcal{D}_2 = \{k : L_h^*(k) = \min_{l\in\mathcal{D}_1} L_h^*(l)\},$$

respectively, where

$$L_h^*(l) = \mathrm{tr}\left((\mathbf{R}^*(l) - \mathbf{A}^*(l))^{-1}\mathbf{C}_{h,0}^*(l)\right) + 2\,\mathrm{tr}\Big(\sum_{s=1}^{h-1}(\mathbf{R}^*(l) - \mathbf{A}^*(l))^{-1}\mathbf{C}_{h,s}^*(l)\Big).$$

To find a model whose index falls with $\mathcal{D}_2$, we suggest using a nonlinear extension of (3.9),

$$\text{MRIC}_h^*(l) = \frac{S_n^{(l)}(\hat{\boldsymbol{\theta}}_{nl})}{N} + \frac{C_n}{n}\widehat{L}_h^*(l), \tag{S4.2}$$

where C_n satisfies (3.10) and (3.11),

$$\widehat{L}_h^*(l) = \text{tr}\left(\left(\widehat{\mathbf{R}}^*(l) - \widehat{\mathbf{A}}^*(l)\right)^{-1}\widehat{\mathbf{C}}_{h,0}^*(l)\right) + 2\,\text{tr}\left(\sum_{s=1}^{h-1}\left(\widehat{\mathbf{R}}^*(l) - \widehat{\mathbf{A}}^*(l)\right)^{-1}\widehat{\mathbf{C}}_{h,s}^*(l)\right)$$

with

$$\begin{aligned}
\widehat{\mathbf{R}}^*(l) &= \frac{1}{N}\sum_{t=1}^{N} D_1 g_{t,h}^{(l)}(\hat{\boldsymbol{\theta}}_{nl}) D_1^\top g_{t,h}^{(l)}(\hat{\boldsymbol{\theta}}_{nl}),\\
\widehat{\mathbf{A}}^*(l) &= \frac{1}{N}\sum_{t=1}^{N} D_2 g_{t,h}^{(l)}(\hat{\boldsymbol{\theta}}_{nl}) \varepsilon_{t,h}^{(l)}(\hat{\boldsymbol{\theta}}_{nl}),\\
\widehat{\mathbf{C}}_{h,s}^*(l) &= \frac{1}{N-s}\sum_{t=1}^{N-s} D_1 g_{t,h}^{(l)}(\hat{\boldsymbol{\theta}}_{nl}) D_1^\top g_{t+s,h}^{(l)}(\hat{\boldsymbol{\theta}}_{nl}) \varepsilon_{t,h}^{(l)}(\hat{\boldsymbol{\theta}}_{nl}) \varepsilon_{t+s,h}^{(l)}(\hat{\boldsymbol{\theta}}_{nl}).
\end{aligned}$$

The next theorem shows that

$$\lim_{n\to\infty} P(\hat{l}_h^* \in \mathcal{D}_2) = 1, \tag{S4.3}$$

where $\hat{l}_h^* = \arg\min_{1\le l\le K} \text{MRIC}_h^*(l)$.

THEOREM S4.2. *Suppose that for each $1 \le l \le K$, $g_{t,h}^{(l)}(\boldsymbol{\theta})$ is continuous on Θ_l and there is $\bar{\delta}_l > 0$ such that $D_1 g_{t,h}^{(l)}(\boldsymbol{\theta})$ is continuously differentiable on $B_{\bar{\delta}_l}(\boldsymbol{\theta}_l^*) \subset \Theta_l$ and each component of $D_2 g_{t,h}^{(l)}(\boldsymbol{\theta})$ is differentiable on $B_{\bar{\delta}_l}(\boldsymbol{\theta}_l^*)$. Assume also that conditions* (E1)–(E7) *in Section S4.2 hold for each candidate model. Then, for $h \ge 1$ and $1 \le l \le K$,*

$$\text{E}(y_{n+h} - g_{n,h}^{(l)}(\hat{\boldsymbol{\theta}}_{nl}))^2 = V_l(\boldsymbol{\theta}_l^*) + n^{-1}(L_h^*(l) + o(1)). \tag{S4.4}$$

Moreover, assume for each $1 \le l \le K$,

$$\frac{1}{n}\sum_{t=1}^{n}(\varepsilon_{t,h}^{(l)}(\boldsymbol{\theta}_l^*))^2 = V_l(\boldsymbol{\theta}_l^*) + O_p(n^{-1/2}), \tag{S4.5}$$

$$\frac{1}{n}\sum_{t=1}^{n} D_2 g_{t,h}^{(l)}(\boldsymbol{\theta}_l^*)\varepsilon_{t,h}^{(l)}(\boldsymbol{\theta}_l^*) = \mathbf{A}^*(l) + o_p(1), \tag{S4.6}$$

and

$$\frac{1}{n}\sum_{t=1}^{n} D_1 g_{t,h}^{(l)}(\boldsymbol{\theta}_l^*) D_1^\top g_{t+s,h}^{(l)}(\boldsymbol{\theta}_l^*)\varepsilon_{t,h}^{(l)}(\boldsymbol{\theta}_l^*)\varepsilon_{t+s,h}^{(l)}(\boldsymbol{\theta}_l^*) = \mathbf{C}_{h,s}^*(l) + o_p(1). \tag{S4.7}$$

Then, (S4.3) *follows.*

Remark S2. Whereas (S4.6) is exclusive to nonlinear regressions, (S4.5) and (S4.7) parallel (3.17) and (3.18) used in the linear case. When $\varepsilon_{t,h}^{(l)}(\boldsymbol{\theta}_l^*)$ and the components of $D_2 g_{t,h}^{(l)}(\boldsymbol{\theta}_l^*)$ and $D_1 g_{t,h}^{(l)}(\boldsymbol{\theta}_l^*)$ are linear processes, a discussion about how assumptions like (S4.5)–(S4.7) are verified has been given in Sections 2, 3 and S2. It is worth mentioning that although model selection criteria, such as GAIC, BIC, GBIC and GBIC$_p$, have been proposed to combat model misspecification under various nonlinear models, none of them has been proven to possess properties like (S4.3) when the FD framework is entertained. Based on the discrepancy between the least squares and weighted least squares estimates when models are misspecified, White (1981) proposed a testing-based approach to conduct model selection for misspecified nonlinear regressions. However, it still seems tricky to justify its asymptotic efficiency within the FD framework.

S4.2. *Conditions (E1)–(E7).* We start by listing (E1)–(E7) as follows.

(E1) $\mathrm{E}\left\|n^{-1/2}\sum_{t=1}^{n}\left(D_1 g_{t,h}(\boldsymbol{\theta}^*)D_1^\top g_{t,h}(\boldsymbol{\theta}^*) - \mathbf{R}^*\right)\right\|^3 = O(1)$.

(E2) $\mathrm{E}\{D_1 g_{t,h}(\boldsymbol{\theta}^*)D_1^\top g_{t+s,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t+s,h}(\boldsymbol{\theta}^*)\} = \mathbf{C}_{h,s}^*$ for all t, and

$$\mathrm{E}\left(D_1 g_{1,h}(\boldsymbol{\theta}^*)D_1^\top g_{n,h}(\boldsymbol{\theta}^*)\varepsilon_{1,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*)\right) = o(n^{-1}).$$

(E3) There exists $q_1 > 6$ such that for $j = 1, 2, 3$,

$$\sup_{-\infty<t<\infty} \mathrm{E}\Big(\sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)} \varepsilon_{t,h}^{2q_1}(\boldsymbol{\theta})\Big) < \infty,$$

$$\sup_{-\infty<t<\infty} \mathrm{E}\left(\sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)} \|D_j g_{t,h}(\boldsymbol{\theta})\|_F^{2q_1}\right) < \infty,$$

where $D_3 g_{t,h}(\boldsymbol{\theta}) = (\partial^3 g_{t,h}(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j\partial\theta_k)_{1\le i,j,k\le m}$ and $\|\mathbf{G}\|_F$ denotes the Frobenius norm of the matrix $\mathbf{G}$. Moreover,

$$\sup_{-\infty<t<\infty} \mathrm{E}(\sup_{\boldsymbol{\theta}\in\Theta} \varepsilon_{t,h}^{q_1}(\boldsymbol{\theta})) < \infty.$$

(E4) $\mathrm{E}\left\|n^{-1/2}\sum_{t=1}^{n} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)\right\|^3 = O(1)$.

(E5) $\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^3 = O(1)$, and there exists a sequence of positive integers, $\{l_n\}$, with $l_n \to \infty$ and $l_n = o(n^{1/2})$ such that $\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n-l_n})\|^3 = o(1)$.

(E6)

$$\sup_{-\infty<t<\infty} \mathrm{E}\left\|\mathrm{E}\Big(D_1 g_{t,h}(\boldsymbol{\theta}^*) D_1^\top g_{t,h}(\boldsymbol{\theta}^*)\Big|\mathcal{F}_{t-k}\Big) - \mathbf{R}^*\right\|^3 = o(1), \text{ as } k\to\infty,$$

$$\sup_{-\infty<t<\infty} \mathrm{E}\left\|\mathrm{E}\Big(D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)\Big|\mathcal{F}_{t-k}\Big)\right\|^6 = o(1), \text{ as } k\to\infty.$$

(E7)

$$\sup_{-\infty<t<\infty} \mathrm{E}\left\|\mathrm{E}\Big(D_2 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)\Big|\mathcal{F}_{t-k}\Big) - \mathbf{A}^*\right\|^3 = o(1), \text{ as } k\to\infty,$$

$$\mathrm{E}\left\|n^{-1/2}\sum_{t=1}^{n}(D_2 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*) - \mathbf{A}^*)\right\|^3 = O(1).$$

Some comments are in order. Conditions (E1)–(E4) and (E6) not only look like (C1)–(C4) and (C6), respectively, but also play a similar role in the proof of Theorem S4.1 to the latter conditions in the proof of Theorem 2.1. (E3) imposes a moment bound on the third-order derivative of $g_{t,h}(\boldsymbol{\theta})$. This type of condition seems quite natural in a rigorous derivation of information criteria under misspecified nonlinear models; see, for example, Lv and Liu (2014). Actually, (E5) and (C5) also parallel each other in their roles in the aforementioned proofs, although they do not take similar forms. To see this, note that (C5), together with (C1) and (C4), yields $\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\beta}}_n(h) - \boldsymbol{\beta}_h)\|^q = O(1)$ and $\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\beta}}_n(h) - \hat{\boldsymbol{\beta}}_{n-l_n}(h))\|^q = o(1)$ for some positive constant q, which are linear counterparts of the identities in (E5). On the other hand, we mention that (E5) is a high-level assumption and its justification is nontrivial and of independent interest; see Section S4.4. Condition (E7) can be understood as a 'nonlinear amendment' of (C1)–(C6), which

vanishes automatically when $g_{t,h}(\boldsymbol{\theta})$ is linear. Finally, we remark that Theorems S4.1 and S4.2 remain valid in the so-call 'asymptotic stationary' case, in which $\mathrm{E}(y_{t+h} - g_{t,h}(\boldsymbol{\theta}))^2$ may vary with t, but converges to $V(\boldsymbol{\theta})$ uniformly over Θ as $t \to \infty$. In this case, $\mathbf{R}^*$, $\mathbf{C}^*_{h,s}$ and $\mathbf{A}^*$ become

$$\lim_{t\to\infty} \mathrm{E}\{D_1 g_{t,h}(\boldsymbol{\theta}^*) D_1^\top g_{t,h}(\boldsymbol{\theta}^*)\},$$
$$\lim_{t\to\infty} \mathrm{E}\{D_1 g_{t,h}(\boldsymbol{\theta}^*) D_1^\top g_{t+s,h}(\boldsymbol{\theta}^*) \varepsilon_{t,h}(\boldsymbol{\theta}^*) \varepsilon_{t+s,h}(\boldsymbol{\theta}^*)\},$$
$$\lim_{t\to\infty} \mathrm{E}\{D_2 g_{t,h}(\boldsymbol{\theta}^*) \varepsilon_{t,h}(\boldsymbol{\theta}^*)\},$$

respectively, where all limits are assumed to be finite. We also need to make some minor changes to (E2), (E6) and (E7), namely, deleting the first statement of (E2) and changing the second one to

$$\mathrm{E}\left(D_1 g_{k,h}(\boldsymbol{\theta}^*) D_1^\top g_{k+n,h}(\boldsymbol{\theta}^*) \varepsilon_{k,h}(\boldsymbol{\theta}^*) \varepsilon_{k+n,h}(\boldsymbol{\theta}^*)\right) = o(n^{-1})$$

for sufficiently large k, and replacing the $\sup_{-\infty<t<\infty}$ in (E6) and (E7) by $\sup_{t\geq H_1}$, where H_1 is some large integer.

S4.3. *Proofs of Theorems S4.1 and S4.2.* Proof of Theorem S4.1. Note first that

(S4.8)
$$n\left\{\mathrm{E}\left(y_{n+h} - g_{n,h}(\hat{\boldsymbol{\theta}}_n)\right)^2 - \mathrm{E}(\varepsilon^2_{n,h}(\boldsymbol{\theta}^*))\right\} = \mathrm{E}\left\{n\left(\varepsilon_{n,h}(\hat{\boldsymbol{\theta}}_n) - \varepsilon_{n,h}(\boldsymbol{\theta}^*)\right)^2\right\}$$
$$+ 2\mathrm{E}\left\{n(\varepsilon_{n,h}(\hat{\boldsymbol{\theta}}_n) - \varepsilon_{n,h}(\boldsymbol{\theta}^*))\varepsilon_{n,h}(\boldsymbol{\theta}^*)\right\} \equiv \mathrm{E(I)} + 2\mathrm{E(II)}.$$

Let $B_n = \{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| < \delta \text{ and } \|\hat{\boldsymbol{\theta}}_{n-\ell_n} - \boldsymbol{\theta}^*\| < \delta\}$, and define $\boldsymbol{w}_n = D_1 g_{n,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*)$. By Taylor's theorem for multivariable functions, we obtain

(S4.9)
$$\mathrm{(II)} = -n\boldsymbol{w}_n^\top(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\mathbf{1}_{B_n} - \frac{n}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^\top\left(D_2 g_{n,h}(\breve{\boldsymbol{\theta}}_n)\varepsilon_{n,h}(\boldsymbol{\theta}^*)\right)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\mathbf{1}_{B_n}$$
$$+ n\left(\varepsilon_{n,h}(\hat{\theta}_n) - \varepsilon_{n,h}(\boldsymbol{\theta}^*)\right)\varepsilon_{n,h}(\boldsymbol{\theta}^*)\mathbf{1}_{B_n^c} \equiv \mathrm{(III)} + \mathrm{(IV)} + \mathrm{(V)},$$

where $\|\breve{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \leq \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|$. We first show that

(S4.10)
$$\mathrm{E(III)} = -\sum_{j=h}^{n-h} \mathrm{E}\left(D_1^\top g_{1,h}(\boldsymbol{\theta}^*)(\mathbf{R}^* - \mathbf{A}^*)^{-1} D_1 g_{1+j,h}(\boldsymbol{\theta}^*)\varepsilon_{1,h}(\boldsymbol{\theta}^*)\varepsilon_{1+j,h}(\boldsymbol{\theta}^*)\right) + o(1).$$

Let $\tilde{\mathbf{R}}_n = n^{-1}\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*) D_1^\top g_{t,h}(\boldsymbol{\theta}^*)$ and $\tilde{\mathbf{A}}_n = n^{-1}\sum_{t=1}^{n-h} D_2 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)$. Then, by the mean value theorem for vector-valued functions, we have on B_n,

$$\mathbf{0} = D_1 S_n(\hat{\boldsymbol{\theta}}_n) = D_1 S_n(\boldsymbol{\theta}^*) + \left\{\int_0^1 D_2 S_n(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*))dr\right\}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*),$$

yielding

(S4.11)

$$\begin{aligned}
&(\mathrm{III}) = -\boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)\mathbf{1}_{B_n} \\
&+ \boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\left\{\sqrt{n}(\tilde{\mathbf{R}}_n - \mathbf{R}^*) - \sqrt{n}(\tilde{\mathbf{A}}_n - \mathbf{A}^*)\right\}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\mathbf{1}_{B_n} \\
&+ \frac{1}{2}\boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\int_0^1 \left[D_2 S_n(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)) - D_2 S_n(\boldsymbol{\theta}^*)\right]dr(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\mathbf{1}_{B_n} \\
&\equiv (\mathrm{VI}){+}(\mathrm{VII}){+}(\mathrm{VIII}).
\end{aligned}$$

Define $\mathbf{R}^*_{n-l_n} = n^{-1}\sum_{t=1}^{n-l_n} D_1 g_{t,h}(\boldsymbol{\theta}^*) D_1^\top g_{t,h}(\boldsymbol{\theta}^*)$ and $\mathbf{A}^*_{n-l_n} = n^{-1}\sum_{t=1}^{n-l_n} D_2 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)$. Then, it follows that

$$\begin{aligned}
&|\mathrm{E}(\mathrm{VII})| \le \|(\mathbf{R}^* - \mathbf{A}^*)^{-1}\| \times \\
&\left\{\mathrm{E}\left(\|\boldsymbol{w}_n\|\left\{\|\sqrt{n}(\tilde{\mathbf{R}}_n - \mathbf{R}^*_{n-l_n})\| + \|\sqrt{n}(\tilde{\mathbf{A}}_n - \mathbf{A}^*_{n-l_n})\|\right\}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|\right)\right. \\
+\ &\mathrm{E}\left(\|\boldsymbol{w}_n\|\left\{\|\sqrt{n}(\mathbf{R}^*_{n-l_n} - \mathbf{R}^*)\| + \|\sqrt{n}(\mathbf{A}^*_{n-l_n} - \mathbf{A}^*)\|\right\}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n-l_n})\|\right) \\
+\ &\left.\mathrm{E}\left[\|E(\boldsymbol{w}_n|\mathcal{F}_{n-l_n})\|\left\{\|\sqrt{n}(\mathbf{R}^*_{n-l_n} - \mathbf{R}^*)\| + \|\sqrt{n}(\mathbf{A}^*_{n-l_n} - \mathbf{A}^*)\|\right\}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_{n-l_n} - \boldsymbol{\theta}^*)\|\right]\right\}.
\end{aligned}$$

This, (E1), (E3), (E5), (E6), (E7), and Hölder's inequality imply

$$|\mathrm{E}(\mathrm{VII})| = o(1). \tag{S4.12}$$

We next show that

$$|\mathrm{E}(\mathrm{VIII})| = o(1), \tag{S4.13}$$

whose proof is somewhat tricky. Express (VIII) as

(S4.14)
$$\begin{aligned}
&\frac{1}{2}\boldsymbol{w}_n^\top(\mathbf{R}^*-\mathbf{A}^*)^{-1}\int_0^1 D_2S_n\left(\boldsymbol{\theta}^*+r(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\right)\\
&-D_2S_n\left(\boldsymbol{\theta}^*+r(\hat{\boldsymbol{\theta}}_{n-l_n}-\boldsymbol{\theta}^*)\right)dr(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\mathbf{1}_{B_n}\\
&+\frac{1}{2}\boldsymbol{w}_n^\top(\mathbf{R}^*-\mathbf{A}^*)^{-1}\int_0^1 D_2S_n\left(\boldsymbol{\theta}^*+r(\hat{\boldsymbol{\theta}}_{n-l_n}-\boldsymbol{\theta}^*)\right)-D_2S_n(\boldsymbol{\theta}^*)dr(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\mathbf{1}_{B_n}\\
&\equiv \text{(IX)+(X)}.
\end{aligned}$$

By making use of

$$\begin{aligned}
(1/2)\left(D_3S_n(\boldsymbol{\theta})\right)_{ijk} &= \sum_{t=1}^{n-h}\Big\{\left(D_2g_{t,h}(\boldsymbol{\theta})\right)_{ij}\left(D_1g_{t,h}(\boldsymbol{\theta})\right)_k+\left(D_2g_{t,h}(\boldsymbol{\theta})\right)_{ik}\left(D_1g_{t,h}(\boldsymbol{\theta})\right)_j\\
&+\left(D_2g_{t,h}(\boldsymbol{\theta})\right)_{jk}\left(D_1g_{t,h}(\boldsymbol{\theta})\right)_i\Big\}-\sum_{t=1}^{n-h}\left(D_3g_{t,h}(\boldsymbol{\theta})\right)_{ijk}\varepsilon_{t,h}(\boldsymbol{\theta})
\end{aligned}$$

and Taylor's theorem for multivariable functions, we have for some $C^*>0$,

$$\begin{aligned}
|\text{(IX)}| \;\le\; & C^*\|\boldsymbol{w}_n\|\Bigg(\max_{j\in\{1,\ldots,3\}}\sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)}n^{-1}\sum_{t=1}^{n-h}\left\|D_jg_{t,h}(\boldsymbol{\theta})\right\|_F^2\\
&+\sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)}n^{-1}\sum_{t=1}^{n-h}\varepsilon_{t,h}^2(\boldsymbol{\theta})\Bigg)\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\|\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\hat{\boldsymbol{\theta}}_{n-l_n})\|,
\end{aligned}$$

which, together with Hölder's inequality, Jensen's inequality, (E3) and (E5), yields

$$\text{E}|\text{(IX)}| = o(1). \tag{S4.15}$$

Assumptions (E3) and (E5) also imply

(S4.16)
$$\begin{aligned}
\text{E(X)} = \text{E}\Bigg\{&\frac{1}{2}\boldsymbol{w}_n^\top(\mathbf{R}^*-\mathbf{A}^*)^{-1}\int_0^1 D_2S_{n-l_n}\left(\boldsymbol{\theta}^*+r(\hat{\boldsymbol{\theta}}_{n-l_n}-\boldsymbol{\theta}^*)\right)\\
&-D_2S_{n-l_n}(\boldsymbol{\theta}^*)dr(\hat{\boldsymbol{\theta}}_{n-l_n}-\boldsymbol{\theta}^*)\Bigg\}+o(1).
\end{aligned}$$

Using (S4.16), (E6) and an argument similar to that used to prove (S4.12) and (S4.15), we obtain $|\mathrm{E(X)}| = o(1)$. In view of this, (S4.14) and (S4.15), (S4.13) follows.

To deal with (VI), note that

$$\begin{aligned}\mathrm{(VI)} &= -\boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*) \\ &\quad + \boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*)\mathbf{1}_{B_n^c} \equiv \mathrm{(XI)+(XII)}.\end{aligned} \tag{S4.17}$$

It follows from (E2) that

$$\mathrm{E(XI)} = -\sum_{j=h}^{n-1}\mathrm{E}\left(D_1^\top g_{1,h}(\boldsymbol{\theta}^*)(\mathbf{R}^* - \mathbf{A}^*)^{-1} D_1 g_{1+j,h}(\boldsymbol{\theta}^*)\varepsilon_{1,h}(\boldsymbol{\theta}^*)\varepsilon_{1+j,h}(\boldsymbol{\theta}^*)\right). \tag{S4.18}$$

Assumptions (E3) and (E5) further yield

$$|\mathrm{E(XII)}| = o(1). \tag{S4.19}$$

Consequently, (S4.10) is guaranteed by (S4.11)–(S4.13) and (S4.17)–(S4.19).

Next, we calculate E(-2(IV)). Straightforward algebraic manipulations yield

$$\mathrm{E(-2(IV))} = \mathrm{E}\left(n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^\top \mathbf{A}^*(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\mathbf{1}_{B_n}\right) + \mathbf{G}_n, \tag{S4.20}$$

where

$$\begin{aligned}|\mathbf{G}_\mathrm{n}| &\le \mathrm{E}\left(n \sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)} \|D_3 g_{n,h}(\boldsymbol{\theta})\|_F \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|^3 |\varepsilon_{n,h}(\boldsymbol{\theta}^*)|\mathbf{1}_{B_n}\right) \\ &+ \mathrm{E}\left\{\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n-l_n})\|\left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\| + \|\sqrt{n}(\hat{\boldsymbol{\theta}}_{n-l_n} - \boldsymbol{\theta}^*)\|\right)\|D_2 g_{n,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*) - \mathbf{A}^*\|\right\} \\ &+ \mathrm{E}\left(\|\mathrm{E}\left(D_2 g_{n,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*)\big|\mathcal{F}_{n-l_n}\right) - \mathbf{A}^*\|\|\sqrt{n}(\hat{\boldsymbol{\theta}}_{n-l_n} - \boldsymbol{\theta}^*)\|^2\right) \\ &+ \mathrm{E}\left(\left\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\right\|^2 \|D_2 g_{n,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*) - \mathbf{A}^*\|\mathbf{1}_{B_n^c}\right).\end{aligned} \tag{S4.21}$$

Let ξ be an arbitrarily small positive number. It is not difficult to see that the first term on the right-hand side of (S4.21) is bounded above by $\delta^{1-\xi}\mathrm{E}(n\sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)}\|D_3 g_{n,h}(\boldsymbol{\theta})\|_F\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*\|^{2+\xi}|\varepsilon_{n,h}(\boldsymbol{\theta}^*)|)$, which in turn converges to 0 in view of (E3) and (E5). Moreover, by (E3), (E5) and (E7), the other three terms on the right-hand side of (S4.21) also vanish asymptotically. As a result,

$$|\mathbf{G}_{\mathrm{n}}| = o(1). \tag{S4.22}$$

On the other hand, we get from (E3), (E4), (E5) and some algebraic manipulations that

$$\begin{aligned}&\mathrm{E}(n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\mathbf{A}^*(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\mathbf{1}_{B_n})\\&=\mathrm{E}\{n^{-1}(\sum_{t=1}^{n-h}D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))V^*(\sum_{t=1}^{n-h}D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))\}+o(1),\end{aligned}$$

where $V^*=(\mathbf{R}^*-\mathbf{A}^*)^{-1}\mathbf{A}^*(\mathbf{R}^*-\mathbf{A}^*)^{-1}$. Combining this with (S4.22) and (S4.20) gives

$$\mathrm{E}((\mathrm{IV})) = \frac{-1}{2}\mathrm{E}\{n^{-1}(\sum_{t=1}^{n-h}D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))V^*(\sum_{t=1}^{n-h}D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))\}+o(1). \tag{S4.23}$$

It follows from (E3), (E5) and Hölder's inequality that $|\mathrm{E(V)}|=o(1)$. With this result and (S4.9), (S4.10) and (S4.23), we obtain

$$\begin{aligned}2\mathrm{E(II)} &= -2\sum_{j=h}^{n-h}\mathrm{E}(D_1^\top g_{1,h}(\boldsymbol{\theta}^*)(\mathbf{R}^*-\mathbf{A}^*)^{-1}D_1 g_{1+j,h}(\boldsymbol{\theta}^*)\varepsilon_{1,h}(\boldsymbol{\theta}^*)\varepsilon_{1+j,h}(\boldsymbol{\theta}^*))\\&-\mathrm{E}\{n^{-1}(\sum_{t=1}^{n-h}D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))V^*(\sum_{t=1}^{n-h}D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))\}+o(1).\end{aligned} \tag{S4.24}$$

Applying Taylor's theorem for multivariable functions again, we have

$$\begin{aligned}(\mathrm{I}) = [&\sqrt{n}(D_1 g_{n,h}(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*))\mathbf{1}_{B_n}+\frac{\sqrt{n}}{2}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)D_2 g_{n,h}(\breve{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\mathbf{1}_{B_n}\\&+\sqrt{n}(g_{n,h}(\hat{\boldsymbol{\theta}}_n)-g_{n,h}(\boldsymbol{\theta}^*))\mathbf{1}_{B_n^c}]^2\equiv[(\mathrm{XIII})+(\mathrm{XIV})+(\mathrm{XV})]^2,\end{aligned}$$

where $\|\breve{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \le \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|$. An argument similar to that used to prove (S4.24) yields

$$\mathrm{E(XIII)}^2 = \mathrm{E}\{n^{-1}(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))Q^*(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))\} + o(1),$$

$$\mathrm{E(XIV)}^2 = o(1),$$

$$\mathrm{E(XV)}^2 = o(1),$$

where $Q^* = (\mathbf{R}^* - \mathbf{A}^*)^{-1}\mathbf{R}^*(\mathbf{R}^* - \mathbf{A}^*)^{-1}$, and hence

(S4.25)

$$\mathrm{E(I)} = \mathrm{E}\{n^{-1}(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))Q^*(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))\} + o(1).$$

Finally, the desired conclusion is ensured by (S4.8), (S4.24), (S4.25) and (E2).

Remark S3. We have obtained a preliminary result, extending Theorem S4.1 to the threshold autoregressive (TAR) model,

$$\text{(S4.26)}\quad y_t = \begin{cases} a_{11}y_{t-1} + \cdots + a_{1p}y_{t-p} + e_t = \mathbf{a}_1^\top \mathbf{y}_{t-1} + e_t, & y_{t-d} \le r, \\ a_{21}y_{t-1} + \cdots + a_{2p}y_{t-p} + e_t = \mathbf{a}_2^\top \mathbf{y}_{t-1} + e_t, & y_{t-d} > r, \end{cases}$$

where $\mathbf{y}_t = (y_t, \ldots, y_{t-p+1})^\top$, r is an unknown threshold parameter, $1 \le d \le p$ is an unknown integer, $\mathbf{a}_1 \ne \mathbf{a}_2$ are unknown p-dimensional vectors, and e_t are i.i.d. random errors with $\mathrm{E}(e_1) = 0$ and $\mathrm{E}e_1^2 = \sigma^2$. Note that model (S4.26) is driven by nonlinear dynamics in that it is nonlinear in both parameters and lagged variables. We propose using

$$\hat{y}_{n+1} = \begin{cases} \hat{\mathbf{a}}_1^\top \mathbf{y}_n, & y_{n+1-\hat{d}} \le \hat{r} \\ \hat{\mathbf{a}}_2^\top \mathbf{y}_n, & y_{n+1-\hat{d}} > \hat{r} \end{cases}$$

to predict y_{n+1}, where $(\hat{\mathbf{a}}_1^\top, \hat{\mathbf{a}}_2^\top, \hat{r}, \hat{d})^\top$, modified slightly from the CLSE of Chan (1993), is a consistent estimate of $(\mathbf{a}_1^\top, \mathbf{a}_2^\top, r, d)^\top$ and shares the same asymptotic distribution as the latter estimate. Let $\pi(\cdot)$ be the density function of y_t, and $[M_-, M_+)$ denote the random interval on which the process $\{\tilde{L}^{(1)}(-z)I_{(z\le 0)} + \tilde{L}^{(2)}(z)I_{(z>0)}, z \in R\}$ attains its global minimum, where $\{\tilde{L}^{(1)}(z), z \ge 0\}$ and $\{\tilde{L}^{(2)}(z), z \ge 0\}$ are

two independent compound Poisson processes, with common rate $\pi(r)$, $\tilde{L}^{(1)}(0) = \tilde{L}^{(2)}(0) = 0$ almost surely, and the distributions of jump being given by the conditional distribution of $[(\mathbf{a}_1 - \mathbf{a}_2)^\top \mathbf{y}_p]^2 + 2(\mathbf{a}_1 - \mathbf{a}_2)^\top \mathbf{y}_p e_{p+1}$ given $y_{p+1-d} = r_-$ and the conditional distribution of $[(\mathbf{a}_2 - \mathbf{a}_1)^\top \mathbf{y}_p]^2 + 2(\mathbf{a}_2 - \mathbf{a}_1)^\top \mathbf{y}_p e_{p+1}$ given $y_{p+1-d} = r_+$, respectively. We show that under correct model specification and some regularity conditions,

$$\begin{aligned} & n\left[\mathrm{E}(y_{n+1} - \hat{y}_{n+1})^2 - \sigma^2\right] \\ & = \mathrm{E}\left[\left\{(\mathbf{a}_2 - \mathbf{a}_1)^\top \mathbf{y}_p\right\}^2 | y_{p+1-d} = r\right] \pi(r)\mathrm{E}|M_-| + 2p\sigma^2. \end{aligned} \tag{S4.27}$$

To the best of our knowledge, (S4.27) is the first result that provides an asymptotic expression for the MSPE of TAR models. Whereas the second term on the right-hand side of (S4.27) is expected in the sense that it is proportional to the number of AR coefficients and contributed by the estimation error of $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$, the first term appears to be unforeseeable because it is essentially due to the estimation error of $\hat{r}$, whose rate of convergence is much faster than those of $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$. Owing to the non-differentiability of the TAR model at the true parameter, most conditions described in (E1)–(E7) are violated. Therefore, (S4.27) and (S4.1) of Theorem S4.1 not only don't resemble each other in statement, but also have very different proofs; see Chi et al. (2017).

The above discussion reminds us of the fact that when regularity conditions like those in Section S4.2 fail to hold, a separate investigation on model selection (based on MSPE) needs to be conducted case by case, that is, for each different nonlinear time series. This task, however, is beyond the scope of the current paper.

Proof of Theorem S4.2. Equation (S4.4) is an immediate consequence of Theorem S4.1. Equation (S4.3) is ensured by

$$\begin{aligned} & \frac{S_n^{(l)}(\hat{\boldsymbol{\theta}}_{nl})}{N} = V_l(\boldsymbol{\theta}_l^*) + O_p(\frac{1}{\sqrt{n}}), \\ & \widehat{\mathbf{C}}_{h,s}^*(l) = \mathbf{C}_{h,s}^*(l) + o_p(1), \\ & \widehat{\mathbf{R}}^*(l) = \mathbf{R}^*(l) + o_p(1), \\ & \widehat{\mathbf{A}}^*(l) = \mathbf{A}^*(l) + o_p(1), \end{aligned} \tag{S4.28}$$

which follow from (S4.5)–(S4.7) and an argument similar to that used to prove Theorem S4.1.

S4.4. *Moment bounds for* $\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|$ *and* $\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n-l_n})\|$. Let $q \geq 1$ and $l_n = o(n^{1/2})$. In this section, we provide sufficient conditions under which

$$\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q = O(1), \tag{S4.29}$$

and

$$\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n-l_n})\|^q = o(1). \tag{S4.30}$$

Define $f_{t,h} = \mathrm{E}(y_{t+h}|\mathcal{F}_t)$ and $\eta_{t,h} = y_{t+h} - f_{t,h}$. It is easy to show that $\eta_{t,h}$ is uncorrelated with $g_{t,h}(\boldsymbol{\theta})$ and $V(\boldsymbol{\theta}) = V_o(\boldsymbol{\theta}) + \mathrm{E}(\eta_{t,h}^2)$, where $V_o(\boldsymbol{\theta}) = \mathrm{E}(f_{t,h} - g_{t,h}(\boldsymbol{\theta}))^2$ is independent of t. In view of the continuity of $V(\boldsymbol{\theta})$ on Θ, it is not difficulty to see that $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* = o_p(1)$ is ensured by

$$\begin{aligned} &\sup_{\boldsymbol{\theta}\in\Theta} |n^{-1}\sum_{t=1}^{n}(g_{t,h}(\boldsymbol{\theta}^*) - g_{t,h}(\boldsymbol{\theta}))\eta_{t,h}| = o_p(1), \\ &\sup_{\boldsymbol{\theta}\in\Theta} |n^{-1}\sum_{t=1}^{n}(f_{t,h} - g_{t,h}(\boldsymbol{\theta}))^2 - V_o(\boldsymbol{\theta})| = o_p(1). \end{aligned} \tag{S4.31}$$

However, to prove (S4.29) and (S4.30), we need a strengthened version of (S4.31) among other conditions.

THEOREM S4.3. *Suppose that* $g_{t,h}(\boldsymbol{\theta})$ *is continuous on* Θ *and there is* $\delta > 0$ *such that* $D_1 g_{t,h}(\boldsymbol{\theta})$ *is continuously differentiable on* $B_\delta(\boldsymbol{\theta}^*)$ *and each component of* $D_2 g_{t,h}(\boldsymbol{\theta})$ *is differentiable on* $B_\delta(\boldsymbol{\theta}^*)$. *Assume that* (E1), (E4) *and the second relation of* (E7) *hold with* 3 *replaced by* q,

$$\sup_{-\infty<t<\infty}\sum_{j=1}^{3}\mathrm{E}\left(\sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)}\left\|D_j g_{t,h}(\boldsymbol{\theta})\right\|_F^{4q}\right) + \sup_{-\infty<t<\infty}\mathrm{E}\left(\sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)}\varepsilon_{t,h}^{4q}(\boldsymbol{\theta})\right) < \infty, \tag{S4.32}$$

$$\begin{aligned} &n^{q/2}\mathrm{P}\left(\sup_{\boldsymbol{\theta}\in\Theta}|n^{-1}\sum_{t=1}^{n}(g_{t,h}(\boldsymbol{\theta}^*) - g_{t,h}(\boldsymbol{\theta}))\eta_{t,h}| > \epsilon\right) = o(1), \\ &n^{q/2}\mathrm{P}\left(\sup_{\boldsymbol{\theta}\in\Theta}|n^{-1}\sum_{t=1}^{n}(f_{t,h} - g_{t,h}(\boldsymbol{\theta}))^2 - V_o(\boldsymbol{\theta})| > \epsilon\right) = o(1), \text{ for any } \epsilon > 0, \end{aligned} \tag{S4.33}$$

and there is $\bar{M} > 0$ *such that*

(S4.34)
$$n^q \mathrm{P}\left(\sup_{\boldsymbol{\theta} \in B_\delta(\boldsymbol{\theta}^*)} n^{-1} \sum_{t=1}^{n} \varepsilon_{t,h}^2(\boldsymbol{\theta}) > \bar{M} \right) = O(1),$$

$$n^q \mathrm{P}\left(\sup_{\boldsymbol{\theta} \in B_\delta(\boldsymbol{\theta}^*)} n^{-1} \sum_{t=1}^{n} \|D_j g_{t,h}(\boldsymbol{\theta})\|_F^2 > \bar{M} \right) = O(1), j = 1, 2, 3.$$

Then, (S4.29) *and* (S4.30) *hold true.*

PROOF. We begin by proving (S4.29). There is $C^* > 0$ such that on the set $\{\hat{\boldsymbol{\theta}}_n \in B_\delta(\boldsymbol{\theta}^*)\}$,

(S4.35)
$$\begin{aligned} &\left\| (2n)^{-1} \int_0^1 D_2 S_n(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)) - D_2 S_n(\boldsymbol{\theta}^*) dr \right\| \\ &\leq C^* \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \Lambda_n, \end{aligned}$$

where

$$\Lambda_n = \max_{j \in \{1,\ldots,3\}} \sup_{\boldsymbol{\theta} \in B_\delta(\boldsymbol{\theta}^*)} n^{-1} \sum_{t=1}^{n-h} \|D_j g_{t,h}(\boldsymbol{\theta})\|_F^2 + \sup_{\boldsymbol{\theta} \in B_\delta(\boldsymbol{\theta}^*)} n^{-1} \sum_{t=1}^{n-h} \varepsilon_{t,h}^2(\boldsymbol{\theta}).$$

Define $Q_n = \{\Lambda_n \leq 2\bar{M}\}$. Let $0 < \delta^* < \min\{\delta, (C^* \|(\mathbf{R}^* - \mathbf{A}^*)^{-1}\| 6\bar{M})^{-1}\}$ and $H_n = \{\hat{\boldsymbol{\theta}}_n \in B_{\delta^*}(\boldsymbol{\theta}^*)\}$. Then, it follows from the mean value theorem for vector-valued functions that

(S4.36)
$$\begin{aligned} &\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n} \\ &\leq 3^q \|(\mathbf{R}^* - \mathbf{A}^*)^{-1}\|^q \Bigg[\left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{N} D_1 g_{t,h}(\boldsymbol{\theta}^*) \varepsilon_{t,h}(\boldsymbol{\theta}^*) \right\|^q \\ &+ \left(\|\sqrt{n}(\tilde{\mathbf{R}}_n^* - \mathbf{R}^*)\| + \|\sqrt{n}(\tilde{\mathbf{A}}_n^* - \mathbf{A}^*)\| \right)^q \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|^q \mathbf{1}_{H_n} \\ &+ \left\| (2n)^{-1} \int_0^1 D_2 S_n(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)) - D_2 S_n(\boldsymbol{\theta}^*) dr \right\|^q \|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n} \Bigg]. \end{aligned}$$

By (S4.35), (S4.36) and the hypotheses associated with (E1), (E4) and (E7), we obtain

(S4.37)
$$\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n} \le O(1) + 3^q \|(\mathbf{R}^* - \mathbf{A}^*)^{-1}\|^q (2\bar{M}C^*\delta^*)^q \|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n}$$
$$+ 3^q \|(\mathbf{R}^* - \mathbf{A}^*)^{-1}\|^q (C^*)^q (\delta^*)^{2q} \mathrm{E}(n^{q/2}\Lambda_n^q \mathbf{1}_{Q_n}).$$

Note that $3^q\|(\mathbf{R}^* - \mathbf{A}^*)^{-1}\|^q(2\bar{M}C^*\delta^*)^q < 1$ and $\mathrm{E}(n^{q/2}\Lambda_n^q \mathbf{1}_{Q_n}) = O(1)$ is ensured by (S4.32) and (S4.34). Combining these with (S4.37) gives

$$\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n} = O(1).$$

Moreover, it follows from (S4.33), the compactness of Θ and the continuity of $V(\boldsymbol{\theta})$ that

$$\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n^c} = O(n^{q/2}\mathrm{P}(H_n^c)) = o(1).$$

This completes the proof of (S4.29). Expressing $D_1 S_n(\hat{\boldsymbol{\theta}}_n)$ as the sum of

$$\begin{aligned} D_1 S_n(\hat{\boldsymbol{\theta}}_{n-l_n}) &= D_1 S_n(\hat{\boldsymbol{\theta}}_{n-l_n}) - D_1 S_{n-l_n}(\hat{\boldsymbol{\theta}}_{n-l_n}) \\ &= -2 \sum_{t=n-h-l_n+1}^{n-h} D_1 g_{t,h}(\hat{\boldsymbol{\theta}}_{n-l_n}) \varepsilon_{t,h}(\hat{\boldsymbol{\theta}}_{n-l_n}) \end{aligned}$$

and a remainder term using the mean value theorem for vector-valued functions, we can prove (S4.30) in a fashion similar to the proof of (S4.29). The details are omitted.

Remark S4. Being an extension of Theorem 2.2 of Chan and Ing (2011), Theorem S4.3 establishes the first result on moment convergence of the least squares estimates in *misspecified nonlinear regressions with dependent observations*. Its applications to prediction and model selection have been illustrated via Theorems S4.1 and S4.2. Note that in the special case of $q = 3$, (S4.32) is weaker than condition (E3). (S4.33) is a strengthened version of (S4.31), and the role of (S4.34) in the proof of Theorem S4.3 is similar to that of (2.13) and (2.14) in the proof of Theorem 2.2 of Chan and Ing (2011). When $f_{t,h}$, $g_{t,h}(\boldsymbol{\theta})$, $\varepsilon_{t,h}(\boldsymbol{\theta})$ and $D_j g_{h,t}(\boldsymbol{\theta})$ are linear processes and the coefficient functions in the latter three satisfy certain smoothness conditions, (S4.33) and (S4.34) can be justified via an argument similar to that used in Lemma B.1 of Chan

and Ing (2011), which is a 'uniform version' of the First Moment Bound Theorem of Findley and Wei (1993). It is worth mentioning that while Theorem 2.2 of Chan and Ing (2011) is proved without imposing assumptions on the third-order derivative of the regression function, some extra *distributional* assumptions like (2.19) on the regression function are needed. Therefore, there exists a trade-off between the smoothness of the regression function and the smoothness of its distribution, which is a subject of further investigation.

S5. Numerical Studies. In this section, the performance of MRIC is illustrated via five simulated examples. The first and second examples focus on linear and nonliner models, respectively, the third and fourth examples address high-dimensional models, and the last one provides a further investigation of Example 2. Throughout this section, the C_n in MRIC is set to n^{α_m} for some $\alpha_m > 0.5$.

Example 1. Let the data be generated according to the following DGPs.

$$y_{t+1} = \beta_1 z_t + \beta_2 w_t + \varepsilon_{t+1}, \tag{S5.1}$$

in which $\varepsilon_t \sim \text{NID}(0,1)$, $z_t = \phi z_{t-1} + \eta_t$ is a stationary AR(1) process, and $w_t = \theta_1 w_{t-1} + \theta_2 w_{t-2} + \delta_t$ is a stationary AR(2) process, with $\eta_t \sim \text{NID}(0, \sigma_\eta^2)$, $\delta_t \sim \text{NID}(0, \sigma_\delta^2)$, and $\{\eta_t\}$, $\{\delta_t\}$ and $\{\varepsilon_t\}$ independent. Here and hereafter $\text{NID}(0, \sigma^2)$ means normally and independently distributed with mean 0 and variance σ^2. We also let

$$\sigma_\eta^2 = 1 - \phi^2, \sigma_\delta^2 = 1 - \theta_2^2 - \{\theta_1^2(1+\theta_2)/(1-\theta_2)\},$$

$\beta_1 = \beta_2 = 1$, and $\phi = \theta_1/(1-\theta_2)$, noting that (S5.1) leads to $\gamma_z(0) = 1 = \gamma_w(0)$, where $\gamma_z(j) = \text{E}(z_t z_{t+j})$ and $\gamma_w(j) = \text{E}(w_t w_{t+j})$. In this study, we consider four different (θ_1, θ_2)'s: (0.15, 0.5), (-0.10, 0.65), (-0.40, -0.60), (0.10, -0.95), which are denoted by DGPs I-IV. With observations up to time n, we are interested in performing h-step prediction, with $h = 2$ and 3, using two candidate models,

$$\begin{aligned} J_1 : &\quad y_{n+h} = \alpha z_n + \varepsilon_{n,h}^{(1)}, \\ J_2 : &\quad y_{n+h} = \beta w_n + \varepsilon_{n,h}^{(2)}, \end{aligned}$$

which are misspecified. The MI and VI of candidate J_l are denoted by $\text{MI}_h(l)$ and $L_h(l)$ with $l = 1, 2$. It is shown in Table S1 that $\text{MI}_2(1) =$

$\mathrm{MI}_2(2)$ in all four DGPs, but $L_2(1) < L_2(2)$ in DGPs I and II and $L_2(1) > L_2(2)$ in DGPs III and IV. Therefore, for the two-step prediction, the better predictive model is J_1 (J_2) under DGP I or II (III or IV). On the other hand, Table S1 reveals that $\mathrm{MI}_3(1) > \mathrm{MI}_3(2)$ in all DGPs, yielding that the better predictive model is always J_2 when $h = 3$. The percentage of MRIC, with $\alpha_m = 0.6$, choosing the better candidate is obtained by using 1,000 simulations for sample sizes $n = 200, 500, 1000, 2000, 3000$; see Table S2 ($h = 2$) and Table S3 ($h = 3$). A data-driven method for choosing α_m is suggested in Section S6. For the sake of comparison, the corresponding percentages of AIC, BIC, GAIC (Konishi and Kitagawa, 1996), GBIC (Lv and Liu, 2014) and GBIC_p (Lv and Liu, 2014) are also reported in Tables S2 and S3, where for candidate J_l,

$$\begin{aligned}
\mathrm{AIC}(l) &= \log \hat{\sigma}_h^2(l) + \frac{2\sharp(J_l)}{n},\\
\mathrm{BIC}(l) &= \log \hat{\sigma}_h^2(l) + \frac{\sharp(J_l)\log n}{n},\\
\mathrm{GAIC}(l) &= \log \hat{\sigma}_h^2(l) + \frac{2\mathrm{tr}(\widehat{H}_h(l))}{n},\\
\mathrm{GBIC}(l) &= \log \hat{\sigma}_h^2(l) + \frac{\sharp(J_l)\log n}{n} - \frac{\log\det(\widehat{H}_h(l))}{n},\\
\mathrm{GBIC}_p(l) &= \log \hat{\sigma}_h^2(l) + \frac{\sharp(J_l)\log n}{n} + \frac{\mathrm{tr}(\widehat{H}_h(l))}{n} - \frac{\log\det(\widehat{H}_h(l))}{n},
\end{aligned}$$

with

$$\widehat{H}_h(l) = \hat{\sigma}_h^{-2}(l)\widehat{\mathbf{R}}^{-1}(l)\widehat{\mathbf{C}}_{h,0}(l),$$

which is a consistent estimator of $\sigma_h^{-2}(l)\mathbf{R}^{-1}(l)\mathbf{C}_{h,0}(l)$. Note first that $\mathrm{MRIC}(l)$ is asymptotically equivalent to

$$\log \hat{\sigma}_h^2(l) + \frac{C_n\hat{\sigma}_h^{-2}(l)\hat{L}_h(l)}{n},$$

which shares a common first term with these five criteria. On the other hand, by featuring a consistent estimator of VI, $\hat{L}_h(l)$, and a suitable penalty term, C_n, the second term of MRIC readily paves the way for a consistent selection of the better predictive model, whether the MIs of candidate models are equal or not. We also mention that this latter property is, in general, not enjoyed by these five criteria because (i)

TABLE S1
The values of $\mathrm{MI}_h(1) - \mathrm{MI}_h(2)$ *and* $L_h(1) - L_h(2)$ *in Example 1, and the corresponding better predictive models*

	DGP			
	I	II	III	IV
		$h = 2$		
$\mathrm{MI}_h(1) - \mathrm{MI}_h(2)$	0.000	0.000	0.000	0.000
$L_h(1) - L_h(2)$	-0.746	-0.999	0.984	1.890
The better predictive model	J_1	J_1	J_2	J_2
		$h = 3$		
$\mathrm{MI}_h(1) - \mathrm{MI}_h(2)$	0.289	0.454	0.246	0.893
$L_h(1) - L_h(2)$	*	*	*	*
The better predictive model	J_2	J_2	J_2	J_2

*: $L_h(1) - L_h(2)$ can be neglected.

the trace of $\widehat{\mathbf{R}}^{-1}(l)\widehat{\mathbf{C}}_{h,0}(l)$ in $\widehat{H}_h(l)$ is a consistent estimator of VI only when $h = 1$ or observations are independent over time, and (ii) the penalty term $\log n$ used in BIC, GBIC and GBIC_p is too weak when misspecified candidates are non-nested (see Sin and White (1996) and Inoue and Kilian (2006) for related discussion). In fact, the criterion values of GAIC (AIC, BIC, GBIC, GBIC_p) for J_1 and J_2 are expected to be close to each other because $\mathrm{MI}_h(1) = \mathrm{MI}_h(2)$, $\sharp(J_1) = \sharp(J_2)$ and $\mathrm{tr}(\mathbf{R}^{-1}(l)\mathbf{C}_{h,0}(l)) = \sharp(J_l)\mathrm{MI}_h(l)$ (under normality). As shown in Table S2, these five criteria behave like a fair coin to choose between two alternatives, and can only select the better candidate about 50% of the time. In contrast, MRIC has a much higher chance of identifying the better model in this difficult situation. Its percentage falls between 67% and 100%, and tends to increase with the sample size and the value of $|L_2(1) - L_2(2)|$.

When $h = 3$, the two competing candidates have different MIs, and hence it becomes much easier to identify the better one. As shown in Table S3, all criteria perform satisfactorily for all sample sizes $n \geq 200$. While in DGPs I and III, MRIC seems slightly worse than the other criteria for $n = 200$, the corresponding percentages are still over 93%.

Example 2. In this example, we consider the following DGP,

$$y_{t+2} = \frac{1}{1 - aB}x_t + \frac{1}{1 - bB}z_t + \varepsilon_{t+2}, \tag{S5.2}$$

in which $|a| < 1$, $|b| < 1$, $\varepsilon_t \sim \mathrm{NID}(0, \sigma_\varepsilon^2)$, $x_t \sim \mathrm{NID}(0, \sigma_x^2)$, $z_t \sim \mathrm{NID}(0, \sigma_z^2)$, and $\{\varepsilon_t\}$, $\{x_t\}$ and $\{z_t\}$ are independent. Note that model

TABLE S2

Percentage of times, across 1,000 simulations, that the better predictive model between J_1 and J_2 of Example 1 is chosen in the case of $h = 2$

		DGPs			
n	Criteria	I	II	III	IV
200	AIC/BIC	51.50	54.50	48.50	46.30
	GAIC	51.40	54.30	49.00	46.70
	GBIC	51.60	54.40	48.50	45.40
	GBICp	51.60	54.40	48.40	46.00
	MRIC	66.80	73.20	76.70	95.80
500	AIC/BIC	51.10	50.70	47.60	49.00
	GAIC	50.80	50.50	47.30	50.90
	GBIC	51.10	50.50	47.60	47.30
	GBICp	51.10	50.70	47.60	49.10
	MRIC	69.80	74.20	85.30	99.70
1000	AIC/BIC	48.10	53.60	53.00	49.40
	GAIC	48.00	53.00	52.40	50.00
	GBIC	48.10	53.50	52.80	49.20
	GBICp	48.10	53.50	53.00	49.40
	MRIC	74.90	80.80	88.70	100.00
2000	AIC/BIC	50.10	49.70	50.80	49.60
	GAIC	50.10	49.50	50.90	49.20
	GBIC	50.30	49.70	50.90	49.30
	GBICp	50.10	49.70	50.80	49.60
	MRIC	78.20	83.90	92.20	100.00
3000	AIC/BIC	51.40	51.20	49.00	50.40
	GAIC	51.40	51.10	48.90	50.60
	GBIC	51.30	51.20	49.00	50.70
	GBICp	51.40	51.20	49.00	50.40
	MRIC	79.80	84.90	93.40	100.00

TABLE S3

Percentage of times, across 1,000 simulations, that the better predictive model between J_1 and J_2 of Example 1 is chosen in the case of $h = 3$

		DGPs			
n	Criteria	I	II	III	IV
200	AIC/BIC	99.30	100.00	99.30	100.00
	GAIC	99.30	100.00	99.10	100.00
	GBIC	99.30	100.00	99.30	100.00
	GBICp	99.20	100.00	99.30	100.00
	MRIC	93.20	97.90	94.70	100.00
500	AIC/BIC	100.00	100.00	100.00	100.00
	GAIC	100.00	100.00	100.00	100.00
	GBIC	100.00	100.00	100.00	100.00
	GBICp	100.00	100.00	100.00	100.00
	MRIC	99.90	100.00	100.00	100.00
1000	AIC/BIC	100.00	100.00	100.00	100.00
	GAIC	100.00	100.00	100.00	100.00
	GBIC	100.00	100.00	100.00	100.00
	GBICp	100.00	100.00	100.00	100.00
	MRIC	100.00	100.00	100.00	100.00
2000	AIC/BIC	100.00	100.00	100.00	100.00
	GAIC	100.00	100.00	100.00	100.00
	GBIC	100.00	100.00	100.00	100.00
	GBICp	100.00	100.00	100.00	100.00
	MRIC	100.00	100.00	100.00	100.00
3000	AIC/BIC	100.00	100.00	100.00	100.00
	GAIC	100.00	100.00	100.00	100.00
	GBIC	100.00	100.00	100.00	100.00
	GBICp	100.00	100.00	100.00	100.00
	MRIC	100.00	100.00	100.00	100.00

(S5.2) is nonlinear in the parameters. With observations up to time n, we are interested in predicting y_{n+2} using a model chosen from

$$J_1: \quad y_{n+2} = \frac{1}{1-\alpha B} x_n + \varepsilon_{n,2}^{(1)}(\alpha) \equiv g_{n,2}^{(1)}(\alpha) + \varepsilon_{n,2}^{(1)}(\alpha),$$
$$J_2: \quad y_{n+2} = \frac{1}{1-\beta B} z_n + \varepsilon_{n,2}^{(2)}(\beta) \equiv g_{n,2}^{(2)}(\beta) + \varepsilon_{n,2}^{(2)}(\beta),$$

both being misspecified. Since $g_{n,2}^{(1)}(\alpha)$ is independent of $\varepsilon_{n,2}^{(1)}(\alpha)$ and $g_{n,2}^{(2)}(\beta)$ is independent of $\varepsilon_{n,2}^{(2)}(\beta)$, J_1 and J_2 are said to be correct up to an independent additive error. In addition, since the initial conditions are set to $x_t = z_t = 0$ for $t < 0$, this example is classified as an asymptotic stationary case discussed at the end of Section S4.2. The parameters in (S5.2) are set to:

DGP I: $(a, b, \sigma_\varepsilon, \sigma_x, \sigma_z) = (0.5, \text{NA}, 1, 1, 1)$,
DGP II: $(a, b, \sigma_\varepsilon, \sigma_x, \sigma_z) = (0.95, 0.65, 1, 0.4109, 1.000)$,
DGP III: $(a, b, \sigma_\varepsilon, \sigma_x, \sigma_z) = (0.4, -0.95, 0.25, 1.4676, 0.5)$,
DGP IV: $(a, b, \sigma_\varepsilon, \sigma_x, \sigma_z) = (0.8, -0.4, 1, 1.3093, 2)$.

In DGP I, $b = \text{NA}$ represents that the true model depends on $\{x_t\}$ only. Let $\text{MI}_2(l)$ and $L_2(l)$ denote the MI and VI of J_l, $l = 1, 2$. It is shown in Table S4 that while $\text{MI}_2(1) < \text{MI}_2(2)$ under DGP I, the two candidates have the same MI for other DGPs, which is caused by $\sigma_x^2/(1-a^2) = \sigma_z^2/(1-b^2)$. Moreover, $L_2(1) < L_2(2)$ for DGPs II and III, but the opposite holds true for DGP IV. Consequently, J_2 is better than J_1 only under DGP IV. In Table S5, we present the performances, based on 1,000 simulations, of MRIC (with $\alpha_m = 0.8$) and the other five criteria described in Example 1. It is worth mentioning that since J_1 and J_2 are correct up to an independent additive error, the $\widehat{\mathbf{A}}^*(l)$ in the nonlinear version of MRIC defined in (S4.2) can be dropped from the formula. In addition, the $\widehat{\mathbf{H}}_h(l)$ in GAIC(l), GBIC(l) and GBIC$_p(l)$ is defined as in Example 1, except that $\widehat{\mathbf{R}}(l)$ and $\widehat{\mathbf{C}}_{h,0}(l)$ are replaced by $\widehat{\mathbf{R}}^*(l)$ and $\widehat{\mathbf{C}}^*_{h,0}(l)$, respectively. The sample size n is again set to $200, 500, 1000, 2000$ and 3000. Note first that since under DGP I, J_1 and J_2 have a substantial difference in MI, all six criteria work well for all sample sizes. However, the performance of these criteria notably deteriorates under DGPs II-IV, in which J_1 and J_2 have the same MI. In particular, all criteria, except for MRIC, can only select the better candidate between 42% and 58% of the time when $n \geq 500$, and the

TABLE S4
The values of $\mathrm{MI}_2(1) - \mathrm{MI}_2(2)$ *and* $L_2(1) - L_2(2)$ *in Example 2, and the corresponding better predictive models*

	DGPs			
	I	II	III	IV
$\mathrm{MI}_2(1) - \mathrm{MI}_2(2)$	-2.333	0.000	0.000	0.000
$L_2(1) - L_2(2)$	$*$	-0.759	-1.311	1.538
The better predictive model	J_1	J_1	J_1	J_2

*: $L_2(1) - L_2(2)$ can be neglected.

percentage seems to be indifferent to the sample size. On the other hand, MRIC tends to perform better with increasing number of data points. More specifically, under DGP II (III, IV), MRIC's percentage of identifying the better candidate increases from 46% (74%, 66%) to 64% (84%, 83%) as n rises from 200 to 3000. The α_m in MRIC is set to 0.8 instead of 0.6. This is because in the nonlinear case, a larger α_m is usually needed to secure a better selection result. For a further investigation on this example, see Example 5.

Example 3. In this example, we evaluate the performance of the high-dimensional model selection method, OGA+HDIC$_h$+Trim, developed in Section 4 using the following one-step predictive model,

$$y_{t+1} = \mathbf{x}_t^\top \boldsymbol{\beta} + x_{t,p+1} + \varepsilon_{t+1}. \tag{S5.3}$$

Here $\{\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})^\top\}$ is a sequence of p-dimensional i.i.d. standard normal random variates, p is allowed to be larger than n, $x_{t,p+1} = x_{t1}x_{t2}$ is the product of the first two regressor variables, $\boldsymbol{\beta}$ is a p-dimensional regression coefficient vector, and $\varepsilon_t = \phi_1\varepsilon_{t-1} + \eta_t$ is an AR(1) process in which $|\phi_1| < 1$ and $\{\eta_t\}$, independent of $\{\mathbf{x}_t\}$, is a sequence i.i.d. random errors with mean 0. Although the data are generated from model (S5.3), we fit a linear regression model without the interaction term $x_{t,p+1}$,

$$y_{t+1} = \mathbf{x}_t^\top \boldsymbol{\beta}^* + \varepsilon_{t+1}^*, \tag{S5.4}$$

which is misspecified in view of (S5.3). In the special case that $\phi_1 = 0$ and $4\eta_t$ is $N(0,1)$ distributed, (S5.3) and (S5.4) have been used in Example 5.1.2 of Lv and Liu (2014) to illustrate the advantage of GBIC$_p$ with respect to AIC, BIC, GAIC and GBIC when $p \geq n$. Following Lv and Liu (2014), we let $\boldsymbol{\beta} = (1, -1.25, 0.75, -0.95, 1.5, \mathbf{0}_{p-5}^\top)^\top$. On the other hand, to demonstrate the broad range of application of

TABLE S5

Percentage of times, across 1,000 simulations, that the better two-step ($h = 2$) predictive model between J_1 and J_2 of Example 2 is chosen

		DGPs			
n	Criteria	I	II	III	IV
200	AIC/BIC	100.00	39.40	60.20	52.50
	GAIC	100.00	39.30	60.80	52.80
	GBIC	100.00	39.20	60.30	51.90
	GBICp	100.00	39.30	60.50	51.90
	MRIC	100.00	46.20	73.50	65.90
500	AIC/BIC	100.00	42.20	58.00	53.10
	GAIC	100.00	42.50	58.30	53.60
	GBIC	100.00	42.10	58.20	52.80
	GBICp	100.00	42.20	58.40	52.90
	MRIC	100.00	51.70	75.10	72.90
1000	AIC/BIC	100.00	45.70	52.00	53.10
	GAIC	100.00	46.10	52.30	53.50
	GBIC	100.00	45.70	52.00	52.20
	GBICp	100.00	45.80	52.20	52.40
	MRIC	100.00	56.10	76.70	78.60
2000	AIC/BIC	100.00	48.50	54.50	51.90
	GAIC	100.00	48.60	54.60	52.20
	GBIC	100.00	48.50	54.50	51.50
	GBICp	100.00	48.60	54.60	51.50
	MRIC	100.00	61.90	81.50	81.20
3000	AIC/BIC	100.00	48.10	54.70	50.70
	GAIC	100.00	48.20	54.80	50.90
	GBIC	100.00	48.10	54.70	50.70
	GBICp	100.00	48.10	54.80	50.70
	MRIC	100.00	63.90	83.50	82.60

OGA+HDIC$_h$+Trim, we set $\phi_1 = 0.8$ and let $4\eta_t$ follow a t-distribution with 8 degrees of freedom (denoted as t_8) in addition to the $N(0,1)$ distribution. Moreover, the (n,p) combinations considered in this example are $\{200,1000\} \times \{100,200,1000\}$, which include both the low-dimensional case ($p \le n$) and the high-dimensional case ($p > n$).

Due to $\min\{q_1,q_2\} < 8$ in the case of $4\eta_t \sim t_8$ (noting that q_1 and q_2 are defined in (F1) and (F2) of Section 4, respectively), $2/q$ in HDIC$_h$ is set to 0.3 (denoted by HDIC$_h$(0.3)) or 0.4 (denoted by HDIC$_h$(0.4)) in our simulation study. In addition, we let $\omega_n = \log n$ and K_n, the maximum number of the OGA iterations, equal $\min\{p, \overline{m}_n\}$, where $\overline{m}_n = 5n^{1/2}/p^{1/4}$. (Note that our (unreported) simulation studies indicate that varying 5 in $\overline{m}_n$ from 3 to 10 leads to similar conclusions.) Since in this example $\theta_1 = \theta_2 = 0$, where θ_1, θ_2 are defined in (F5) of Section 4, it is not difficult to see that the above specifications on OGA and HDIC$_h$ fulfill (4.11) and (4.12), which are required to establish the selection consistency of OGA+HDIC$_h$+Trim; see Theorem 4.2. A data-driven method for determining $2/q$ is provided in Section S6.

For the purpose of comparison, we also evaluate the performance of GBIC$_p$, AIC, BIC, GAIC and GBIC in the same simulation setting. Since it is unrealistic to implement best subset regression due to $p \ge 100$, these five criteria are used to select models along the OGA path, which can substantially relieve computational burden. Note that Theorem 4.1 ensures that OGA can asymptotically include $J^* = \{1,\ldots,5\}$, the most parsimonious model among those having the lowest MI value.

We evaluate the performance of a model selection criterion (which selects variable set $\hat{J}^{(i)} \subset \{1,\ldots,p\}$ in the ith simulation) using three different measures,

$$\text{expected number of true positives (ENTP)} : \frac{1}{1000}\sum_{i=1}^{1000} \sharp(\hat{J}^{(i)} \bigcap J^*),$$

$$\text{expected number of false positives (ENFP)} : \frac{1}{1000}\sum_{i=1}^{1000} \sharp(\hat{J}^{(i)} \bigcap J^{*^c}),$$

$$\text{selection probability (SP)} : \frac{\sum_{i=1}^{1000} I_{\{\hat{J}^{(i)}=J^*\}}}{1000},$$

where $J^{*^c} = \{1,\ldots,p\} - J^*$. We summarize in Table S6 the performance of the seven criteria mentioned in previous paragraphs in Table S6. As observed, all criteria have ENTP values equal to $5 = \sharp(J^*)$,

except for the case of $(n, p) = (200, 1000)$ in which the ENTP values of OGA+HDIC$_h(0.4)$+Trim fall between 4.998 and 4.999. This result not only suggests that OGA is a very powerful tool for including relevant variables in finite samples, but also reveals that all criteria, in conjunction with OGA, enjoy high true positive rates. In addition, all ENFP and SP values of OGA+HDIC$_h(r)$+Trim, with $r = 0.3$ and 0.4, are close to 0 and 1, respectively. These findings indicate that the model selection method proposed in Section 4 is not only reliable in both low- and high-dimensional cases, but also robust to the tail behavior of the error distribution. On the other hand, while the overfitting problem associated with OGA+BIC-type criteria (OGA+BIC, OGA+GBIC and OGA+GBIC$_p$) is not severe in terms of ENFP when $p \leq 200$, their SP values are substantially smaller than 1. Moreover, suffering from considerable overfitting, OGA+BIC-type criteria in the case of $p = 1000$ and OGA+AIC-type criteria (OGA+AIC and OGA+GAIC) in all cases have SP values very close to 0.

Example 4. In this example, we demonstrate the performance of OGA+HDIC$_{h,l}$+Trim+MRIC under high-dimensional misspecified models. We first generate $(y_i, z_i, w_i), i = 1, \ldots, n$, according to model (S5.1), with $\beta_1 = \beta_2 = 6$, $(\theta_1, \theta_2) = (0.8, -0.95)$, $\phi = \theta_1/(1 - \theta_2)$, and ε_t being i.i.d. random variables from a t_8 distribution. Suppose that there are two forecasters, F1 and F2, who are interested in predicting y_{n+h}, with $h = 2, 3$, using the following high-dimensional models,

$$\text{F1}: \quad y_{n+h} = \sum_{j=1}^{p_1} \beta_{j,h}^{(1)} x_{n,j}^{(1)} + \varepsilon_{n,h}^{(1)} = \beta_{1,h}^{(1)} x_{n,1}^{(1)} + \mathbf{x}_{t,-1}^{\top} \boldsymbol{\beta}_{-1} + \varepsilon_{n,h}^{(1)},$$

$$\text{F2}: \quad y_{n+h} = \sum_{j=1}^{p_2} \beta_{j,h}^{(2)} x_{n,j}^{(2)} + \varepsilon_{n,h}^{(2)} = \beta_{1,h}^{(2)} x_{n,1}^{(2)} + \mathbf{x}_{t,-2}^{\top} \boldsymbol{\beta}_{-2} + \varepsilon_{n,h}^{(2)},$$

where $p_1 > n, p_2 > n$, $x_{t,1}^{(1)} = z_t$, $x_{t,1}^{(2)} = w_t$, $\mathbf{x}_{t,-l} \sim \text{N}(\mathbf{0}, 0.25\text{I}_{p_l - 1}), l = 1, 2$, and $\{\mathbf{x}_{t,-1}\}$, $\{\mathbf{x}_{t,-2}\}$, $\{z_t\}$, $\{w_t\}$, and $\{\varepsilon_t\}$ are independent. Recall the definitions of $N_h^{(l)}$, $\text{MI}_{h,l}(N_h^{(l)})$, $L_{h,l}(N_h^{(l)})$, $M_{A,h}$, $M_{B,h}$, and $M_{C,h}$ given in Section 4. Since $\{\mathbf{x}_{t,-l}, l = 1, 2\}$, are independent of $\{z_t, w_t, \varepsilon_t\}$, it is easy to see that for $h = 2$ and 3, $N_h^{(1)} = \{1\}$ and $N_h^{(2)} = \{1\}$. Straightforward calculations give $\text{MI}_{2,1}(N_2^{(1)})$=$\text{MI}_{2,2}(N_2^{(2)})$ and $L_{2,1}(N_2^{(1)}) - L_{2,2}(N_2^{(2)}) = 37.8$, yielding $M_{A,2} = \{1, 2\}$, $M_{B,2} = \{2\}$,

TABLE S6

Expected numbers of true positives (ENTP), expected numbers of false positives (ENFP), and selection probabilities (SP), across 1,000 simulations, of the model selection criteria in Example 3 under the misspecified model (S5.4)

		True Model: (S5.3) with $\varepsilon_t = 0.8\varepsilon_{t-1} + \eta_t$ and $4\eta_t$ following a $N(0,1)$ distribution								
		ENTP	ENFP	SP	ENTP	ENFP	SP	ENTP	ENFP	SP
n	Criteria	p=100			p=200			p=1000		
200	OGA+AIC	5.000	17.425	0.000	5.000	14.000	0.000	5.000	8.000	0.000
	OGA+BIC	5.000	2.903	0.074	5.000	8.014	0.004	5.000	8.000	0.000
	OGA+GAIC	5.000	17.454	0.000	5.000	13.999	0.000	5.000	8.000	0.000
	OGA+GBIC	5.000	2.632	0.093	5.000	6.947	0.009	5.000	8.000	0.000
	OGA+GBICp	5.000	1.563	0.246	5.000	3.715	0.051	5.000	8.000	0.000
	OGA+HDIC$_h$(0.3)+Trim	5.000	0.043	0.958	5.000	0.032	0.970	5.000	0.016	0.986
	OGA+HDIC$_h$(0.4)+Trim	5.000	0.006	0.994	5.000	0.002	0.999	4.998	0.002	0.996
1000	OGA+AIC	5.000	16.262	0.000	5.000	34.460	0.000	5.000	24.000	0.000
	OGA+BIC	5.000	0.863	0.437	5.000	1.830	0.166	5.000	12.116	0.000
	OGA+GAIC	5.000	16.524	0.000	5.000	34.454	0.000	5.000	24.000	0.000
	OGA+GBIC	5.000	0.857	0.439	5.000	1.799	0.163	5.000	11.483	0.000
	OGA+GBICp	5.000	0.517	0.604	5.000	1.024	0.361	5.000	6.164	0.006
	OGA+HDIC$_h$(0.3)+Trim	5.000	0.000	1.000	5.000	0.000	1.000	5.000	0.000	1.000
	OGA+HDIC$_h$(0.4)+Trim	5.000	0.000	1.000	5.000	0.000	1.000	5.000	0.000	1.000
		True Model: (S5.3) with $\varepsilon_t = 0.8\varepsilon_{t-1} + \eta_t$ and $4\eta_t$ following a t_8 distribution								
		ENTP	ENFP	SP	ENTP	ENFP	SP	ENTP	ENFP	SP
n	Criteria	p=100			p=200			p=1000		
200	OGA+AIC	5.000	17.486	0.000	5.000	14.000	0.000	5.000	8.000	0.000
	OGA+BIC	5.000	3.041	0.084	5.000	7.920	0.000	5.000	8.000	0.000
	OGA+GAIC	5.000	17.478	0.000	5.000	14.000	0.000	5.000	8.000	0.000
	OGA+GBIC	5.000	2.729	0.115	5.000	6.977	0.005	5.000	8.000	0.000
	OGA+GBICp	5.000	1.630	0.259	5.000	3.886	0.051	5.000	7.997	0.000
	OGA+HDIC$_h$(0.3)+Trim	5.000	0.037	0.965	5.000	0.033	0.968	5.000	0.014	0.987
	OGA+HDIC$_h$(0.4)+Trim	5.000	0.005	0.995	5.000	0.003	0.997	4.999	0.001	0.998
1000	OGA+AIC	5.000	16.062	0.000	5.000	34.207	0.000	5.000	24.000	0.000
	OGA+BIC	5.000	0.791	0.461	5.000	1.747	0.192	5.000	12.106	0.000
	OGA+GAIC	5.000	16.298	0.000	5.000	34.356	0.000	5.000	24.000	0.000
	OGA+GBIC	5.000	0.782	0.464	5.000	1.715	0.198	5.000	11.535	0.000
	OGA+GBICp	5.000	0.457	0.629	5.000	0.969	0.387	5.000	6.080	0.006
	OGA+HDIC$_h$(0.3)+Trim	5.000	0.000	1.000	5.000	0.000	1.000	5.000	0.000	1.000
	OGA+HDIC$_h$(0.4)+Trim	5.000	0.000	1.000	5.000	0.000	1.000	5.000	0.000	1.000

Note: all values are rounded off to the nearest thousandths.

TABLE S7
Percentage of times, across 1,000 simulations, that $M_{C,h}$ in Example 4 is chosen

Criteria	h=2		h=3	
	$n = 200$	$n = 500$	$n = 200$	$n = 500$
OGA+HDIC$_h$(0.3)+Trim+AIC	47.90	48.40	99.30	100.00
OGA+HDIC$_h$(0.3)+Trim+BIC	47.90	48.40	99.30	100.00
OGA+HDIC$_h$(0.3)+Trim+GAIC	48.30	48.40	99.30	100.00
OGA+HDIC$_h$(0.3)+Trim+GBIC	47.60	48.20	99.30	100.00
OGA+HDIC$_h$(0.3)+Trim+GBIC$_p$	47.80	48.40	99.30	100.00
OGA+HDIC$_h$(0.3)+Trim+MRIC	66.60	76.60	98.50	100.00

and $M_{C,2} = \{(2, N_2^{(2)})\}$. Moreover, $\text{MI}_{3,1}(N_3^{(1)}) - \text{MI}_{3,2}(N_3^{(2)}) = 12.899$, implying $M_{A,3} = \{2\}$, $M_{B,3} = \{2\}$, and $M_{C,3} = \{(2, N_3^{(2)})\}$. In other words, the subset of model F2, which contains one variable $x_{n,1}^{(2)}$ only, is the asymptotically best model for predicting y_{n+2} and y_{n+3}. In Table S7, the percentage of OGA+HDIC$_{h,l}$+Trim+MRIC choosing $M_{C,h}$, with $h = 2, 3$, is reported by using 1,000 simulations for $n = 200, 500$ and $p_1 = p_2 = p = 1001$. In view of Example 3, the tuning parameters in OGA and HDIC$_{h,l}$ are given by $K_n^{(l)} = \min\{p, 5n^{1/2}/p^{1/4}\}$, $\omega_n^{(l)} = \log n$, and $2/q_l = 0.3$ for $l = 1, 2$. Following Example 1, the α_m in MRIC$_{h,l}$ is set to 0.6 for $l = 1, 2$. We also evaluate the performance of OGA+HDIC$_{h,l}$+Trim+AIC (BIC, GAIC, GBIC, or GBIC$_p$) in choosing $M_{C,h}$, with $h = 2, 3$, and record the results in Table S7. As shown in the table, all criteria perform satisfactorily in the case of $h = 3$ even when $n = 200$. This is not only because OGA+HDIC$_{3,l}$+Trim can consistently select $N_3^{(l)}, l = 1, 2$, but because the notable difference between $\text{MI}_{3,1}(N_3^{(1)})$ and $\text{MI}_{3,2}(N_3^{(2)})$ makes all criteria easy to identify the better predictive model. On the other hand, while in the case of $h = 2$, the percentage of OGA+HDIC$_{2,l}$+Trim selecting $N_2^{(l)}, l = 1, 2$, is still very high, all criteria except for OGA+HDIC$_{2,l}$+Trim+MRIC have only about a 50% chance of choosing $M_{C,2}$ due to $\text{MI}_{2,1}(N_2^{(1)})$=$\text{MI}_{2,2}(N_2^{(2)})$ and $\sharp(N_2^{(1)}) = \sharp(N_2^{(2)}) = 1$. In contrast, OGA+HDIC$_{2,l}$+Trim+MRIC has a 66% percent chance of choosing $M_{C,2}$ when $n = 200$, and the percentage grows to 76% when n increases to 500.

Example 5. Suppose that data are generated from model (S5.2) and the candidate models used for predicting y_{n+2} are J_1 and

$$J_2' : \qquad y_{n+2} = \frac{r_2}{1 - \beta B} z_n + \varepsilon_{n,2}^{(2)}(r_2, \beta).$$

Unlike J_2 in Example 2, J_2' does not assume that prior information about r_2 is available, and hence the parameter needs to be estimated

from the data. It can be show that

(S5.5)
$$\mathrm{MI}_2(1) = \sigma_z^2/(1-b^2) + \sigma_\varepsilon^2 \text{ and } \mathrm{MI}_2^*(2) = \sigma_x^2/(1-a^2) + \sigma_\varepsilon^2 = \mathrm{MI}_2(2),$$

where $\mathrm{MI}_2^*(2)$ is the MI for $J_2^{'}$. Moreover, we have

$$L_2(1) = \frac{\sigma_z^2}{1-b^2} + \sigma_\varepsilon^2 + \frac{4ab\sigma_z^2}{(1+a^2)(1-b^2)},$$

and

$$L_2^*(2) = 2\Big(\frac{\sigma_x^2}{1-a^2} + \sigma_\varepsilon^2\Big) + \frac{4ab\sigma_x^2}{1-a^2}. \tag{S5.6}$$

Therefore, when $\mathrm{MI}_2(1)$=$\mathrm{MI}_2^*(2)$ (i.e., $\sigma_z^2/(1-b^2) = \sigma_x^2/(1-a^2)$), $L_2(1)$ is always smaller than $L_2^*(2)$. This is in sharp contrast to the situation described in Example 2, where $L_2(1)$ can sometimes be greater than

$$L_2(2) = \frac{\sigma_x^2}{1-a^2} + \sigma_\varepsilon^2 + \frac{4ab\sigma_x^2}{(1+b^2)(1-a^2)}$$

under $\sigma_z^2/(1-b^2) = \sigma_x^2/(1-a^2)$.

On the other hand, if J_1 is replaced by

$$J_1^{'}: \qquad y_{n+2} = \frac{r_1}{1-\alpha B} x_n + \varepsilon_{n,2}^{(1)}(r_1, \alpha),$$

where both r_1 and α are unknown, then

$$\mathrm{MI}_2^*(1) = \sigma_z^2/(1-b^2) + \sigma_\varepsilon^2 \tag{S5.7}$$

and

$$L_2^*(1) = 2\Big(\frac{\sigma_z^2}{1-b^2} + \sigma_\varepsilon^2\Big) + \frac{4ab\sigma_z^2}{1-b^2}, \tag{S5.8}$$

where $\mathrm{MI}_2^*(1)$ and $L_2^*(1)$ are the MI and VI for $J_1^{'}$. In view of (S5.5)–(S5.8), $J_1^{'}$ and $J_2^{'}$ share the same VI value, provided their MI values are the same.

In Table S8, we present the performances, based on 1,000 simulations, of the six criteria described in Example 2 when they are used to choose between J_1 and $J_2^{'}$ and between $J_1^{'}$ and $J_2^{'}$ under

DGP V: (a, b, $\sigma_\varepsilon, \sigma_x, \sigma_z$)=(-0.2,0.7,0.25, 1.372,1),

yielding $\mathrm{MI}_2(1)$=$\mathrm{MI}_2^*(1)$=$\mathrm{MI}_2^*(2)$ = 2.0233, $L_2(1) - L_2^*(2) = -1.9811$, and $L_2^*(1) - L_2^*(2) = 0$. It is shown in the upper panel of Table S8 that MRIC can successfully identify the better model, J_1, between J_1 and $J_2^{'}$, and its finite sample performance under DGP V is even better than under DGPs III and IV. On the other hand, all other criteria do not possess this advantage due to $\mathrm{MI}_2(1)$=$\mathrm{MI}_2^*(2)$. The lower panel of Table S8 reveals that all six criteria tend to randomly choose between $J_1^{'}$ and $J_2^{'}$. This result, however, should not be an overriding concern because $J_1^{'}$ and $J_2^{'}$ are indistinguishable in terms of both MI and VI.

TABLE S8

Percentage of times, across 1,000 simulations, that the better two-step ($h = 2$) predictive model between J_1 and $J_2^{'}$ (or between $J_1^{'}$ and $J_2^{'}$) is chosen under DGP V in Example 5

		J_1 vs. $J_2^{'}$					
h	n	AIC	BIC	GAIC	GBIC	GBICp	MRIC
2	200	52.20	55.60	52.20	55.60	56.90	89.80
2	500	52.90	55.30	53.00	55.40	55.70	95.10
2	1000	48.80	50.50	48.70	50.50	51.00	98.60
2	2000	50.80	52.90	50.70	52.90	53.10	99.60
2	3000	49.70	51.30	49.70	51.30	51.80	99.90
		$J_1^{'}$ vs. $J_2^{'}$					
h	n	AIC	BIC	GAIC	GBIC	GBICp	MRIC
2	200	52.40	52.40	52.50	52.50	52.40	51.90
2	500	51.40	51.40	51.60	51.50	51.40	50.50
2	1000	49.30	49.30	49.30	49.30	49.30	49.10
2	2000	50.50	50.50	50.50	50.50	50.50	50.60
2	3000	49.20	49.20	49.20	49.20	49.20	49.50

S6. Real Data Analysis: Two Cases. In this section, we compare the performance of the criteria discussed in Section S5 using two real datasets. The first dataset is the monthly life insurance data recording the net number of new personal life insurances for a large insurance company from January 1964 to December 1980; see Claeskens et al. (2007) for more details. Following Claeskens et al. (2007), we took the first and the seasonal differences of the log-transformed data to get a (possibly) stationary series; see Figure S1 for the time plot as well as the sample ACF/PACF plot of the resultant series, denoted by $\{S_t\}, 1 \le t \le 191$. The goal of this study is to investigate the prediction performance of the criteria considered in Example 1 of Section S5 when they are applied to $\{S_t\}$. For the sake of completeness, our assessment also includes FIC (Claeskens et al., 2007), whose performance on $\{S_t\}$ has been explored in the same paper. Specifying the candidate models as AR(1), ..., AR(15) and retaining the latest $\lfloor nd \rfloor$ observations in $\{S_t\}$

for performance evaluation, we measure the prediction capability of a criterion by the empirical MSPE (EMSPE),

$$\text{EMSPE} = \frac{1}{\lfloor nd \rfloor} \sum_{t=n-h-\lfloor nd \rfloor+1}^{n-h} (S_{t+h} - \hat{S}_{t+h})^2, \tag{S6.1}$$

where $d = 0.3$, $\hat{S}_{t+h}$ is the predictor of S_{t+h} whose order is selected by the criteria and parameters estimated by least squares using observations up to time t. In this connection, we also compute

$$\text{EMSPE}_0 = \min_{1 \le k \le 15} \frac{1}{\lfloor nd \rfloor} \sum_{t=n-h-\lfloor nd \rfloor+1}^{n-h} (S_{t+h} - \hat{S}_{t+h}(k))^2, \tag{S6.2}$$

where $\hat{S}_{t+h}(k)$ is the least squares predictor of S_{t+h} whose order is fixed at k and parameters are estimated by least squares using observations up to time t. Note that EMSPE_0 serves as a convenient benchmark for comparing the EMSPEs derived from different criteria. Note also that the α_m in the $C_n = n^{\alpha_m}$ of MRIC is chosen by minimizing the in-sample counterpart of (S6.1),

$$\frac{1}{\lfloor nd \rfloor} \sum_{t=n-2\lfloor nd \rfloor-h+1}^{n-\lfloor nd \rfloor-h} (S_{t+h} - \hat{S}_{t+h}^{(\alpha_m)})^2, \tag{S6.3}$$

over $\alpha_m \in \{0.1, \ldots, 0.8\}$, where $\hat{S}_{t+h}^{(\alpha_m)}$ is $\hat{S}_{t+h}$ with order selected by MRIC using penalty of $C_n = n^{\alpha_m}$. Since the candidate models are nested, any $\alpha_m \in \{0.1, \ldots, 0.8\}$ leads to an asymptotically efficient MRIC, in view of Remark 2. For the sake of convenience, once an α_m is determined by (S6.3), it will be used throughout the period for forecast evaluation.

The values obtained from (S6.1) and (S6.2), with $h = 1, \ldots, 5$, are summarized in Table S9. As shown in Table S9, MRIC appears to perform favorably compared to all other criteria. In particular, its EMSPE values are almost identical to the values of EMSPE_0 for all $h = 1, \ldots, 5$. The performance of FIC, AIC and GAIC is also reasonably good. The EMSPE of FIC is even a little bit smaller than EMSPE_0 in the case of $h = 2$ and 3. However, FIC may seem inferior to MRIC, AIC and GAIC when $h = 4$ and 5. AIC and GAIC have performance close to that of MRIC, but their EMSPE values are either equal or greater than

MRIC's. All BIC-type criteria, BIC, GBIC and GBIC$_p$, suffer from relatively large EMSPE values, and hence are surpassed by the former four criteria. Finally, we remark that our conclusion on FIC, AIC and BIC is not necessarily coincident with the one provided by Claeskens et al. (2007). This may be due to fact that the performance measure used by the latter paper is EMSPE with d close to 0.5 instead of 0.3.

The second dataset contains three weakly time series of length $n = 508$ for cardiovascular mortality (M_t), temperature (T_t) and particulate pollution (P_t) in Los Angeles County over the 10 year period 1970-1979; see Shumway et al. (1988) or Example 2.2 of Shumway and Stoffer (2011) for details. The time series plots shown in Figure S2 of Shumway and Stoffer (2011) reveal that there are strong *contemporaneous* co-movements between these series. These authors therefore built the following model to describe the effects of T_t and P_t on M_t,

$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - \bar{T}) + \beta_3(T_t - \bar{T})^2 + \beta_4 P_t + w_t, \tag{S6.4}$$

where $\{w_t\}$ is a stationary AR(2) model and $\bar{T}$ is the sample mean of $\{T_t\}$. However, it seems difficult to use (S6.4) to predict M_{t+h} when its contemporaneous explanatory variables, T_{t+h} and P_{t+h}, are not available. To bypass this dilemma, we devise a (purely) predictive model,

$$\begin{aligned} M_{t+h} &= \beta_0 + \beta_1(t+h) + \sum_{i=1}^{L} \beta_{3,i} M_{t+1-i} \\ &+ \sum_{i=1}^{L} \beta_{4,i} T_{t+1-i} + \sum_{i=1}^{L} \beta_{5,i} T_{t+1-i}^2 + \sum_{i=1}^{L} \beta_{6,i} P_{t+1-i} + \sum_{i=1}^{L} \beta_{7,i} \log P_{t+1-i} \\ &+ \epsilon_{t,h}, t = L, \ldots, n-h, \end{aligned} \tag{S6.5}$$

where $\epsilon_{t,h}$ denotes the error term. In this study, L is set to 156, namely, all dependent variables lagged up to three years are included. The reason why we adopt so many lagged variables is that the sample ACFs of $\{M_t\}, \{T_t\}$ and $\{P_t\}$ are still significantly bounded away from 0 even after lag 150; see Figure S2. We also include $\log P_t$ and its lagged values because $\log P_t$ has been used by Shumway et al. (1988) as an explanatory variable for M_t. Due to the inclusion of the lagged variables and the retention of the latest $O_1 = 35$ observations for forecast evaluation, the sample size for model selection is reduced to $N_1 = n-h-L+1-O_1$. On the other hand, the number of candidate variables in model (S6.5)

is $p_1 = 5L + 1 = 781$, noting that the intercept β_0 is always included in our study. Because p_1 is much greater than N_1, following Example 3 of Section S5, we first use OGA to sequentially select $K_n = 5N_1^{1/2}/p_1^{1/4}$ variables, and then choose models along the OGA path using the criteria considered in the same example. Their performance is evaluated by

$$\widetilde{\text{EMSPE}} = \frac{1}{O_1} \sum_{t=n-h-O_1+1}^{n-h} (M_{t+h} - \hat{M}_{t+h})^2, \tag{S6.6}$$

where $\hat{M}_{t+h}$ is the least squares predictor of M_{t+h} based on the model selected at time $n-h-O_1+1$ and the parameters estimated at time t. Although the estimates of the unknown parameters are continuously updated throughout the period of forecast evaluation, we choose not to update the model once it is determined at time $n-h-O_1+1$ because O_1 is relatively small compared to n. For HDIC_h, ω_n is still given by $\log n$ as in Example 3 of Section S5. On the other hand, instead of setting $2/q = 0.3$ or 0.4, we choose $2/q$ to minimize

$$\frac{1}{O_1} \sum_{t=n-h-2O_1+1}^{n-h-O_1} (M_{t+h} - \hat{M}_{t+h}^{(2/q)})^2,$$

over $2/q \in \{0.3, 0.4, \ldots, 0.9\}$, where $\hat{M}_{t+h}^{(2/q)}$ is the predictor of M_{t+h} based on the model selected by OGA+$\text{HDIC}_h(2/q)$+Trim at time $n-h-2O_1+1$ and the parameters estimated least squares at time t. The values of $\widetilde{\text{EMSPE}}$, with $1 \le h \le 5$, are documented in Table S10. For the purpose of comparison, we also compute the following benchmark value,

$$\widetilde{\text{EMSPE}}_0 = \min_{1 \le k \le K_n} \frac{1}{O_1} \sum_{t=n-h-O_1+1}^{n-h} (M_{t+h} - \hat{M}_{t+h}(k))^2, \tag{S6.7}$$

where $\hat{M}_{t+h}(k)$ is the h-step least squares predictor of M_{t+h} based on the model determined by the first k OGA iterations at time $n-h-O_1+1$ and the parameters estimated at time t.

As shown in Table S10, the performance of OGA+AIC, OGA+GAIC, and OGA+BIC is exactly the same in terms of $\widetilde{\text{EMSPE}}$. In addition, the $\widetilde{\text{EMSPE}}$ value of OGA+GBIC is smaller than (the same as) that

TABLE S9
The values of EMSPE and EMSPE$_0$ derived from series $\{S_t\}$.

	EMSPE							EMSPE$_0$
h	AIC	BIC	FIC	MRIC	GAIC	GBIC	GBIC$_p$	
1	0.0409	0.0533	0.0395	0.0393	0.0393	0.0533	0.0533	0.0393
2	0.0609	0.0756	0.0577	0.0594	0.0598	0.0756	0.0756	0.0593
3	0.0586	0.0764	0.0569	0.0575	0.0580	0.0763	0.0763	0.0574
4	0.0599	0.0817	0.0623	0.0589	0.0589	0.0817	0.0817	0.0589
5	0.0583	0.0815	0.0654	0.0583	0.0595	0.0815	0.0815	0.0583

of OGA+AIC when $h = 1$ ($h > 1$). Except for the case of $h = 1$, OGA+HDIC$_h$+Trim obviously outperforms the other criteria, and the difference between its $\widetilde{\text{EMSPE}}$ and the corresponding benchmark value, $\widetilde{\text{EMSPE}}_0$, does not seem to be sizeable. When $h = 1$, OGA+GBIC$_p$ performs best among all criteria. For $1 < h \leq 5$, its performance generally lies between OGA+HDIC$_h$+Trim and the other four non-HDIC$_h$ criteria.

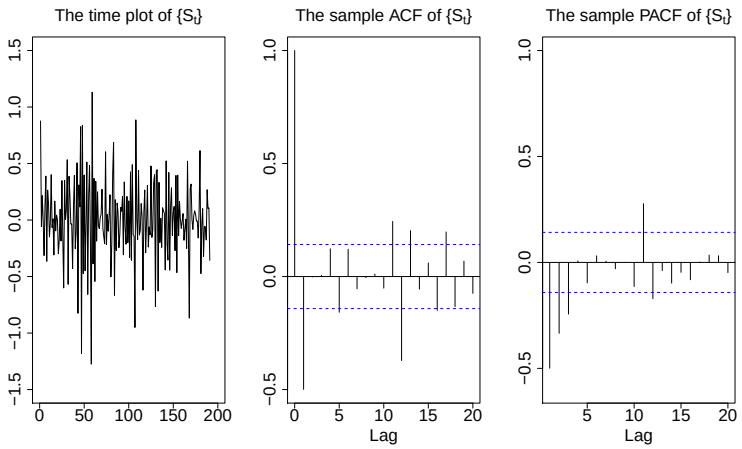

FIG S1. *Plots of series $\{S_t\}$ and its sample ACF and PACF*

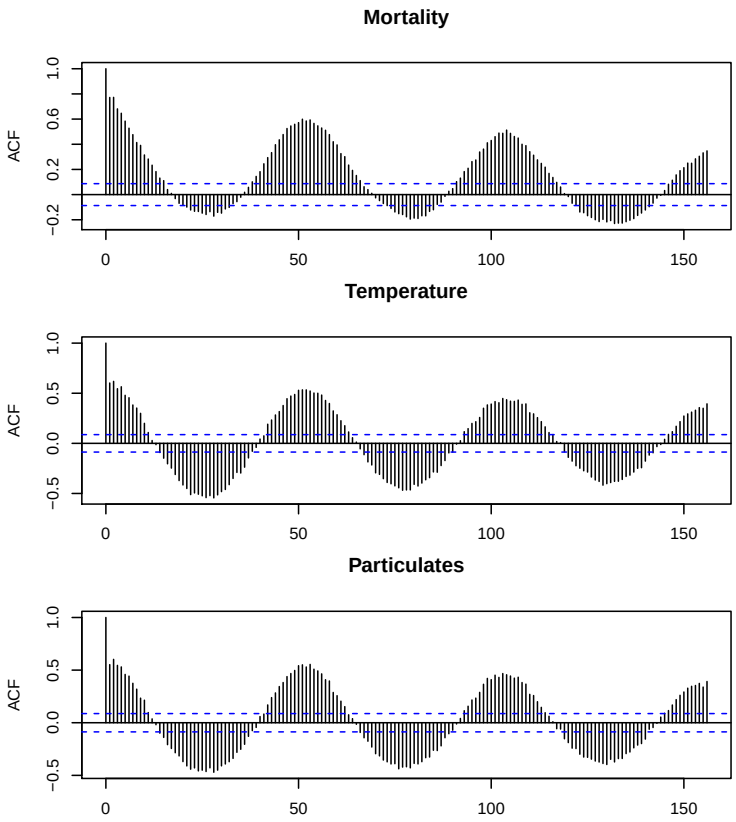

FIG S2. *Sample ACFs of weakly time series for cardiovascular mortality (top), temperature (middle) and particulate pollution (bottom) in Los Angeles County from 1970-1979.*

References.

Chan, N. H., and Ing, C.-K. (2011). Uniform moment bounds of Fisher's information with applications to time series. *Ann. Statist.*, **39**, 1526–1550.

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Ann. Statist.*, **21**, 520–533.

Chi, C.-M., Ing, C.-K., and Tong, H. (2017). Mean squared prediction error of a threshold autoregressive model. *Manuscript*.

Claeskens, G., Croux, C., and Kerckhoven, J. V. (2007). Perdiction-focused model selection for autoregressive models. *Aust. N. Z. J. Stat.*, **49**, 359–379.

Findley, D. F., and Wei, C. Z. (1993). Moment bounds for deriving time series CLT's and model selection procedures. *Statist. Sinica*, **3**, 453–480.

Hsu, H.-L., Ing, C.-K., and Tong, H. (2018). On model selection from a finite family of possibly misspecified time series models, accepted by *Ann. Statist.*.

Inoue, A. and Kilian, L. (2006). On the selection of forecasting models. *J. Econometrics*, **130**, 273–306.

TABLE S10
The values of $\widetilde{EMSPE}$ and $\widetilde{EMSPE}_0$ derived from the mortality data of Shumway et al. (1988).

	$\widetilde{\text{EMSPE}}$						$\widetilde{\text{EMSPE}}_0$
h	OGA+AIC	OGA+BIC	OGA+HDIC$_h$+Trim	OGA+GAIC	OGA+GBIC	OGA+GBIC$_p$	
1	20.75	20.75	18.99	20.75	18.38	17.60	16.79
2	26.63	26.63	21.89	26.63	26.63	26.77	16.68
3	36.02	36.02	22.90	36.02	36.02	31.29	16.54
4	27.51	27.51	23.45	27.51	27.51	24.08	16.31
5	31.48	31.48	24.22	31.48	31.48	31.48	16.55

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **76**, 141–167.

Shumway, R. H., Azari, A. S. and Pawitan, Y. (1988). Modeling mortality uctuations in Los Angeles as functions of pollution and weather effects. *Environ. Res.*, **45**, 224–241.

Shumway, R. H. and Stoffer, D. S. (2011). *Time series analysis and its applications: with R examples* (3rd ed.), New York: Springer.

Sin, C. Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *J. Econometrics*, **71**, 207–225.

Temlyakov, V. N. (2000). Weak greedy algorithms. *Adv. Comput. Math.*, **12**, 213–227.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc.*, **76**, 419–433.

White, H. (1984). Nonlinear regression with dependent observations. *Econometrica*, **52**, 143–162.