The Annals of Statistics
2019, Vol. 47, No. 2, 1061–1087
https://doi.org/10.1214/18-AOS1706

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

BY HSIANG-LING HSU[*,1], CHING-KANG ING[†,2] AND HOWELL TONG[‡,§,3]

National University of Kaohsiung[*]*, National Tsing Hua University*[†]*, University of Electronic Science & Technology*[‡] *and London School of Economics*[§]

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 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,

Received November 2017; revised March 2018.

[1]Supported in part by the Ministry of Science and Technology of Taiwan Grants MOST 103-2118-M-390-004-MY2.

[2]Supported in part by the Science Vanguard Research Program of the Ministry of Science and Technology, Taiwan.

[3]Supported in part by a special research grant, University of Electronic Science and Technology, Chengdu, China.

MSC2010 subject classifications. Primary 63M30; secondary 62F07, 62F12.

Key words and phrases. AIC, BIC, misspecification-resistant information criterion, multistep prediction error, high-dimensional misspecified models, orthogonal greedy algorithm.

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, for example, 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, Grünwald and de Rooij (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

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

Criteria	Case I:	Case II:	Case III:
	The true model is included as a candidate. Goal: Consistency	**The true model is NOT included as a candidate. Goal: Asymp. efficiency for prediction.**	**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

that Akaike was originally trying to solve. He said (Akaike (1978), page 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, for example, 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 S5 of the Supplementary Material for this paper [Hsu, Ing and Tong (2019)]. 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.

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

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 investigated 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 the Appendix. All proofs and an extension of MRIC to a class of nonlinear models are relegated to Hsu, Ing and Tong (2019). 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, Ing and Tong (2019).

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 the Appendix.

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

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

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,

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

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,

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

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 $\{\mathbf{x}_k, k \leq t\}$. 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 S1 of the Supplementary Material Hsu, Ing and Tong (2019).

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 of 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 $1 \le i, j \le m$,

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

(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 $\mathbf{f} = (f_1, \ldots, f_m)^\top$, $\|\mathbf{f}\|^2 = \sum_{t=1}^m f_t^2$.

(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 $q > 0$,

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

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

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

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

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 nonrandom 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

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

$$\begin{aligned} &\mathrm{E}\big(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\top|\mathcal{G}_{t-1}\big) = \boldsymbol{\Sigma} \quad \text{and} \\ &\sup_{-\infty<t<\infty} \mathrm{E}\big(\|\boldsymbol{\epsilon}_t\|^{q^*}|\mathcal{G}_{t-1}\big) < C_{q^*} \qquad \text{with probability 1,} \end{aligned} \tag{2.18}$$

where $\boldsymbol{\Sigma}$ is a positive definite nonrandom 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 $\boldsymbol{\Sigma}$. 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

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

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 $\|\mathbf{v}\| = 1$,

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

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, Ing and Tong (2019) 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,

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

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

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

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

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

and

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

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

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

where

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

with

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

A model selection criterion is said to be asymptotically efficient if (3.6) is fulfilled. Section S5 of Hsu, Ing and Tong (2019) 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 $L_h(l)$,

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

and

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

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

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

and

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

yielding

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

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 $L_h(l_1) \neq L_h(l_2)$,

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

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* $0 \le s \le h-1$,

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

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

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

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

and

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

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, Ing and Tong (2019). 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-called "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 nonnested misspecified models with the same MI may have different VIs even if they share the same number of parameters; see Section S5 of Hsu, Ing and Tong (2019). 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, Ing and Tong (2019).

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

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

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

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 $\hat{\mathbf{l}}_H = (\hat{l}_1, \ldots, \hat{l}_H)^\top$ 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, Ing and Tong (2019).

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 $\mathrm{E}(\boldsymbol{\delta}_t(p)\boldsymbol{\delta}_t^\top(p)) = \Sigma_p$, 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 $|z| < 1$,

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

and

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

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

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

with $J_v^{(l)}, 0 \leq v \leq p$, being given finite sets of nonnegative 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 $v_{t+2} = \epsilon_{t+2} + a\epsilon_{t+1} + \delta_{t+1}^{(1)}$. Since the model involves the infinite past $s_t^{(1)}, s_{t-1}^{(1)}, \ldots,$ any candidate model containing only a finite number of the lagged variables of $s_t^{(1)}$ is misspecified.

We aim at finding a data-driven method to choose among the candidate models such that (3.6) is satisfied. Let $\varepsilon_{t,h}^{(l)}$, $\hat{y}_{n+h}(l)$, $\mathrm{MI}_h(l)$, $L_h(l)$, 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 $\psi_j^{(v)}, \theta_j$ and $c_j^{(v)}$.

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

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

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

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

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

$$|\theta_j| \leq c_1 (j+1)^{-s} \quad \textit{and} \quad |\psi_j^{(v)}| + |c_j^{(v)}| \leq c_1 (j+1)^{-s}, \qquad 1 \leq v \leq p. \tag{3.29}$$

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

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

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

REMARK 4. Assumption (3.29) allows the component of $\mathbf{x}_t^{(l)}$ to not only be a short-memory ARMA process, but also belong to some important classes of long-memory processes, for example, 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, Ing and Tong (2019). 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. *Consistency 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 nonrandom 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

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

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 $\hat{J}_0 = \varnothing$. 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

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

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

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

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) $N_h \neq \varnothing$ 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, Ritov and Tsybakov (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:

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

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

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

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

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

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, Ing and Tong (2019) 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,

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

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, that is, 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

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

respectively, where

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

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

$$\begin{aligned}
M_{A,h} &= \Big\{l : 1 \le l \le K, \mathrm{MI}_{h,l}\big(N_h^{(l)}\big) = \min_{1\le j\le K} \mathrm{MI}_{h,j}\big(N_h^{(j)}\big)\Big\},\\
M_{B,h} &= \Big\{l : L_{h,l}\big(N_h^{(l)}\big) = \min_{j\in M_{A,h}} L_{h,j}\big(N_h^{(j)}\big)\Big\}
\end{aligned}$$

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

with

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

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

$$\hat{\varepsilon}_{t,h}^{(l)}(J) = y_{t+h} - \hat{\boldsymbol{\beta}}_{h,l}^\top(J)\mathbf{x}_t^{(l)}(J),$$

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

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),

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

and

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

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

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

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

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

Then

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

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, Ing and Tong (2019).

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 multistep 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., nonlikelihood-based) approaches are available and beginning to attract attention; see, for example, Davies (2008), Xia and Tong (2011) and others. Therefore, it remains a future challenge to develop a model selection criterion via a nonlikelihood-based approach.

APPENDIX: 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 $-\infty < t < \infty$,

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

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 $\mathcal{G}_t = \sigma(x_j, j \le t)$. Then

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

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

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

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

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

where $\mathcal{G}_t' = \sigma(x_j, w_j, j \le t)$, 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 improved the presentation of this paper.

SUPPLEMENTARY MATERIAL

Supplement to "On model selection from a finite family of possibly misspecified time series models" (DOI: 10.1214/18-AOS1706SUPP; .pdf). 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. Automat. Control* **19** 716–723. MR0423716

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. MR3357870

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BOZDOGAN, H. (2000). Akaike's information criterion and recent developments in information complexity. *J. Math. Psych.* **44** 62–91. MR1770002

BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer, New York. MR1919620

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. MR2850211

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. MR2530322

FINDLEY, D. F. (1991). Counterexamples to parsimony and BIC. *Ann. Inst. Statist. Math.* **43** 505–514. MR1143638

FINDLEY, D. F. and WEI, C. Z. (1993). Moment bounds for deriving time series CLTs and model selection procedures. *Statist. Sinica* **3** 453–480. MR1243396

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. MR1945962

GREENAWAY-MCGREVY, R. (2013). Multistep prediction of panel vector autoregressive processes. *Econometric Theory* **29** 699–734. MR3092461

GREENAWAY-MCGREVY, R. (2015). Evaluating panel data forecasts under independent realization. *J. Multivariate Anal.* **136** 108–125. MR3321483

HSU, H.-L., ING, C.-K. and TONG, H. (2019). Supplement to "On model selection from a finite family of possibly misspecified time series models." DOI:10.1214/18-AOS1706SUPP.

ING, C.-K. (2003). Multistep prediction in autoregressive processes. *Econometric Theory* **19** 254–279. MR1966030

ING, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Ann. Statist.* **35** 1238–1277. MR2341705

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. MR2895106

ING, C.-K. and WEI, C.-Z. (2003). On same-realization prediction in an infinite-order autoregressive process. *J. Multivariate Anal.* **85** 130–155. MR1978181

ING, C.-K. and WEI, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *Ann. Statist.* **33** 2423–2474. MR2211091

INOUE, A. and KILIAN, L. (2006). On the selection of forecasting models. *J. Econometrics* **130** 273–306. MR2211796

KONISHI, S. and KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83** 875–890. MR1440051

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. MR0902239

LIU, W. and YANG, Y. (2011). Parametric or nonparametric? A parametricness index for model selection. *Ann. Statist.* **39** 2074–2102. MR2893862

LV, J. and LIU, J. S. (2014). Model selection principles in misspecified models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 141–167. MR3153937

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. MR0740928

RAO, C. R. and WU, Y. H. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76** 369–374. MR1016028

SCHORFHEIDE, F. (2005). VAR forecasting under misspecification. *J. Econometrics* **128** 99–136. MR2022928

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHAO, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. MR1466682

SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126. MR0403130

SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164. MR0557560

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54. MR0614940

SIN, C.-Y. and WHITE, H. (1996). Information criteria for selecting possibly misspecified parametric models. *J. Econometrics* **71** 207–225. MR1381082

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. MR0443204

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

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN ERVEN, T., GRÜNWALD, P. 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. MR2925369

WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **104** 1512–1524. MR2750576

WEI, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20** 1–42. MR1150333

WU, W.-B. and WU, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.* **10** 352–379. MR3466186

XIA, Y. and TONG, H. (2011). Feature matching (with discussion). *Statist. Sci.* **26** 21–46.

YANG, Y. (2007). Prediction/estimation with simple linear models: Is it really that simple? *Econometric Theory* **23** 1–36. MR2338950

ZHANG, Y. and YANG, Y. (2015). Cross-validation for selecting a model selection procedure. *J. Econometrics* **187** 95–112. MR3347297

H.-L. HSU
INSTITUTE OF STATISTICS
NATIONAL UNIVERSITY OF KAOHSIUNG
700 KAOHSIUNG ROAD
KAOHSIUNG 811
TAIWAN
E-MAIL: hsuhl@nuk.edu.tw

C.-K. ING
INSTITUTE OF STATISTICS
NATIONAL TSING HUA UNIVERSITY
101, SECTION 2, KUANG FU ROAD
HSINCHU 30013
TAIWAN
E-MAIL: cking@stat.nthu.edu.tw

H. TONG
UNIVERSITY OF ELECTRONIC SCIENCE & TECHNOLOGY
4, SECTION 2, NORTH JIANSHE ROAD
CHENGDU, SICHUAN 610054
CHINA
AND
DEPARTMENT OF STATISTICS
LONDON SCHOOL OF ECONOMICS
HOUGHTON STREET
LONDON WC2A 2AE
UNITED KINGDOM
E-MAIL: howell.tong@gmail.com