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

Hsiang-Ling Hsu

National University of Kaohsiung, Taiwan

E-mail: hsuhl@nuk.edu.tw

Ching-Kang Ing

Academia Sinica and National Taiwan University, Taiwan

E-mail: cking@stat.sinica.edu.tw

Howell Tong

University of Electronic Science & Technology, China and London School of Economics, United Kingdom

E-mail: howell.tong@gmail.com

Summary. Consider finite parametric time series models. In many practical situations, a really fundamental problem is that of selecting a model from a *finite and fixed* 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. Instead of addressing the above problem, model selection problems to-date tend to be classified into two categories according to whether the true DGP is included among the candidate models. The first category assumes that the true DGP is among the candidates, and the objective of model selection is simply to choose this DGP. The second category assumes that the true DGP is not among the candidates. In this case, one primary objective is to choose the model that has the best predictive capability in some asymptotic sense by allowing (at least in theory if not in practice) the number of candidates to increase indefinitely with the sample size. However, most existing model selection criteria can only perform well in at most one category, and hence when the underlying category is unknown, the choice of selection criteria becomes a serious point of contention. In this article, we propose a misspecification-resistant information criterion (MRIC) to address this difficult problem, by working within the fixed-dimensional framework that requires that the number of candidate finite-parametric models is fixed irrespective of the sample size. We prove the asymptotic efficiency of MRIC whether the true DGP is among the candidates or not. We also illustrate MRIC's finite-sample performance through simulated and real data, including some high-dimensional cases.

Keywords: AIC, BIC, High-dimensional models, Misspecification-resistant information criterion, Model selection, Multi-step Prediction error, Orthogonal greedy algorithm

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, 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 errors vanishes ultimately, Shibata (1981) and Li (1987) showed that AIC and Mallows' C_p possess asymptotic efficiency (AE), 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 AE 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 AE 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 recently 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 AE 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).

Although these recent efforts to resolve the controversy between AIC and BIC are novel, they mainly contribute to the increasing-dimensional (ID) framework, which allows the number of candidate predictor variables to grow to infinity with the sample size. The ID framework, however, may not be applicable to situations where collecting an increasing number of predictor variables is expensive, technically infeasible, or unnecessary according to domain knowledge (constraints). To cite a trivial example of the last, Kepler's third law asserts that the ratio of the square of the revolutionary period to the cube of the orbital axis is the same for all the planets of the solar system. Therefore, if we wish to establish a statistical model for a planet's period of revolution around the sun, predictor variables other than its orbital axis appear to be unessential, even when more data become available.

It can be argued that the really fundamental question is the following. 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. 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 the sample size, thus filling a serious lacuna in the vast literature on model selection.

It is already well known that when category I holds, BIC is consistent under both ID and FD frameworks; see Shao (1997) and Ing (2007). On the other hand, when category II holds instead of category I, AIC is AE under the ID framework but fails to carry over to the FD one, as illustrated in Section 5. (Note that the definitions of AE in the FD and the ID frameworks are slightly different but similar in spirit; see Section 3.) Sin and White (1996) and Inoue and Kilian (2006) have shown that a BIC-type criterion has the so-call 'strong parsimony property' under the FD framework in the sense that it will asymptotically choose the most parsimonious model among those candidates having the smallest population MSPE; see Sections 2 and 3 for further discussion. However, misspecified models can behave quite differently from correctly specified ones. For example, as shown by Findley (1991), when two misspecified models have the same population MSPE, the one with fewer parameters does not necessarily lead to a smaller finite-sample MSPE, which is the sum of the population MSPE and a term accounting for the estimation error. Moreover, it is shown in Section 5 that when two competing non-nested models are misspecified and share the same population MSPE and the same number of parameters, both BIC and AIC tend to randomly choose between the two alternatives instead of selecting the one having the smaller finite-sample MSPE. As a result, AE is also not achievable by the BIC-type criteria under Category II and within the FD framework. Indeed, there are already several criteria 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 AE under the FD framework; see also Section 5.

In this article, we propose a misspecification-resistant information criterion (MRIC). Specifically, we prove that MRIC, within the FD framework, possesses AE 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 AE 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 AE on its own without the help of additional/auxiliary criteria. Third, by incorporating some screening methods, MRIC also performs

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 (AE).	Case III: No info. on whether the true model is included. Goal: Consistency when the true model is included + AE 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: AE	Case III: Consistency + AE
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

satisfactorily in high-dimensional models; see Sections 5 and 6.

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, Table 2 for the FD framework and both tables focus on the case of $h = 1$, in which AE is equivalent to selection consistency under category I; see Section 3 for further details.

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 estimator of the expression obtained in Section 2.1, we propose our MRIC and prove its AE under the FD framework in Section 3.1. Applications of MRIC to misspecified ARX models are given in Section 3.2. In Section 4.1, the results in Sections 2 and 3 are extended to nonlinear models. Specifically, our nonlinear extension of MRIC involves an additional term accounting for the joint effect of nonlinearity and model misspecification. In Section 4.2, we furnish the technical conditions used in Section 4.1 and compare them with the conditions presented in Section 2.2. In Section 5, we conduct a careful comparison of the finite sample performance of MRIC, AIC, BIC, GAIC, GBIC and GBIC$_p$ through simulated data generated from some linear, nonlinear and high-dimensional models. In Section 6, the advantage of MRIC is demonstrated using two real datasets. We conclude in Section 7. All proofs are relegated to Appendices A to E.

2. Mean squared prediction error

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,

$$y_{t+h} = \alpha_h + \boldsymbol{\beta}_h^\top \mathbf{x}_t + \varepsilon_{t,h}, \tag{1}$$

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 when $\varepsilon_{t,1} = \epsilon_t$ is a white noise process, $\mathbf{x}_t = (y_t, \ldots, y_{t+1-m})^\top$, and $1 - \beta_{1,1}z - \cdots - \beta_{m,1}z^m \neq 0$ for all $|z| \leq 1$, model (1), with $h = 1$, is the (linear) autoregressive (AR) model of order m satisfying $\mathrm{E}(\mathbf{x}_k \varepsilon_{t,1}) = \mathbf{0}$ for all $k \leq t$. However, since 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 $k \neq t$, model (1) actually represents much more general situations, including, for example, multistep prediction in (possibly) misspecified AR or autoregressive exogenous (ARX) models. Having observed $y_1, \ldots, y_n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we may replace y_t by $y_t - \bar{y}$ and $\mathbf{x}_t$ by $\mathbf{x}_t - \bar{\mathbf{x}}$, where $\bar{\mathbf{x}} = n^{-1}\sum_{t=1}^n \mathbf{x}_t$ and $\bar{y} = n^{-1}\sum_{t=1}^n y_t$, and assume, without loss of generality, that $\mathrm{E}(y_t) = 0$ and $\mathrm{E}(\mathbf{x}_t) = \mathbf{0}$. Hence (1) becomes

$$y_{t+h} = \boldsymbol{\beta}_h^\top \mathbf{x}_t + \varepsilon_{t,h}. \tag{2}$$

Using the least squares estimator (LSE),

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

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 $\widehat{\mathbf{R}}_k = k^{-1}\sum_{t=1}^k \mathbf{x}_t \mathbf{x}_t^\top$ for positive integer k.

In the next theorem, we shall provide an asymptotic expression for the finite-sample mean squared prediction error (MSPE) of $\hat{y}_{n+h}$, namely $\mathrm{E}(y_{n+h} - \hat{y}_{n+h})^2$. (For the sake of simplicity, we will refer to finite-sample MSPE as 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; see Section 3 for further details. The proof of Theorem 1 is given in Appendix A.

THEOREM 1. *Assume* (2) *and conditions* (C1)–(C6) *in Section 2.2. Then, for any* $h \geq 1$,

$$\mathrm{E}\left(y_{n+h} - \hat{y}_{n+h}\right)^2 = \mathrm{E}\left(\varepsilon_{n,h}^2\right) + n^{-1}(L_h + o(1)), \tag{3}$$

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 (3), referred to as the population MSPE, can be viewed as a measure of the goodness fit of model (2), whereas the second term on the right-hand side of (3) is related to the estimation error of $\hat{\boldsymbol{\beta}}_n(h)$. To appreciate the novelty of Theorem 1, assume that y_t is a stationary AR(m) model,

$$y_{t+1} = \sum_{i=1}^{m} a_i y_{t+1-i} + \epsilon_{t+1}, \tag{4}$$

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 (4), a correctly specified model for the h-step, $h \geq 1$, prediction is given by

$$y_{t+h} = \boldsymbol{\beta}_h^\top \mathbf{x}_t + \varepsilon_{t,h}, \tag{5}$$

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

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

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 1 and some algebraic manipulations,

$$\lim_{n\to\infty} n\{\mathrm{E}\left(y_{n+h} - \hat{y}_{n+h}\right)^2 - \mathrm{E}\left(\varepsilon_{n,h}^2\right)\} = L_h = \mathrm{tr}\left(\mathbf{R}^{-1}\mathrm{cov}\left(\sum_{j=0}^{h-1} b_j \mathbf{x}_{1+j}\right)\right)\sigma^2, \tag{6}$$

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,

$$\begin{aligned} & n\{\mathrm{E}\left(y_{n+h} - \hat{y}_{n+h}\right)^2 - \mathrm{E}(\varepsilon_{n,h}^2)\} \\ & = -2\mathrm{E}\{\varepsilon_{n,h}\mathbf{x}_n^\top \widehat{\mathbf{R}}_N^{-1} \sum_{t=1}^{N} \mathbf{x}_t \varepsilon_{t,h}\} + \mathrm{E}(\mathbf{x}_n^\top \widehat{\mathbf{R}}_N^{-1} N^{-1/2} \sum_{t=1}^{N} \mathbf{x}_t \varepsilon_{t,h})^2, \end{aligned} \tag{7}$$

may have a nonnegligible "cross-product" term, $-2\mathrm{E}\{\varepsilon_{n,h}\mathbf{x}_n^\top \widehat{\mathbf{R}}_N^{-1} \sum_{t=1}^{N} \mathbf{x}_t \varepsilon_{t,h}\}$, 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 (6) is solely attributed to the second term on the right-hand side of (7). 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 cancellation between some components in the first and the second terms on the right-hand side of (7); see Appendix A for details.

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

$$\mathrm{E}\left|(n_2 - n_1 + 1)^{-1/2} \sum_{t=n_1}^{n_2} x_{t,i}x_{t,j} - \mathrm{E}\left(x_{t,i}x_{t,j}\right)\right|^{q_1} \le C_1. \tag{8}$$

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

$$\mathrm{E}\left(x_{1,i}x_{n,j}\varepsilon_{1,h}\varepsilon_{n,h}\right) = o(n^{-1}). \tag{9}$$

(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 $1 \le n_1 < n_2 \le n - h$,

$$\mathrm{E}\left\|(n_2 - n_1 + 1)^{-1/2} \sum_{t=n_1}^{n_2} \mathbf{x}_t\varepsilon_{t,h}\right\|^5 < C_2. \tag{10}$$

(C5) For any $q > 0$,

$$\mathrm{E}\|\widehat{\mathbf{R}}_n^{-1}\|^q = O(1), \tag{11}$$

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

$$\sup_{-\infty<t<\infty} \mathrm{E}\left\|\mathrm{E}\left(\mathbf{x}_t\mathbf{x}_t^{\top}\middle|\mathcal{F}_{t-k}\right) - \mathbf{R}\right\|^3 = o(1), \tag{12}$$

$$\sup_{-\infty<t<\infty} \mathrm{E}\left\|\mathrm{E}\left(\mathbf{x}_t\varepsilon_{t,h}|\mathcal{F}_{t-k}\right)\right\|^3 = o(1), \tag{13}$$

as $k \to \infty$.

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

$$x_{t,i} = \sum_{s=0}^{\infty} \mathbf{a}_{s,i}^{\top}\,\boldsymbol{\epsilon}_{t-s}, \tag{14}$$

and

$$\varepsilon_{t,h} = \sum_{s=0}^{\infty} \mathbf{b}_s^{\top}\,\boldsymbol{\epsilon}_{t+h-s}, \tag{15}$$

where $\boldsymbol{\epsilon}_t = (\epsilon_t, \epsilon_t^{(1)}, \dots, \epsilon_t^{(m)})^\top$ 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, (8) and (10) hold true, provided

$$\sum_{k=-\infty}^{\infty} (\gamma_1^2(k) + \cdots + \gamma_m^2(k)) < \infty, \ \sum_{k=-\infty}^{\infty} \gamma^2(h,k) < \infty, \tag{16}$$

$$\mathrm{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\top | \mathcal{G}_{t-1}) = \boldsymbol{\Sigma} \text{ and } \sup_{-\infty<t<\infty} \mathrm{E}(\|\boldsymbol{\epsilon}_t\|^{q^*} | \mathcal{G}_{t-1}) < C_{q^*} \text{ with probability 1,} \tag{17}$$

where $\boldsymbol{\Sigma}$ 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 (14)–(17) lead to (10) and (8), 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 (14)–(16) are fulfilled by not only short-memory autoregressive moving average (ARMA) processes but also some long-memory processes; see Section 3.2 for more details. While it is possible to justify (8) and (10) under more general conditions, such an investigation is beyond the scope of the present article.

Condition (C2) leads to an unexpected cancellation associated with the right-hand side of (7) 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. This latter requirement appears to be mild for most stationary time series; see also Section 3.2 for further discussion.

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. This kind of assumption is also quite natural for most stationary time series. Conditions (C5) and (C6) are used to show that the first and second terms on the right-hand side of (7) are asymptotically equivalent to

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

respectively, which facilitate mathematical analysis. According to Theorem 2.1 of Chan and Ing (2011), (11) 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 $\|\mathbf{v}\| = 1$,

$$P(s_1 < \mathbf{v}^\top\mathbf{x}_t \le s_2 | \mathcal{F}_{t-D}) \le M\,(s_2 - s_1)^\alpha \text{ almost surely.} \tag{18}$$

Note that (18) is flexible enough to allow for a variety of time series applications; see Section 3 of Chan and Ing (2011) and Section 3.2 below. Moreover, in the special case of (4), (18) can be superseded by a simpler condition,

$$P(s_1 < \epsilon_t \le s_2) \le M\,(s_2 - s_1)^\alpha; \tag{19}$$

see Ing and Wei (2003). 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

3.1. Asymptotic efficiency of MRIC

Being the population MSPE of model (2), the first term on the right-hand side of (3) 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 (3) is refereed to as *variability index* (VI) because it is contributed by the sampling variability of $\hat{y}_{n+h} = \hat{\boldsymbol{\beta}}_n^\top(h)\mathbf{x}_n$; see the proof of Theorem 1. As revealed by (3), selecting the model with the smallest MSPE amounts to selecting the model with the smallest VI among those with the smallest MI when n is large enough.

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

$$y_{n+h} = \boldsymbol{\beta}_{h,l}^\top \mathbf{x}_n(J_l) + \varepsilon_{n,h}^{(l)}, l = 1, \dots, K, \tag{20}$$

where J_l is a set of positive integers, $\mathbf{x}_t(J_l) = (x_{t,j}, j \in J_l)$, $\boldsymbol{\beta}_{h,l}^\top \mathbf{x}_t(J_l)$ is the best linear predictor of y_{t+h} based on $\mathbf{x}_t(J_l)$, and $\varepsilon_{t,h}^{(l)} = y_{t+h} - \boldsymbol{\beta}_{h,l}^\top \mathbf{x}_t(J_l)$. In the following, we shall call J_l a 'model' whenever confusion is unlikely to occur. Let

$$\hat{y}_{n+h}(l) = \hat{\boldsymbol{\beta}}_{n,l}^\top(h)\mathbf{x}_n(J_l) \tag{21}$$

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

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

Define for each model l

$$\mathrm{MI}_h(l) = \mathrm{E}(\varepsilon_{1,h}^{(l)})^2 \text{ and } L_h(l) = \lim_{n\to\infty} n \left\{ \mathrm{E}\left(y_{n+h} - \hat{y}_{n+h}(l)\right)^2 - \mathrm{E}(\varepsilon_{1,h}^{(l)})^2 \right\},$$

which are the MI and the VI respectively for model J_l. As mentioned, our goal is to find a model $\widehat{J} \in \{J_1, \dots, J_K\}$ in a data-driven fashion such that

$$\lim_{n\to\infty} \mathrm{P}\left(\widehat{J} \in M_2\right) = 1, \tag{22}$$

where

$$M_2 = \left\{ J_k : J_k \in M_1, L_h(k) = \min_{J_l \in M_1} L_h(l) \right\},$$

with

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

A model selection criterion is said to be asymptotically efficient if (22) is fulfilled. In Section 5, we provide several interesting examples showing that to achieve (22), 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. Our examples also reveal that the best predictive model may vary with the prediction lead time h, raising another subtle issue.

Inspired by (3), our strategy to achieve (22) is to first construct the method of moments estimators of $\mathrm{MI}_h(l)$ and $L_h(l)$,

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

and

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

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

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

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

$$\mathrm{MRIC}_h(l) = \hat{\sigma}_h^2(l) + \frac{C_n}{n}\widehat{L}_h(l), \tag{23}$$

with

$$\frac{C_n}{n^{1/2}} \to \infty, \tag{24}$$

and

$$\frac{C_n}{n} \to 0. \tag{25}$$

Finally, we choose model $J_{\hat{l}_h}$, in which $\hat{l}_h = \arg\min_{1\le l\le K}\mathrm{MRIC}_h(l)$. The major difference between $\mathrm{MRIC}_h(l)$ and the natural estimator $\hat{\sigma}_h^2(l) + n^{-1}\widehat{L}_h(l)$ of $\mathrm{E}\left(y_{n+h} - \hat{y}_{n+h}(l)\right)^2$ (cf.(3)) is that the former contains an additional penalty factor C_n. 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 (C1)–(C6) and two additional assumptions, (31) and (32) (see below), we have

$$\hat{\sigma}_h^2(l) = \mathrm{MI}_h(l) + O_p(n^{-1/2}), \tag{26}$$

and

$$\widehat{L}_h(l) = L_h(l) + o_p(1), \tag{27}$$

yielding

$$\mathrm{MRIC}_h(l) = \mathrm{MI}_h(l) + O_p(n^{-1/2}) + \frac{C_n}{n}L_h(l) + o_p\left(\frac{C_n}{n}\right). \tag{28}$$

In view of (25), property (28) immediately implies

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

Moreover, it follows from (28) and (24) that for $J_{l_1}, J_{l_2} \in M_1$ with $L_h(l_1) \neq L_h(l_2)$,

$$\lim_{n\to\infty} P(\mathrm{sign}(\mathrm{MRIC}_h(l_1) - \mathrm{MRIC}_h(l_2)) = \mathrm{sign}(L_h(l_1) - L_h(l_2))) = 1, \tag{29}$$

and hence

$$\lim_{n\to\infty} P(J_{\hat{l}_h} \in M_2) = 1. \tag{30}$$

The above discussion is summarized in the next theorem the proof is sketched in Appendix B.

THEOREM 2. *Assume that* (C1)–(C6) *hold for each candidate model* $J_l, l = 1, \dots, K$. *Suppose*

$$n^{-1}\sum_{t=1}^{n}(\varepsilon_{t,h}^{(l)})^2 = \mathrm{E}(\varepsilon_{1,h}^{(l)})^2 + O_p(n^{-1/2}), \tag{31}$$

and for each $0 \le s \le h-1$,

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t(J_l)\mathbf{x}_{t+s}^{\top}(J_l)\varepsilon_{t,h}^{(l)}\varepsilon_{t+s,h}^{(l)} = \mathbf{C}_{h,s}(l) + o_p(1), \tag{32}$$

where $\mathbf{C}_{h,s}(l) = \mathrm{E}(\mathbf{x}_1(J_l)\mathbf{x}_{1+s}^{\top}(J_l)\varepsilon_{1,h}^{(l)}\varepsilon_{1+s,h}^{(l)})$. *Then,* (26) *and* (27) *hold. As a result,* (30) *follows.*

Remark 1. Assumptions (31) and (32) seem moderate when $\mathbf{x}_t(J_l)$ and $\varepsilon_{t,h}^{(l)}$ are linear process obeying (14) and (15), respectively; see Section 3.2 and Appendix C for a further discussion.

Remark 2. The restriction on C_n given in (24) can be dropped if M_1 only contains one element, and weakened to

$$C_n \to \infty \tag{33}$$

if the elements in M_1 are nested. To see this, assume $J_{l_1}, J_{l_2} \in M_1$ with $J_{l_1} \subset J_{l_2}$ and $L_h(l_1) \neq L_h(l_2)$. 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 (33) yield (29), and hence the desired conclusion.

It would be of interest to further explore the relationship between AE defined above and selection consistency. Model J_l is said to be true if

$$\mathrm{E}(\varepsilon_{t,h}^{(l)}|\mathcal{F}_t) = 0 \text{ a.s.} \tag{34}$$

Let $\tilde{M}_1$ denote the set of true models. Also define $\tilde{M}_2^* = \{J_k : J_k \in \tilde{M}_1, \sharp(J_k) = \min_{J_l\in\tilde{M}_1}\sharp(J_l)\}$, provided $\tilde{M}_1 \neq \emptyset$ (i.e., Category I holds), and $\tilde{M}_2 = \{J_k : J_k \in M_1, \sharp(J_k) = \min_{J_l\in M_1}\sharp(J_l)\}$. A model selection criterion is said to possess the 'strong parsimony property' if it can choose a model falling in $\tilde{M}_2$ with probability tending to 1 as $n \to \infty$. When $\tilde{M}_1 \neq \emptyset$, we say that a model selection criterion is consistent if it can choose a model belonging to $\tilde{M}_2^*$ with probability tending to 1 as $n \to \infty$. Since $\tilde{M}_1 = M_1$ in the case of $\tilde{M}_1 \neq \emptyset$, we have $\tilde{M}_2^* = \tilde{M}_2$, yielding that strong parsimony property is equivalent to consistency when at least one true model is contained among the candidate models.

When $\tilde{M}_1 = \emptyset$, some examples in Section 5 show that $\tilde{M}_2 \neq M_2$, and hence strong parsimony property and AE have fundamentally different goals to pursue. Is strong parsimony property/consistency equivalent to AE when $\tilde{M}_1 \neq \emptyset$? By (3) and (34), it is not difficult to prove that for $h = 1$ and $\tilde{M}_1 \neq \emptyset$,

$$\tilde{M}_2^* = M_2, \tag{35}$$

provided there is a positive constant $G_0(1)$ such that for all $J_l \in M_1 = \tilde{M}_1$,

$$\mathrm{E}((\varepsilon_{1,1}^{(l)})^2 | \mathcal{F}_1) = G_0(1) \text{ a.s.} \tag{36}$$

Therefore, pursuing AE in the case of $h = 1$ is identical to pursuing consistency/strong parsimony property in situations where $\tilde{M}_1 \neq \emptyset$ and (36) holds true. This fact and Theorem 2 also lead to the last row of Table 2. However, for $h > 1$ and $\tilde{M}_1 \neq \emptyset$, even assuming an extension of (36), namely

$$\mathrm{E}(\varepsilon_{1,h}^{(l)} \varepsilon_{1+s,h}^{(l)} | \mathcal{F}_{1+s}) = G_s(h) \text{ a.s.},$$

where $G_s(h), 0 \leq s \leq h-1$, are some constants with $G_0(h) > 0$, (35) does not necessarily hold, unless the true model is an AR model; see Ing (2003).

We close this section with a brief comparison of the AEs in the ID and the FD frameworks when category II holds true. To simplify the discussion, we shall focus on the case of $h = 1$ and nested candidates. For the ID case, assume there are K_n candidates models $J_1 \subset J_2 \cdots \subset J_{K_n}$, with $K_n \to \infty$ and $K_n = o(n)$, such that $\mathrm{E}(\varepsilon_{1,1}^{(K_n)})^2 \to Q_0(1) = \mathrm{E}(y_2 - \mathrm{E}(y_2|\mathcal{F}_1))^2 > 0$, noting that under category II, $\mathrm{E}(\varepsilon_{1,1}^{(l)})^2 \neq Q_0(1)$ for all $1 \leq l \leq K_n$. A model selection criterion is said to be AE in the ID framework if it can choose a model $J_{\hat{k}}$, with $\hat{k} \in \{1, \ldots, K_n\}$, such that $\mathrm{E}(y_{n+1} - \hat{y}_{n+1}(\hat{k}))^2 - Q_0(1)$ is asymptotically equivalent to $\min_{1 \leq l \leq K_n} \{\mathrm{E}(y_{n+1} - \hat{y}_{n+1}(l))^2 - Q_0(1)\}$. Since

$$\mathrm{E}(y_{n+1} - \hat{y}_{n+1}(l))^2 - Q_0(1) = \{\mathrm{E}(\varepsilon_{1,1}^{(l)})^2 - Q_0(1)\} + \{\mathrm{E}(y_{n+1} - \hat{y}_{n+1}(l))^2 - \mathrm{E}(\varepsilon_{1,1}^{(l)})^2\}, \tag{37}$$

a potential candidate must satisfy $B^2(l) = \mathrm{E}(\varepsilon_{1,1}^{(l)})^2 - Q_0(1) \to 0$ and strike a good balance between $B^2(l)$ and the second term on the right-hand side of (37), which is closely related the VI of model J_l. However, in the FD framework where $K_n = K$ is a fixed integer, $B^2(l) > 0, 1 \leq l \leq K$, always dominates the other term, and hence striking a balance between these two terms is no longer the key point of contention. That is why our definition of AE for the FD framework is different from the one for the ID framework, although both are from the MSPE point of view.

3.2. Applications to ARX models

First let B denote the back shift operator such that $By_t = y_{t-1}$. In this section, we assume that data are generated by the following ARX model,

$$\boldsymbol{\phi}(B) y_{t+1} = \sum_{v=1}^{p} \sum_{j=0}^{r_v} \eta_j^{(v)} s_{t-j}^{(v)} + \epsilon_{t+1}, \tag{38}$$

where 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 $\sum_{j=0}^{\infty} |\phi_j| < \infty$, $\eta_j^{(v)}$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 and for any $|z| < 1$,

$$\boldsymbol{\phi}^{-1}(z) = \boldsymbol{\theta}(z) = \sum_{j=0}^{\infty} \theta_j z^j, \text{ with } \sum_{j=0}^{\infty} \theta_j^2 < \infty, \tag{39}$$

$$\boldsymbol{\theta}(e^{-i\lambda}) = \sum_{j=0}^{\infty} \theta_j e^{-ij\lambda} \neq 0, -\pi \leq \lambda \leq \pi, \tag{40}$$

and

$$\sum_{j=0}^{\infty} (c_j^{(v)})^2 < \infty, \text{ with } c_j^{(v)} = \sum_{k=0}^{j} \psi_k^{(v)} \theta_{j-k}, 1 \leq v \leq p. \tag{41}$$

These specifications yield that the spectral density, $f_y(\lambda)$, of y_t obeys

$$f_y(\lambda) = \frac{1}{2\pi}\Big\{ \Big(\boldsymbol{\eta}_1(e^{-i\lambda})\boldsymbol{C}_1(e^{-i\lambda}), \ldots, \boldsymbol{\eta}_p(e^{-i\lambda})\boldsymbol{C}_p(e^{-i\lambda})\Big) \Sigma_p \\ \Big(\boldsymbol{\eta}_1(e^{i\lambda})\boldsymbol{C}_1(e^{i\lambda}), \ldots, \boldsymbol{\eta}_p(e^{i\lambda})\boldsymbol{C}_p(e^{i\lambda})\Big)^\top + |\boldsymbol{\theta}(e^{-i\lambda})|^2\sigma^2\Big\} > 0, -\pi \leq \lambda \leq \pi,$$

where $\boldsymbol{\eta}_v(z) = \sum_{j=1}^{r_v} \eta_j^{(v)} z^{j-1}$ and $\boldsymbol{C}_v(z) = \sum_{j=0}^{\infty} c_j^{(v)} z^j$, with $1 \leq v \leq p$.

Having observed $y_1, \ldots, y_n$ and $s_1^{(v)}, \ldots, s_n^{(v)}, 1 \leq v \leq p^* \leq p$, we are interested in predicting $y_{n+h}, h \geq 1$, using one of the candidate models $J_1, \ldots, J_K$, where for $1 \leq l \leq K$, $J_l = J_0^{(l)} \times J_1^{(l)} \times \cdots \times J_{p^*}^{(l)}$, with $J_v^{(l)}, 0 \leq v \leq p^*$, being a given finite set of non-negative integers. Note that the regressor corresponding to model J_l at time t is

$$\mathbf{x}_t(J_l) = (y_{t-j}, j \in J_0^{(l)}, s_{t-j}^{(v)}, j \in J_v^{(l)}, 1 \leq v \leq p^*)^\top, \tag{42}$$

and all candidate models are subject to misspecification, in view of (38) and (42). Without loss of generality, we shall assume that $\{y_{1-j}, j \in J_0^{(l)}\}$ are observed in order to make $\mathbf{x}_t(J_l)$ available for all $1 \leq t \leq n$. We aim at finding a data-driven method to choose among $J_1, \ldots, J_K$ such that (22) is satisfied. With (42), let $\hat{y}_{n+h}(l)$, $\varepsilon_{t,h}^{(l)}$, $\mathrm{MI}_h(l), L_h(l), M_1$ and M_2 be defined as in Section 3.1. The next theorem shows that MRIC, introduced in (23)–(25), 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. *Assume that* (38)–(41) *hold. Suppose that the fourth moments of* $\{\mathbf{v}_t\}$ *are independent of* t,

$$\sup_{-\infty<t<\infty} \mathrm{E}\|\mathbf{v}_t\|^\theta < \infty, \textit{ for some } \theta > 10, \tag{43}$$

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

$$\sup_{\|\mathbf{a}\|=1} \mathrm{P}\left(u < \mathbf{a}^\top \mathbf{v}_t \leq w\right) \leq K_1(w-u)^\nu. \tag{44}$$

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

$$|\theta_j| \leq c_1(j+1)^{-s} \textit{ and } |\psi_j^{(v)}| + |c_j^{(v)}| \leq c_1(j+1)^{-s}, 1 \leq v \leq p. \tag{45}$$

Then, (C1)–(C6) *hold for* $\mathbf{x}_t = \mathbf{x}_t(J_l)$, $\varepsilon_{t,h} = \varepsilon_{t,h}^{(l)}$, *and* $\mathcal{F}_t = \sigma(\mathbf{v}_t, \mathbf{v}_{t-1}, \ldots)$. *Moreover,* (31) *and* (32) *follow, and hence* (30) *holds true.*

Remark 3. Assumption (45) allows the component of $\mathbf{x}_t(J_l)$ 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 $0 < d < 1/4$. As is clear from the proof of Theorem 3 given in Appendix C, (45) is crucial for verifying (C.2) and (32) and can hardly be weakened.

Remark 4. Assumption (44) is used to prove (18), which in turn leads to (C.5) according to Chan and Ing (2011). For more details, see Lemma C.1 in Appendix C. Note also that (C.5) 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 Greenway-McGrevy (2013, 2015). Based on assumptions like (44), most of these papers prove (C.5) in situations where the regressor vector only contains endogenous variables. To the best of our knowledge, this is the first work that shows how to justify (C.5) when the regressor vector consists of both endogenous and exogenous variables.

4. A nonlinear extension

4.1. A nonlinear extension of MRIC and its asymptotic efficiency

In this section, we generalize the results obtained previously 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 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^\top$ 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. 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 (3). Define $D_1 g_{t,h}(\boldsymbol{\theta}) = (\partial g_{t,h}(\boldsymbol{\theta})/\partial\theta_1 \dots \partial g_{t,h}(\boldsymbol{\theta})/\partial\theta_m)^\top$ 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 4. *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 $B_\delta(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta\} \subset \Theta$, 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 4.2 hold. Then, for $h \ge 1$,*

$$\mathrm{E}(y_{n+h} - g_{n,h}(\hat{\boldsymbol{\theta}}_n))^2 = V(\boldsymbol{\theta}^*) + n^{-1}(L_h^* + o(1)), \tag{46}$$

where $L_h^* = tr((\mathbf{R}^* - \mathbf{A}^*)^{-1}\mathbf{C}_{h,0}^*) + 2\sum_{s=1}^{h-1} tr((\mathbf{R}^* - \mathbf{A}^*)^{-1}\mathbf{C}_{h,s}^*)$, 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 5. There is a striking resemblance between (46) and (3). In particular, (46) reduces to (3) when $g_{t,h}(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$. Compared to the L_h in (3), 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 4 is also of independent interest because it provides the first result revealing that the simple MSPE formula (6) obtained in correctly specified AR models carries over (after a very 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). Finally, we note that the similarities and dissimilarities between (E1)–(E7) and (C1)–(C6) will be discussed in Section 4.2.

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 $1 \leq l \leq K$, $V_l(\boldsymbol{\theta}) \equiv \mathrm{E}(y_{t+h} - g_{t,h}^{(l)}(\boldsymbol{\theta}))^2$ 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 4, the nonlinear counterparts of M_1 and M_2 are given by

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

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 (23),

$$\mathrm{MRIC}_h^*(l) = \frac{S_n^{(l)}(\hat{\boldsymbol{\theta}}_{nl})}{N} + \frac{C_n}{n}\widehat{L}_h^*(l), \tag{47}$$

where C_n satisfies (24) and (25),

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

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 establishes the desired property of MRIC_h^*,

$$\lim_{n\to\infty} P(\hat{l}_h^* \in \mathcal{D}_2) = 1. \tag{48}$$

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

THEOREM 5. *Suppose that for each $1 \leq l \leq K$, $g^{(l)}_{t,h}(\boldsymbol{\theta})$ is continuous on Θ_l and there is $\bar{\delta}_l > 0$ such that $D_1 g^{(l)}_{t,h}(\boldsymbol{\theta})$ is continuously differentiable on $B_{\bar{\delta}_l}(\boldsymbol{\theta}^*_l) \subset \Theta_l$ and each component of $D_2 g^{(l)}_{t,h}(\boldsymbol{\theta})$ is differentiable on $B_{\bar{\delta}_l}(\boldsymbol{\theta}^*_l)$. Assume also that conditions* (E1)–(E7) *in Section 4.2 hold for each candidate models. Then, for $h \geq 1$ and $1 \leq l \leq K$,*

$$\mathrm{E}(y_{n+h} - g^{(l)}_{n,h}(\hat{\boldsymbol{\theta}}_{nl}))^2 = V_l(\boldsymbol{\theta}^*_l) + n^{-1}(L^*_h(l) + o(1)). \tag{49}$$

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

$$\frac{1}{n}\sum_{t=1}^{n}(\varepsilon^{(l)}_{t,h}(\boldsymbol{\theta}^*_l))^2 = V_l(\boldsymbol{\theta}^*_l) + O_p(n^{-1/2}), \tag{50}$$

$$\frac{1}{n}\sum_{t=1}^{n} D_2 g^{(l)}_{t,h}(\boldsymbol{\theta}^*_l)\varepsilon^{(l)}_{t,h}(\boldsymbol{\theta}^*_l) = \mathbf{A}^*(l) + o_p(1), \tag{51}$$

and

$$\frac{1}{n}\sum_{t=1}^{n} D_1 g^{(l)}_{t,h}(\boldsymbol{\theta}^*_l) D_1^\top g^{(l)}_{t+s,h}(\boldsymbol{\theta}^*_l)\varepsilon^{(l)}_{t,h}(\boldsymbol{\theta}^*_l)\varepsilon^{(l)}_{t+s,h}(\boldsymbol{\theta}^*_l) = \mathbf{C}^*_{h,s}(l) + o_p(1). \tag{52}$$

Then, (48) *follows.*

Remark 6. Whereas (51) is exclusive for nonlinear regressions, (50) and (52) parallel (31) and (32) used in the linear case. When $\varepsilon^{(l)}_{t,h}(\boldsymbol{\theta}^*_l)$ and the components of $D_2 g^{(l)}_{t,h}(\boldsymbol{\theta}^*_l)$ and $D_1 g^{(l)}_{t,h}(\boldsymbol{\theta}^*_l)$ are linear processes, a discussion about how assumptions like (50)–(52) are verified has been given in Section 3 and Appendix C. 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 (48) 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 AE under the FD framework.

4.2. 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

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

for $j = 1, 2, 3$, 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 proofs of Theorems 4 and 5 to the latter conditions in the proofs of Theorems 1 and 2. (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 Appendix E. 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 4 and 5 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 converge to $V(\boldsymbol{\theta})$ uniformly over Θ as $t \to \infty$. In this case, $\mathbf{R}^*$, $\mathbf{C}_{h,s}^*$ and $\mathbf{A}^*$ become $\lim_{t\to\infty}\mathrm{E}\{D_1 g_{t,h}(\boldsymbol{\theta}^*)D_1^\top g_{t,h}(\boldsymbol{\theta}^*)\}$, $\lim_{t\to\infty}\mathrm{E}\{D_1 g_{t,h}(\boldsymbol{\theta}^*)D_1^\top g_{t+s,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t+s,h}(\boldsymbol{\theta}^*)\}$, and $\lim_{t\to\infty}\mathrm{E}\{D_2 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)\}$, 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 $\sup_{t\ge H_1}$, where H_1 is some large integer.

5. Numerical studies

In this section, the performance of MRIC is illustrated via three simulated examples. The first and second examples focus on linear and nonliner models, respectively, whereas the third one addresses high-dimensional models. 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 true DGPs.

$$y_{t+1} = \beta_1 z_t + \beta_2 w_t + \varepsilon_{t+1}, \tag{53}$$

in which $\varepsilon_t \sim \mathrm{NID}(0,1)$, $z_t = \phi z_{t-1} + \eta_t$ 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 $\eta_t \sim \mathrm{NID}(0, \sigma_\eta^2)$, $\delta_t \sim \mathrm{NID}(0, \sigma_w^2)$, and $\{\eta_t\}$, $\{\delta_t\}$ and $\{\varepsilon_t\}$ mutually independent. We also let

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

$\beta_1 = \beta_2 = 1$, and $\phi = \theta_1/(1-\theta_2)$, noting that (53) leads to $\gamma_z(0) = 1 = \gamma_w(0)$, where $\gamma_z(j) = \mathrm{E}(z_t z_{t+j})$ and $\gamma_w(j) = \mathrm{E}(w_t w_{t+j})$. 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 $\mathrm{MI}_h(l)$ and $L_h(l)$ with $l = 1, 2$. It is shown in Table 3 that $\mathrm{MI}_2(1) = \mathrm{MI}_2(2)$ 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 3 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 4 ($h = 2$) and Table 5 ($h = 3$). 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 4 and 5, 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

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

Table 3. 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
$L_h(1) - L_h(2)$	-0.716	-0.966	0.959	1.873
The better predictive model	J_1	J_1	J_2	J_2
	$h = 3$			
$\mathrm{MI}_h(1) - \mathrm{MI}_h(2)$	0.269	0.428	0.232	0.882
$L_h(1) - L_h(2)$	*	*	*	*
The better predictive model	J_2	J_2	J_2	J_2

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

which is a consistent estimator of $\sigma_h^{-2}(l)\mathbf{R}^{-1}(l)\mathbf{C}_{h,0}(l)$. Note first that 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) the trace of $\widehat{\mathbf{R}}_N^{-1}(l)\widehat{\mathbf{C}}_{h,0}(l)$ 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 4, 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 5, 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,

$$y_{t+2} = \frac{1}{1 - aB}x_t + \frac{1}{1 - bB}z_t + \varepsilon_{t+2}, \tag{54}$$

Table 4. 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 5. 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

in which $|a| < 1$, $|b| < 1$, $\varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2)$, $x_t \sim \text{NID}(0, \sigma_x^2)$, $z_t \sim \text{NID}(0, \sigma_z^2)$, and $\{\varepsilon_t\}$, $\{x_t\}$ and $\{z_t\}$ are independent. Note that model (54) is nonlinear in the parameters. With obersvations up to time n, we are interested in predicting y_{n+2} using a model chosen from

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

both being misspecified. Since $g_{n,2}^{(1)}(\alpha)$ is independent of $\varepsilon_{n,2}^{(1)}$ and $g_{n,2}^{(2)}(\beta)$ is independent of $\varepsilon_{n,2}^{(2)}$, 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 4.2. The coefficients in (54) are set to:

DGP I: $(a, b, \sigma_\varepsilon^2, \sigma_x^2, \sigma_z^2) = (0.5, \text{NA}, 1, 1, 1)$,

DGP II: $(a, b, \sigma_\varepsilon^2, \sigma_x^2, \sigma_z^2) = (0.95, 0.65, 1, 0.4109, 1.000)$,

DGP III: $(a, b, \sigma_\varepsilon^2, \sigma_x^2, \sigma_z^2) = (0.4, -0.95, 0.25, 1.4676, 0.5)$,

DGP IV: $(a, b, \sigma_\varepsilon^2, \sigma_x^2, \sigma_z^2) = (0.8, -0.4, 1, 1.3093, 2)$.

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 6 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 DGP 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 7, 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 (47) 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 $\widehat{\mathbf{R}}_N(l)$ 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 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. Finally, we mention that 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.

Example 3. To evaluate the performance of the criteria mentioned in the previous examples in

Table 6. 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.

Table 7. 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

high-dimensional misspecified models, we consider the following one-step predictive model,

$$y_{t+1} = \mathbf{x}_t^\top \boldsymbol{\beta} + x_{t,p+1} + \varepsilon_{t+1}, \tag{55}$$

where $\mathbf{x}_t = (x_{t1}, \ldots, x_{tp})^\top$ is a p-dimensional explanatory vector and i.i.d. $N(\mathbf{0}_p, I_p)$ distributed with $\mathbf{0}_p$ and I_p denoting the p-dimensional vector of zeros and the p-dimensional identity matrix, p is allowed to be larger than n, $x_{t,p+1} = x_{t1}x_{t2}$ is an interaction term which 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 $|\phi_1| < 1$, $\eta_t \sim \mathrm{NID}(0, \sigma^2)$, and $\{\eta_t\}$ is independent of $\{\mathbf{x}_t\}$. Although the data are generated from model (55), we fit a linear regression model without interaction,

$$y_{t+1} = \mathbf{x}_t^\top \boldsymbol{\beta}^* + \varepsilon_{t+1}^*, \tag{56}$$

as in Example 5.1.2 of Lv and Liu (2014). In the special case of $\phi_1 = 0$, (55) and (56) have been used by Lv and Liu (2014) to illustrate the advantage of GBIC$_p$ with respect to AIC, BIC, GAIC and GBIC when $p \geq n$. To highlight MRIC's efficacy in dealing with dependent data, ϕ_1 is set to 0.8 here. On the other hand, following Lv and Liu (2014), we let $\boldsymbol{\beta} = (1, -1.25, 0.75, -0.95, 1.5, \mathbf{0}_{p-5}^\top)^\top$ and $\sigma = 0.25$, which, together with $\phi_1 = 0.8$, yields $\sigma_\varepsilon = 0.417$, where σ_ε is the standard error of ε_t.

The (n, p) combinations considered in this example are $\{200, 500, 1000\} \times \{100, 200, 1000\}$. Since it is unrealistic to implement best subset regression due to $p \geq 100$, we use the orthogonal greedy algorithm (OGA) of Ing and Lai (2011) to *sequentially* include K_n variables, where $K_n = 5\sqrt{n/\log p}$ is suggested in Section 5 of the same paper. We then apply MRIC (with $\alpha_m = 0.6, 0.7$ and denoting the MRIC by MRIC(0.6) and MRIC(0.7) respectively later in this example), and other five criteria to choose models along the OGA path. It is not difficult to show that $J^* = \{1, \ldots, 5\}$ is the "oracle" working model in the sense that

$$\begin{aligned}
&\mathrm{MI}_1(J^*) \leq \mathrm{MI}_1(J) \text{ for any } J \subset \{1, \ldots, p\} \text{ with } \sharp(J) \leq K_n, \\
&\mathrm{L}_1(J^*) < \mathrm{L}_1(J) \text{ for any } J \neq J^* \text{ with } \mathrm{MI}_1(J^*) = \mathrm{MI}_1(J),
\end{aligned}$$

where $\mathrm{MI}_1(J)$ and $\mathrm{L}_1(J)$ denote the MI and VI of model J respectively.

In view of the optimality of J^*, 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,

$$\begin{aligned}
&\text{expected number of true positives (ENTP)}: \frac{1}{1000} \sum_{i=1}^{1000} \sharp(\hat{J}^{(i)} \bigcap J^*), \\
&\text{expected number of true negatives (ENTN)}: \frac{1}{1000} \sum_{i=1}^{1000} \sharp(\hat{J}^{(i)} \bigcap J^{*^c}), \\
&\text{selection probability (SP)}: \frac{\sum_{i=1}^{1000} I_{\{\hat{J}^{(i)} = J^*\}}}{1000},
\end{aligned}$$

where $J^{*^c} = \{1, \ldots, p\} - J^*$. We also summarize the performance of the seven criteria mentioned in the previous paragraph in Table 8. As observed in Table 8, all criteria have ENTP values equal to $5 = \sharp(J^*)$, except for the few cases of (n, p) in which the ENTP values of MRIC(0.6) and MRIC(0.7) fall between 4.980 and 4.999. This result not only suggests that all criteria, in conjunction with OGA, enjoy high true positive rates, but also reveals that OGA possesses the so-called "sure screening property" (Fan and Lv, 2008), meaning that the probability of the OGA path containing all relevant variables in the candidate set approaches 1 as $n \to \infty$. In fact, this property has been already established for OGA under high-dimensional misspecified models without assuming that the relevant variables omitted from the model are uncorrelated with candidate variables; see Ing et al. (2016). Except in the case of $(n, p) = (200, 1000)$, the ENTN values of MRIC(0.6) and MRIC(0.7) are close to 0 and smaller than those of the BIC-type criteria (BIC, GBIC, GBIC_p), which in turn are much smaller than those of AIC and GAIC. The SP values of MRIC(0.6) and MRIC(0.7) are larger than 0.996 when $n \geq 500$, between 0.743 and 0.939 when $n = 200$ and $p \leq 200$, and near 0 when $(n, p) = (200, 1000)$ due to a severe overfitting problem. While the overfitting problem associated with BIC and GBIC is generally not severe in terms of ENTN when $p \leq 200$, their SP values are substantially smaller than the corresponding values of MRIC(0.6) and MRIC(0.7). Suffering from considerable overfitting, the BIC-type criteria in the case of $p = 1000$ and AIC and GAIC in all cases have SP values very close to 0. When $p \leq 200$, although GBIC_p outperforms BIC and GBIC in terms of SP and ENTN, it is obviously surpassed by MRIC(0.6) and MRIC(0.7). Finally, we remark that when MRIC is used along with OGA, its α_m value is suggested to lie between 0.6 and 0.8.

6. Real data analysis: two cases

In this section, we compare the performance of MRIC and other competing methods using two real datasets. The C_n in MRIC is set to n^{α_m}, where $0 < \alpha_m < 1$ is chosen in a data-driven fashion. 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 1 for the time plot as well as the sample ACF/PACF plot of the resultant series, denoted by $\{S_t\}, 1 \leq t \leq 191$. The goal of this study is to investigate the prediction performance of the criteria considered in Section 5 when they are applied to $\{S_t\}$. For the sake of completeness, our assessment also includes FIC_p (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),

$$\mathrm{EMSPE} = \frac{1}{\lfloor nd \rfloor} \sum_{t=n-h-\lfloor nd \rfloor+1}^{n-h} (S_{t+h} - \hat{S}_{t+h})^2, \tag{57}$$

where d is set to 0.3, $\hat{S}_{t+h}$ is the h-step least squares predictor of S_{t+h} whose order is selected by the criterion and parameters are estimated by least squares using observations up to time t. In this

Table 8. Expected numbers of true positives (ENTP), expected numbers of true negatives (ENTN), selection probabilities (SP), across 1,000 simulations, of the model selection criteria considered in Example 3 when they are used in conjunction with the OGA under model misspecification

		ENTP	ENTN	SP	ENTP	ENTN	SP	ENTP	ENTN	SP
n	Criteria		p=100			p=200			p=1000	
200	AIC	5.000	23.232	0.000	5.000	26.000	0.000	5.000	22.000	0.000
	BIC	5.000	2.937	0.079	5.000	8.719	0.005	5.000	22.000	0.000
	GAIC	5.000	27.630	0.000	5.000	26.000	0.000	5.000	22.000	0.000
	GBIC	5.000	2.268	0.117	5.000	5.283	0.009	5.000	21.736	0.000
	GBICp	5.000	1.538	0.261	5.000	3.890	0.040	5.000	21.751	0.000
	MRIC(0.6)	4.996	0.170	0.859	4.998	0.466	0.743	5.000	21.939	0.001
	MRIC(0.7)	4.986	0.053	0.939	4.980	0.120	0.888	4.999	18.202	0.132
500	AIC	5.000	17.857	0.000	5.000	40.799	0.000	5.000	38.000	0.000
	BIC	5.000	1.400	0.259	5.000	3.026	0.055	5.000	32.100	0.000
	GAIC	5.000	21.336	0.000	5.000	43.776	0.000	5.000	38.000	0.000
	GBIC	5.000	1.375	0.253	5.000	2.856	0.057	5.000	19.631	0.000
	GBICp	5.000	0.783	0.465	5.000	1.662	0.189	5.000	13.942	0.000
	MRIC(0.6)	5.000	0.000	1.000	5.000	0.003	0.997	5.000	0.004	0.996
	MRIC(0.7)	5.000	0.000	1.000	5.000	0.000	1.000	5.000	0.000	1.000
1000	AIC	5.000	16.344	0.000	5.000	36.249	0.000	5.000	55.000	0.000
	BIC	5.000	0.863	0.444	5.000	1.821	0.173	5.000	12.139	0.002
	GAIC	5.000	15.949	0.000	5.000	40.783	0.000	5.000	55.000	0.000
	GBIC	5.000	0.899	0.428	5.000	1.855	0.162	5.000	10.928	0.002
	GBICp	5.000	0.488	0.625	5.000	1.031	0.370	5.000	6.317	0.005
	MRIC(0.6)	5.000	0.000	1.000	5.000	0.000	1.000	5.000	0.000	1.000
	MRIC(0.7)	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.

connection, we also compute

$$\text{EMSPE}_0 = \min_{1\leq k\leq 15} \frac{1}{\lfloor nd\rfloor} \sum_{t=n-h-\lfloor nd\rfloor+1}^{n-h} (S_{t+h} - \hat{S}_{t+h}(k))^2, \tag{58}$$

where $\hat{S}_{t+h}(k)$ is the h-step 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 we choose α_m for MRIC by minimizing the in-sample counterpart of (57),

$$\frac{1}{\lfloor nd\rfloor} \sum_{t=n-2\lfloor nd\rfloor-h+1}^{n-\lfloor nd\rfloor-h} (S_{t+h} - \hat{S}_{t+h}^{(\alpha_m)})^2, \tag{59}$$

over $\alpha_m \in \{0.1, \ldots, 0.8\}$, where $\hat{S}_{t+h}^{(\alpha_m)}$ is $\hat{S}_{t+h}$ with the order selected by MRIC with the corresponding α_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 (59), it will be used throughout the period for forecast evaluation.

The values obtained from (57) and (58), with $h = 1, \ldots, 5$, are summarized in Table 9. As shown in Table 9, 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 2.2 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,

$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - \bar{T}) + \beta_3(T_t - \bar{T})^2 + \beta_4 P_t + w_t, \tag{60}$$

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 (60) 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,

$$\begin{aligned} M_{t+h} &= \beta_0 + \beta_1(t+h) + \sum_{i=1}^{L} \beta_{3,i} M_{t+1-i} \\ &+ \sum_{i=1}^{L} \beta_{4,i} T_{t+1-i} + \sum_{i=1}^{L} \beta_{5,i} T_{t+1-i}^2 + \sum_{i=1}^{L} \beta_{6,i} P_{t+1-i} + \sum_{i=1}^{L} \beta_{7,i} \log P_{t+1-i} + \epsilon_{t,h}, t = L, \ldots, n-h, \end{aligned} \tag{61}$$

where $\epsilon_{t,h}$ denotes the error term. In this study, L is set to 156. 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 2. 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 (61) 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 5, we first use OGA to sequentially select $K_n = 5\sqrt{N_1/\log p_1}$ variables, and then choose models along the OGA path using the criteria considered in the same example. Their performance is evaluated by

$$\widetilde{\text{EMSPE}} = \frac{1}{O_1} \sum_{t=n-h-O_1+1}^{n-h} (M_{t+h} - \hat{M}_{t+h})^2, \tag{62}$$

where $\hat{M}_{t+h}$ is the h-step 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 the purpose of comparison, we also compute the corresponding benchmark value,

$$\widetilde{\text{EMSPE}}_0 = \min_{1 \le k \le K_n} \frac{1}{O_1} \sum_{t=n-h-O_1+1}^{n-h} (M_{t+h} - \hat{M}_{t+h}(k))^2, \tag{63}$$

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. Moreover, as suggested by Example 3 of Section 5, the α_m in MRIC is chosen from among $\{0.5, \ldots, 0.8\}$ using the in-sample counterpart of (62). The resultant values of $\widetilde{\text{EMSPE}}_0$ and $\widetilde{\text{EMSPE}}$, with $1 \le h \le 5$, are documented in Table 10.

As shown in Table 10, the performance of AIC and GAIC is exactly the same in terms of $\widetilde{\text{EMSPE}}$. In addition, BIC and GBIC also behave similarly, and have $\widetilde{\text{EMSPE}}$ values smaller than (close to) those of AIC and GAIC when $h = 1$ and 3 ($h = 2, 4$ and 5). The $\widetilde{\text{EMSPE}}$ values of these four criteria, however, are substantially larger than the values of $\widetilde{\text{EMSPE}}_0$ for all $1 \le h \le 5$, and the discrepancy tends to quickly grow with increasing lead-time h. MRIC obviously outperforms the other criteria from the $\widetilde{\text{EMSPE}}$ point of view, and the difference between its $\widetilde{\text{EMSPE}}$ and the corresponding benchmark value does not seem to be sizeable. In particular, it has the smallest $\widetilde{\text{EMSPE}}$ for $2 \le h \le 5$. GBIC$_p$ is slightly superior to MRIC

Table 9. 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

when $h = 1$. Unfortunately, its performance deteriorates fast as h increases although it still generally overshadows the other four non-MRIC criteria for $h = 2, 3$ and 5.

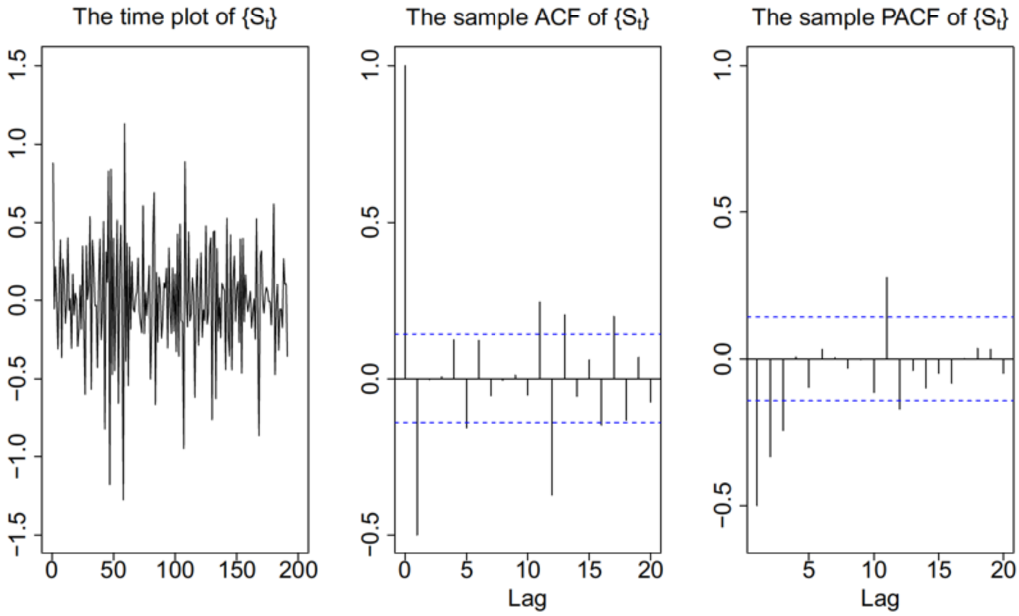

Fig. 1. Plots of series $\{S_t\}$ and its sample ACF and PACF

7. 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 collection of candidate models to increase indefinitely with the sample size. 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 (23), for the penalty on model misspecification and the penalty on sampling variability, which addresses not only the one-step ahead prediction but also

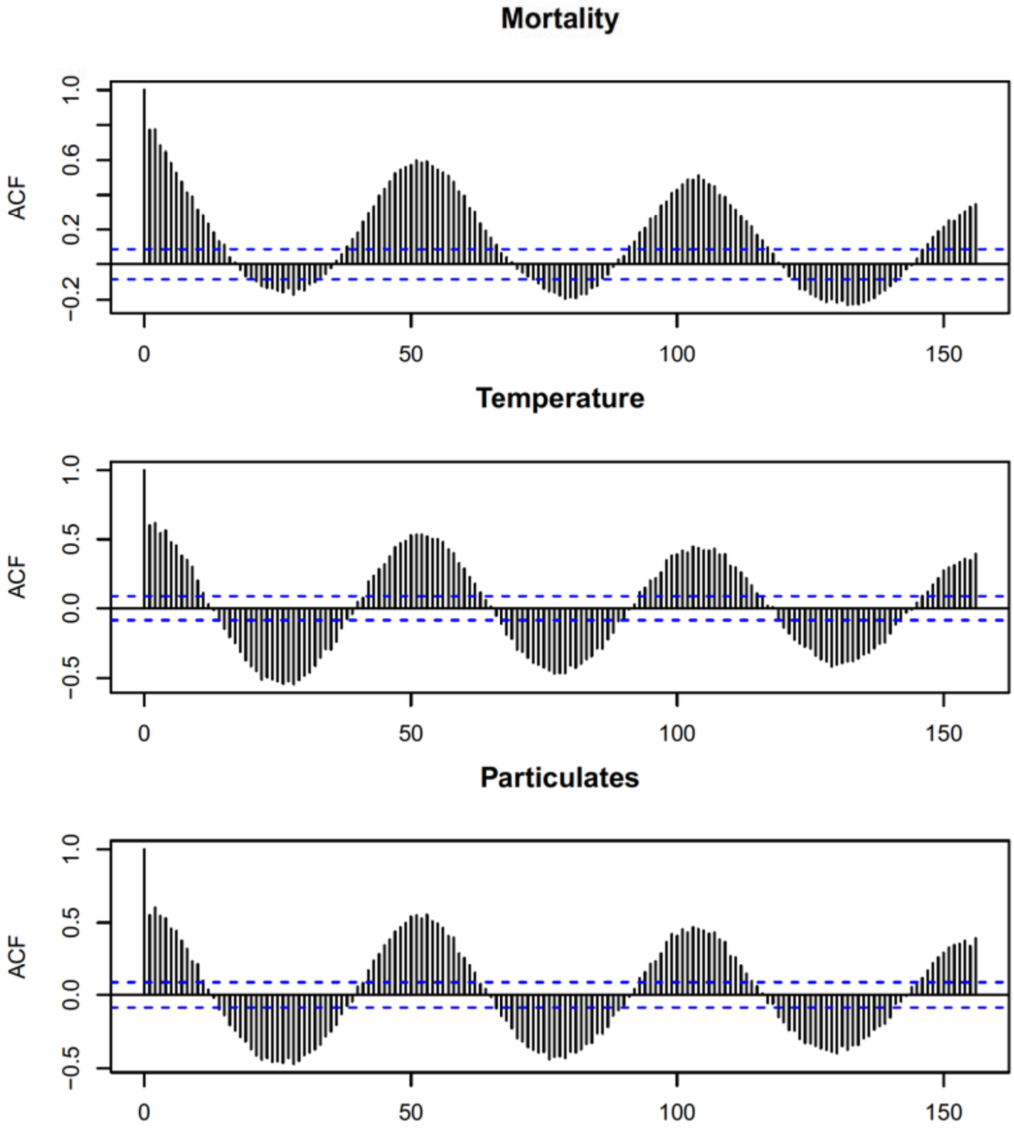

Fig. 2. Sample ACFs of weakly time series for cardiovascular mortality (top), temperature (middle) and particulate pollution (bottom) in Los Angeles County from 1970-1979.

the multi-step case. We have shown how we can compute the explicit expressions and given detailed theoretical underpinnings. The impact of nonlinearity is clarified in equation (47). We have illustrated how the MRIC has addressed the fundamental issue successfully with simulated and real data, including both high-dimensional cases and nonlinear cases. 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

Table 10. The values of $\widetilde{\text{EMSPE}}$ and $\widetilde{\text{EMSPE}}_0$ derived from the mortality data of Shumway et al. (1988).

	$\widetilde{\text{EMSPE}}$						$\widetilde{\text{EMSPE}}_0$
h	AIC	BIC	MRIC	GAIC	GBIC	GBIC$_p$	
1	28.35	22.24	18.99	28.35	22.24	17.60	16.79
2	26.21	27.34	21.89	26.21	27.34	26.77	16.68
3	41.38	38.80	22.90	41.38	38.80	31.29	16.54
4	37.12	37.12	23.45	37.12	37.77	38.33	16.31
5	46.83	46.83	24.22	46.83	46.83	44.28	16.55

(2011) and a recent Workshop entitled *Non-likelihood based statistical modelling* held at the Centre for Research in Statistical Methodology, Warwick University UK, 7-9 September 2015. Therefore, it remains a future challenge to develop a model selection criterion via a non-likelihood-based approach.

A. Proof of Theorem 1

In view of (7), we have

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

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 A.1 below that

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

and

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

By (C1) and (C2), it follows that

$$\begin{aligned}(\mathrm{IV})&=\frac{1}{N}\mathrm{E}\left(\sum_{t=1}^N\mathbf{x}_t^\top\mathbf{R}^{-1}\mathbf{x}_t\varepsilon_{t,h}^2\right)+\frac{2}{N}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\mathrm{E}\left(\mathbf{x}_j^\top\mathbf{R}^{-1}\mathbf{x}_k\varepsilon_{j,h}\varepsilon_{k,h}\right)\\&=\mathrm{tr}\left\{\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_1^\top\varepsilon_{1,h}^2\right)\right\}+\frac{2}{N}\left\{\sum_{j=1}^{N-1}(N-j)\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right)\right\}.\end{aligned}\tag{A.4}$$

Similarly,

$$(\mathrm{III})=-2\left\{\sum_{j=h}^{n-1}\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right)\right\}.\tag{A.5}$$

By (A.1)–(A.5) and (C2),

$$N\left\{\mathrm{E}\left(y_{n+h}-\hat{\boldsymbol{\beta}}_n^\top(h)\mathbf{x}_n\right)^2-\mathrm{E}\left(\varepsilon_{n,h}^2\right)\right\}$$

$$=\operatorname{tr}\left\{\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_1^\top\varepsilon_{1,h}^2\right)\right\}+2\sum_{s=1}^{h-1}\operatorname{tr}\left\{\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_{1+s}^\top\varepsilon_{1,h}\varepsilon_{1+s,h}\right)\right\}$$

$$-\frac{2}{N}\sum_{j=1}^{N-1}j\,\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right)-2\sum_{j=N}^{n-1}\mathrm{E}\left(\mathbf{x}_1^\top\mathbf{R}^{-1}\mathbf{x}_{j+1}\varepsilon_{1,h}\varepsilon_{j+1,h}\right)+o(1)$$

$$=\operatorname{tr}\left(\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_1^\top\varepsilon_{1,h}^2\right)\right)+2\sum_{s=1}^{h-1}\operatorname{tr}\left(\mathbf{R}^{-1}\mathrm{E}\left(\mathbf{x}_1\mathbf{x}_{1+s}^\top\varepsilon_{1,h}\varepsilon_{1+s,h}\right)\right)+o(1),$$

yielding the desired conclusion. □

Lemma A.1. *Under the assumptions of Theorem 1,* (A.2) *and* (A.3) *follow.*

Proof. It suffices for (A.2) to prove that

$$\mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^\top\left(\widehat{\mathbf{R}}_N^{-1}-\mathbf{R}^{-1}\right)\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right)=o(1). \tag{A.6}$$

To show (A.6), first observe that

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

$$=\mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^\top\left(\widehat{\mathbf{R}}_N^{-1}-\widehat{\mathbf{R}}_{n-l_n}^{-1}\right)\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right)+\mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^\top\left(\widehat{\mathbf{R}}_{n-l_n}^{-1}-\mathbf{R}^{-1}\right)\sum_{t=n-l_n-h+1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right) \tag{A.7}$$

$$+\mathrm{E}\left(\varepsilon_{n,h}\mathbf{x}_n^\top\left(\widehat{\mathbf{R}}_{n-l_n}^{-1}-\mathbf{R}^{-1}\right)\sum_{t=1}^{n-l_n-h}\mathbf{x}_t\varepsilon_{t,h}\right):=\text{(I)}+\text{(II)}+\text{(III)}.$$

Since for large n,

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

This, together with (11) (in (C5)), (8) (in (C1)), Hölder's inequality and the hypothesis that $l_n=o(n^{1/2})$ (see (C6)), yields for any $0<\gamma\leq 5$,

$$\mathrm{E}\|\widehat{\mathbf{R}}_N^{-1}-\widehat{\mathbf{R}}_{n-l_n}^{-1}\|^\gamma=O\left((l_n/n)^\gamma\right). \tag{A.8}$$

Applying Hölder's inequality again, we have

$$\text{(I)}\leq\left(\mathrm{E}|\varepsilon_{n,h}|^6\right)^{1/6}\left(\|\mathbf{x}_n\|^6\right)^{1/6}\left(\mathrm{E}\|\widehat{\mathbf{R}}_N^{-1}-\widehat{\mathbf{R}}_{n-l_n}^{-1}\|^3\right)^{1/3}\left(\mathrm{E}\left\|\sum_{t=1}^{N}\mathbf{x}_t\varepsilon_{t,h}\right\|^3\right)^{1/3}. \tag{A.9}$$

By (A.8), (A.9), (C3) and (C4), it holds that

$$\text{(I)}=o(l_n/n^{1/2})=o(1). \tag{A.10}$$

An argument similar to that used to prove (A.8) yields for any $0<\gamma\leq 5$,

$$\mathrm{E}\|\widehat{\mathbf{R}}_N^{-1}-\mathbf{R}^{-1}\|^\gamma=O(n^{-\gamma/2}),\quad \mathrm{E}\|\widehat{\mathbf{R}}_{n-l_n}^{-1}-\mathbf{R}^{-1}\|^\gamma=O(n^{-\gamma/2}). \tag{A.11}$$

By making use of (A.11), Hölder's inequality, (C3), (C4) and (13) (in (C6)), we obtain

$$(\text{II}) = O\left((l_n/n)^{1/2}\right) = o(1), \tag{A.12}$$

and

$$(\text{III}) \leq \left(\mathrm{E}\|\widehat{\mathbf{R}}_{n-l_n}^{-1} - \mathbf{R}^{-1}\|^3\right)^{1/3} \left(\mathrm{E}\left\|\sum_{t=1}^{n-l_n-h} \mathbf{x}_t\varepsilon_{n,h}\right\|^3\right)^{1/3} \left(\mathrm{E}\left(\|\mathrm{E}\left(\mathbf{x}_n\varepsilon_{n,h}|\mathcal{F}_{n-l_n}\right)\|^3\right)\right)^{1/3} = o(1). \tag{A.13}$$

Consequently, (A.6) follows from (A.7), (A.10), (A.12) and (A.13).

To show (A.3), let

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

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

It follows that

$$\mathrm{E}(\mathbf{x}_n^\top \widehat{\mathbf{R}}_N^{-1} N^{-1/2}\sum_{t=1}^{N} \mathbf{x}_n\varepsilon_{t,h})^2 = \mathrm{E}(\mathbf{M}_1^2) + \mathrm{E}(\mathbf{M}_2^2) + 2\mathrm{E}(\mathbf{M}_1\mathbf{M}_2). \tag{A.14}$$

Moreover, by (C3), (C4), (A.11), (12) (in (C6)) and Hölder's inequality, we have

$$\mathrm{E}(\mathbf{M}_1^2) \leq (\mathrm{E}\|\mathbf{x}_n\|^{10})^{1/5}(\mathrm{E}\|\widehat{\mathbf{R}}_N^{-1} - \mathbf{R}^{-1}\|^5)^{2/5}(\mathrm{E}\|N^{-1/2}\sum_{t=1}^{N} \mathbf{x}_t\varepsilon_{t,h}\|^5)^{2/5} = O(n^{-1}), \tag{A.15}$$

and

$$\begin{aligned}
\mathrm{E}(\mathbf{M}_2^2) &= \mathrm{E}\left(\mathbf{M}_3\right) + \mathrm{E}\left\{\mathbf{u}_n^\top \mathbf{R}^{-1}\left(\mathrm{E}\left(\mathbf{x}_n\mathbf{x}_n^\top|\mathcal{F}_{n-l_n}\right) - \mathbf{R}\right)\mathbf{R}^{-1}\mathbf{u}_n\right\} + O((l_n/n)^{1/2}) \\
&= \mathrm{E}(\mathbf{M}_3) + O\left(\left\{\mathrm{E}\|\mathbf{u}_n\|^3\right\}^{2/3}\left\{\mathrm{E}\left\|\mathrm{E}\left(\mathbf{x}_n\mathbf{x}_n^\top|\mathcal{F}_{n-l_n}\right) - \mathbf{R}\right\|^3\right\}^{1/3}\right) + O((l_n/n)^{1/2}) \qquad (\text{A.16}) \\
&= \mathrm{E}(\mathbf{M}_3) + o(1).
\end{aligned}$$

Consequently, (A.3) follows from (A.14)–(A.16). □

B. Proof of Theorem 2.

It suffices to show (26) and (27). To show (26), first note that

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

In addition, by (C1) and the positive definiteness of $\mathbf{R}(l) = \mathrm{E}(\mathbf{x}_t(J_l)\mathbf{x}_t^\top(J_l))$ (as ensured by (C5)),

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

which, together with (C4), yields $N^{-2}(\sum_{t=1}^{N}\mathbf{x}_t^\top(J_l)\varepsilon_{t,h}^{(l)})\widehat{\mathbf{R}}_N^{-1}(l)(\sum_{t=1}^{N}\mathbf{x}_t(J_l)\varepsilon_{t,h}^{(l)}) = O_p(n^{-1})$. Thus, (B.1) and (31) in turn imply (26).

To show (27), we first dissect $\widehat{\mathbf{C}}_{h,s}(J_l)$ as

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

By (B.2), (C1), (C3) and (C4), it can be shown that the last three terms on the right-hand side of (B.3) are of order $o_p(1)$. Combining this with (32) leads to the desired conclusion (27). □

C. Proof of Theorem 3

For notational simplicity, we shall suppress (J_l) in $\mathbf{x}_t(J_l)$ and (l) in $\varepsilon_{t,h}^{(l)}$, $J_v^{(l)}$ and $\mathbf{C}_{h,s}(l)$. It follows from (38)–(41) and (45) that

$$y_t = \sum_{j=0}^{\infty}\mathbf{w}_{j,y}^\top\mathbf{v}_{t-j},\ s_t^{(v)} = \sum_{j=0}^{\infty}\mathbf{w}_{j,v}^\top\mathbf{v}_{t-j},\ \varepsilon_{t,h} = \sum_{j=0}^{\infty}\mathbf{w}_{j,0}^\top\mathbf{v}_{t+h-j},\tag{C.1}$$

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

$$\|\mathbf{w}_{j,y}\| \le c^*(j+1)^{-s},\ \|\mathbf{w}_{j,v}\| \le c^*(j+1)^{-s},\tag{C.2}$$

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

$$\mathrm{E}(\varepsilon_{t,h}y_{t-j}) = 0, j \in J_0 \text{ and } \mathrm{E}(\varepsilon_{t,h}s_{t-j}^{(v)}) = 0, j \in J_v, 1 \le v \le p^*,\tag{C.3}$$

it holds that

$$\begin{aligned}&\sum_{k=0}^{\infty}\mathbf{w}_{k,y}^\top\Lambda\mathbf{w}_{k+h+j,0} = 0,\ j \in J_0,\\ &\sum_{k=0}^{\infty}\mathbf{w}_{k,v}^\top\Lambda\mathbf{w}_{k+h+j,0} = 0,\ j \in J_v, 1 \le v \le p^*,\end{aligned}\tag{C.4}$$

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

By (43) and (C.1)–(C.3), it is not difficult to show that conditions (C3) and (C6) follow. Moreover, since (C.2) ensures that the autocovariance functions of y_t, $s_t^{(v)}$ and $\varepsilon_{t,h}$ are square summable, by (43), (C.1), (C.3) 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 C.1 below. The first statement of condition (C2) is obviously guaranteed by (C.1) and the hypothesis that the fourth moments of $\{\mathbf{v}_t\}$ are independent of t, whereas the second one holds if

$$\text{each component of } n\mathrm{E}(\mathbf{x}_1\mathbf{x}_n^\top\varepsilon_{1,h}\varepsilon_{n,h}) \text{ converges to } 0, \tag{C.5}$$

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

In the following, we shall prove

$$\mathrm{E}(\varepsilon_{1,h}y_1\varepsilon_{n,h}y_n) = o(n^{-1}) \tag{C.6}$$

instead of (C.5) 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

$$\begin{aligned}
&\mathrm{E}(\varepsilon_{1,h}y_1\varepsilon_{n,h}y_n) = \mathrm{E}(\varepsilon_{1,h}y_1\tilde{\varepsilon}_{n,h}\tilde{y}_n)\\
&= \mathrm{E}\{\sum_{j_1=-\infty}^{1+h}\sum_{j_2=-\infty}^{1+h}\sum_{j_3=-\infty}^{1+h}\sum_{j_4=-\infty}^{1+h}\mathbf{w}_{1-j_1,y}^\top\mathbf{v}_{j_1}\mathbf{w}_{1+h-j_2,0}^\top\mathbf{v}_{j_2}\mathbf{w}_{n-j_3,y}^\top\mathbf{v}_{j_3}\mathbf{w}_{n+h-j_4,0}^\top\mathbf{v}_{j_4}\}\\
&= \mathrm{E}\{\sum_{j=-\infty}^{1+h}\mathbf{w}_{1-j,y}^\top\mathbf{v}_j\mathbf{w}_{1+h-j,0}^\top\mathbf{v}_j\mathbf{w}_{n-j,y}^\top\mathbf{v}_j\mathbf{w}_{n+h-j,0}^\top\mathbf{v}_j\}\\
&+ \mathrm{E}\{\sum_{\substack{-\infty<m,k\leq 1+h\\ m\neq k}}\mathbf{w}_{1-m,y}^\top\mathbf{v}_m\mathbf{v}_m^\top\mathbf{w}_{1+h-m,0}\mathbf{w}_{n-k,y}^\top\mathbf{v}_k\mathbf{v}_k^\top\mathbf{w}_{n+h-k,0}\}\\
&+ \mathrm{E}\{\sum_{\substack{-\infty<m,k\leq 1+h\\ m\neq k}}\mathbf{w}_{1-m,y}^\top\mathbf{v}_m\mathbf{v}_m^\top\mathbf{w}_{n-m,y}\mathbf{w}_{1+h-k,0}^\top\mathbf{v}_k\mathbf{v}_k^\top\mathbf{w}_{n+h-k,0}\}\\
&+ \mathrm{E}\{\sum_{\substack{-\infty<m,k\leq 1+h\\ m\neq k}}\mathbf{w}_{1-m,y}^\top\mathbf{v}_m\mathbf{v}_m^\top\mathbf{w}_{n+h-m,0}\mathbf{w}_{1+h-k,0}^\top\mathbf{v}_k\mathbf{v}_k^\top\mathbf{w}_{n-k,y}\}\\
&= (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}) + (\mathrm{IV}),
\end{aligned} \tag{C.7}$$

where $\mathbf{w}_{s,y}$ is set to $\mathbf{0}$ when $s < 0$. By (43) and (C.2), it holds that

$$|(\mathrm{I})| \leq \sum_{j=-\infty}^{1+h}\mathrm{E}\|\mathbf{v}_j\|^4(\|\mathbf{w}_{1-j,y}\|\|\mathbf{w}_{1+h-j,0}\|\|\mathbf{w}_{n-j,y}\|\|\mathbf{w}_{n+h-j,0}\|) = O(n^{-2s}) = o(n^{-1}). \tag{C.8}$$

Using the first relation of (C.4) with $j = 0$, we obtain

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

Therefore,

$$|(\mathrm{II})| \leq \|\Lambda\|^2\sum_{k=-\infty}^{1+h}\|\mathbf{w}_{n-k,y}\|\|\mathbf{w}_{n+h-k,0}\|\|\mathbf{w}_{1-k,y}\|\|\mathbf{w}_{1+h-k,0}\| = O(n^{-2s}) = o(n^{-1}). \tag{C.9}$$

Straightforward calculations and (C.2) yield

$$\begin{aligned}
|(\mathrm{III})| &= O\left(\sum_{k=-\infty}^{1+h}\|\mathbf{w}_{1+h-k,0}\|\|\mathbf{w}_{n+h-k,0}\|(\sum_{m=-\infty,m\neq k}^{1+h}\|\mathbf{w}_{1-m,y}\|\|\mathbf{w}_{n-m,y}\|)\right)\\
&= O(n^{-4s+2}) = o(n^{-1}).
\end{aligned} \tag{C.10}$$

Similarly,

$$|(\mathrm{IV})| = O(n^{-4s+2}) = o(n^{-1}). \tag{C.11}$$

The desired conclusion (C.6) (and hence (C.5)) now follows from (C.7)–(C.11).

It remains to show that (31) and (32) hold true. Note first that (31) is an immediate consequence of (C.1), (C.2), and the First Moment Bound Theorem of Findley and Wei (1993). To prove (32), 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 (43), (C.1), (C.2) and tedious algebraic manipulations, we obtain for each $1 \le i,j \le S_{p^*} \equiv \sum_{v=0}^{p^*} \sharp(J_v)$,

$$\sup_{|m-k|=r} |\mathrm{E}(D_{h,s}^{(m)}(i,j) D_{h,s}^{(k)}(i,j))| \to 0, \text{ as } r \to \infty, \tag{C.12}$$

yielding $n^{-1} \sum_{t=1}^{n} D_{h,s}^{(t)}(i,j) = o_p(1)$, which in turn implies (32). Consequently, the proof of Theorem 3 is complete. □

Lemma C.1. *Assume* (38)–(41) *and* (44). *Then* (C5) *follows.*

Proof. In view of Theorem 2.1 of Chan and Ing (2011), it suffices for (C5) to show that (18) holds true. By (C.1), 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, it is not difficult to see 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 exist a positive integer D such that for all $t \ge D$,

$$\lambda_{\min}\left(\mathrm{E}(\mathbf{x}_{t,D} \mathbf{x}_{t,D}^\top)\right) > \delta_1^*, \tag{C.13}$$

where $\mathbf{x}_{t,D} = \sum_{j=0}^{D-1} H_j \mathbf{v}_{t-j}$. Since (C.13) 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

$$\boldsymbol{s}^\top H_{j(\boldsymbol{s})} \Lambda H_{j(\boldsymbol{s})}^\top \boldsymbol{s} \ge \delta_1^*/D. \tag{C.14}$$

Define $\mathcal{F}_{t,j(\boldsymbol{s})} = \left\{\mathbf{v}_t, \ldots, \mathbf{v}_{t-j(\boldsymbol{s})+1}, \mathbf{v}_{t-j(\boldsymbol{s})-1}, \ldots\right\}$ and $\eta_{j(\boldsymbol{s})} = \boldsymbol{s}^\top H_{j(\boldsymbol{s})} H_{j(\boldsymbol{s})}^\top \boldsymbol{s}$. Then, by (44) and (C.14), 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 (44),

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

recalling that $\mathcal{F}_t = \sigma(\mathbf{v}_t, \mathbf{v}_{t-1}, \ldots)$. Consequently, (18) 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.

D. Proofs and Theorems 4 and 5

Proof of Theorem 4. Note first that

$$\begin{aligned} n\left\{\mathrm{E}\left(y_{n+h}-g_{n,h}(\hat{\boldsymbol{\theta}}_n)\right)^2-\mathrm{E}(\varepsilon^2_{n,h}(\boldsymbol{\theta}^*))\right\} &= \mathrm{E}\left\{n\left(\varepsilon_{n,h}(\hat{\boldsymbol{\theta}}_n)-\varepsilon_{n,h}(\boldsymbol{\theta}^*)\right)^2\right\} \\ +2\mathrm{E}\left\{n(\varepsilon_{n,h}(\hat{\boldsymbol{\theta}}_n)-\varepsilon_{n,h}(\boldsymbol{\theta}^*))\varepsilon_{n,h}(\boldsymbol{\theta}^*)\right\} &\equiv \mathrm{E(I)}+2\mathrm{E(II)} \end{aligned} \tag{D.1}$$

Let $B_n=\{\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*\|<\delta \text{ and } \|\hat{\boldsymbol{\theta}}_{n-\ell_n}-\boldsymbol{\theta}^*\|<\delta\}$, and define $\boldsymbol{w}_n=D_1g_{n,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*)$. By Taylor's theorem for multivariable functions, we obtain

$$\begin{aligned} \mathrm{(II)} &= -n\boldsymbol{w}_n^\top(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\mathbf{1}_{B_n}-\frac{n}{2}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)^\top\left(D_2g_{n,h}(\breve{\boldsymbol{\theta}}_n)\varepsilon_{n,h}(\boldsymbol{\theta}^*)\right)(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\mathbf{1}_{B_n} \\ &+n\left(\varepsilon_{n,h}(\hat{\boldsymbol{\theta}}_n)-\varepsilon_{n,h}(\boldsymbol{\theta}^*)\right)\varepsilon_{n,h}(\boldsymbol{\theta}^*)\mathbf{1}_{B_n^c}\equiv\mathrm{(III)}+\mathrm{(IV)}+\mathrm{(V)}, \end{aligned} \tag{D.2}$$

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

$$\mathrm{E(III)} \;=\; -\sum_{j=h}^{n-h}\mathrm{E}\Big(D_1^\top g_{1,h}(\boldsymbol{\theta}^*)(\mathbf{R}^*-\mathbf{A}^*)^{-1}D_1g_{1+j,h}(\boldsymbol{\theta}^*)\varepsilon_{1,h}(\boldsymbol{\theta}^*)\varepsilon_{1+j,h}(\boldsymbol{\theta}^*)\Big)+o(1). \tag{D.3}$$

Let $\tilde{\mathbf{R}}_n=n^{-1}\sum_{t=1}^{n-h}D_1g_{t,h}(\boldsymbol{\theta}^*)D_1^\top g_{t,h}(\boldsymbol{\theta}^*)$ and $\tilde{\mathbf{A}}_n=n^{-1}\sum_{t=1}^{n-h}D_2g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)$. Then, by the mean value theorem for vector-valued functions, we have on B_n,

$$\mathbf{0}=D_1S_n(\hat{\boldsymbol{\theta}}_n)=D_1S_n(\boldsymbol{\theta}^*)+\left\{\int_0^1 D_2S_n(\boldsymbol{\theta}^*+r(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*))dr\right\}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*),$$

yielding

$$\begin{aligned} \mathrm{(III)} &= -\boldsymbol{w}_n^\top(\mathbf{R}^*-\mathbf{A}^*)^{-1}\sum_{t=1}^{n-h}D_1g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)\mathbf{1}_{B_n} \\ &+\boldsymbol{w}_n^\top(\mathbf{R}^*-\mathbf{A}^*)^{-1}\left\{\sqrt{n}(\tilde{\mathbf{R}}_n-\mathbf{R}^*)-\sqrt{n}(\tilde{\mathbf{A}}_n-\mathbf{A}^*)\right\}\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\mathbf{1}_{B_n} \\ &+\frac{1}{2}\boldsymbol{w}_n^\top(\mathbf{R}^*-\mathbf{A}^*)^{-1}\int_0^1\Big[D_2S_n(\boldsymbol{\theta}^*+r(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*))-D_2S_n(\boldsymbol{\theta}^*)\Big]dr(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*)\mathbf{1}_{B_n} \\ &\equiv \mathrm{(VI)}+\mathrm{(VII)}+\mathrm{(VIII)}. \end{aligned} \tag{D.4}$$

Define $\mathbf{R}^*_{n-l_n}=n^{-1}\sum_{t=1}^{n-l_n}D_1g_{t,h}(\boldsymbol{\theta}^*)D_1^\top g_{t,h}(\boldsymbol{\theta}^*)$ and $\mathbf{A}^*_{n-l_n}=n^{-1}\sum_{t=1}^{n-l_n}D_2g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*)$. Then, it follows that

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

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

$$|\mathrm{E(VII)}|=o(1). \tag{D.5}$$

We next show that

$$|\mathrm{E}(\mathrm{VIII})| = o(1), \tag{D.6}$$

whose proof is somewhat tricky. Express (VIII) as

$$\begin{aligned}
&\frac{1}{2}\boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\int_0^1 D_2S_n\left(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\right) - D_2S_n\left(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_{n-l_n} - \boldsymbol{\theta}^*)\right)dr(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\mathbf{1}_{B_n}\\
&+\frac{1}{2}\boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\int_0^1 D_2S_n\left(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_{n-l_n} - \boldsymbol{\theta}^*)\right) - D_2S_n(\boldsymbol{\theta}^*)dr(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\mathbf{1}_{B_n}\\
&\equiv (\mathrm{IX}) + (\mathrm{X})
\end{aligned} \tag{D.7}$$

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 $C^* > 0$,

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

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

$$\mathrm{E}|(\mathrm{IX})| = o(1). \tag{D.8}$$

Assumptions (E3) and (E5) also imply

$$\begin{aligned}
\mathrm{E}(\mathrm{X}) = \mathrm{E}\left\{\frac{1}{2}\boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\int_0^1 D_2S_{n-l_n}\left(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_{n-l_n} - \boldsymbol{\theta}^*)\right) - D_2S_{n-l_n}(\boldsymbol{\theta}^*)dr(\hat{\boldsymbol{\theta}}_{n-l_n} - \boldsymbol{\theta}^*)\right\}\\
+\,o(1)
\end{aligned} \tag{D.9}$$

Using (D.9), (E6) and an argument similar to that used to prove (D.5) and (D.8), we obtain $|\mathrm{E}(\mathrm{X})| = o(1)$. In view of this, (D.7) and (D.8), (D.6) follows.

To deal with (VI), note that

$$\begin{aligned}
(\mathrm{VI}) = -\boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\sum_{t=1}^{n-h}D_1g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*) + \boldsymbol{w}_n^\top(\mathbf{R}^* - \mathbf{A}^*)^{-1}\sum_{t=1}^{n-h}D_1g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*)\mathbf{1}_{B_n^c}\\
\equiv (\mathrm{XI}) + (\mathrm{XII})
\end{aligned} \tag{D.10}$$

It follows from (E2) that

$$\mathrm{E}(\mathrm{XI}) = -\sum_{j=h}^{n-1}\mathrm{E}\left(D_1^\top g_{1,h}(\boldsymbol{\theta}^*)(\mathbf{R}^* - \mathbf{A}^*)^{-1}D_1g_{1+j,h}(\boldsymbol{\theta}^*)\varepsilon_{1,h}(\boldsymbol{\theta}^*)\varepsilon_{1+j,h}(\boldsymbol{\theta}^*)\right). \tag{D.11}$$

Assumptions (E3) and (E5) further yield

$$|\mathrm{E}(\mathrm{XII})| = o(1). \tag{D.12}$$

Consequently, (D.3) is guaranteed by (D.4)–(D.6) and (D.10)–(D.12).

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

$$\mathrm{E}(\text{-2(IV)}) = \mathrm{E}\left(n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^\top \mathbf{A}^*(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\mathbf{1}_{B_n}\right) + \mathbf{G}_n, \tag{D.13}$$

where

$$\begin{aligned}
|\mathbf{G}_{\mathrm{n}}| \le\ & \mathrm{E}\left(n \sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)} \|D_3 g_{n,h}(\boldsymbol{\theta})\|_F \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|^3 |\varepsilon_{n,h}(\boldsymbol{\theta}^*)| \mathbf{1}_{B_n}\right) \\
& + \mathrm{E}\left\{\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n-l_n})\| \left(\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\| + \|\sqrt{n}(\hat{\boldsymbol{\theta}}_{n-l_n} - \boldsymbol{\theta}^*)\|\right) \|D_2 g_{n,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*) - \mathbf{A}^*\|\right\} \\
& + \mathrm{E}\left(\|\mathrm{E}\left(D_2 g_{n,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*)\middle|\mathcal{F}_{n-l_n}\right) - \mathbf{A}^*\| \|\sqrt{n}(\hat{\boldsymbol{\theta}}_{n-l_n} - \boldsymbol{\theta}^*)\|^2\right) \\
& + \mathrm{E}\left(\left\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\right\|^2 \|D_2 g_{n,h}(\boldsymbol{\theta}^*)\varepsilon_{n,h}(\boldsymbol{\theta}^*) - \mathbf{A}^*\| \mathbf{1}_{B_n^c}\right).
\end{aligned} \tag{D.14}$$

Let ξ be an arbitrarily small positive number. It is not difficult to see that the first term on the right-hand side of (D.14) 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 rest three terms on the right-hand side of (D.14) also vanish asymptotically. As a result,

$$|\mathbf{G}_{\mathrm{n}}| = o(1). \tag{D.15}$$

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 (D.15) and (D.13) gives

$$\mathrm{E}((\mathrm{IV})) = \frac{-1}{2}\mathrm{E}\{n^{-1}(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))V^*(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))\} + o(1). \tag{D.16}$$

It follows from (E3), (E5) and Hölder's inequality that $|\mathrm{E(V)}| = o(1)$. In view of this, (D.2), (D.3) and (D.16), we obtain

$$\begin{aligned}
2\mathrm{E(II)} = & -2\sum_{j=h}^{n-h} \mathrm{E}(D_1^\top g_{1,h}(\boldsymbol{\theta}^*)(\mathbf{R}^* - \mathbf{A}^*)^{-1} D_1 g_{1+j,h}(\boldsymbol{\theta}^*)\varepsilon_{1,h}(\boldsymbol{\theta}^*)\varepsilon_{1+j,h}(\boldsymbol{\theta}^*)) \\
& - \mathrm{E}\{n^{-1}(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))V^*(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))\} + o(1).
\end{aligned} \tag{D.17}$$

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

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

where $\|\breve{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \leq \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|$. An argument similar to that used to prove (D.17) yields

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

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

$$\mathrm{E}(\mathrm{I}) = \mathrm{E}\{n^{-1}(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))Q^*(\sum_{t=1}^{n-h} D_1 g_{t,h}(\boldsymbol{\theta}^*)\varepsilon_{t,h}(\boldsymbol{\theta}^*))\} + o(1). \tag{D.18}$$

Finally, the desired conclusion is ensured by (D.1), (D.17), (D.18) and (E2).

Proof of Theorem 5. Equation (49) is an immediate consequence of Theorem 4. Equation (48) is ensured by

$$\begin{aligned}
&\frac{S_n^{(l)}(\hat{\boldsymbol{\theta}}_{nl})}{N} = V_l(\boldsymbol{\theta}_l^*) + O_p(\frac{1}{\sqrt{n}}),\\
&\widehat{\mathbf{C}}_{h,s}^*(l) = \mathbf{C}_{h,s}^*(l) + o_p(1),\\
&\widehat{\mathbf{R}}^*(l) = \mathbf{R}^*(l) + o_p(1),\\
&\widehat{\mathbf{A}}^*(l) = \mathbf{A}^*(l) + o_p(1),
\end{aligned} \tag{D.19}$$

which follow from (50)–(52) and an argument similar to that used to prove Theorem 4.

E. 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

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

and

$$\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_{n-l_n})\|^q = o(1). \tag{E.2}$$

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

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta} |n^{-1}\sum_{t=1}^{n}(g_{t,h}(\boldsymbol{\theta}^*) - g_{t,h}(\boldsymbol{\theta}))\eta_{t,h}| = o_p(1),\\
&\sup_{\boldsymbol{\theta}\in\Theta} |n^{-1}\sum_{t=1}^{n}(f_{t,h} - g_{t,h}(\boldsymbol{\theta}))^2 - V_o(\boldsymbol{\theta})| = o_p(1).
\end{aligned} \tag{E.3}$$

However, to prove (E.1) and (E.2), we need a strengthened version of (E.3) among other conditions.

THEOREM 6. *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,

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

$$\begin{aligned} &n^{q/2}\mathrm{P}\left(\sup_{\boldsymbol{\theta}\in\Theta} |n^{-1} \sum_{t=1}^{n} (g_{t,h}(\boldsymbol{\theta}^*) - g_{t,h}(\boldsymbol{\theta}))\eta_{t,h}| > \epsilon \right) = o(1), \\ &n^{q/2}\mathrm{P}\left(\sup_{\boldsymbol{\theta}\in\Theta} |n^{-1} \sum_{t=1}^{n} (f_{t,h} - g_{t,h}(\boldsymbol{\theta}))^2 - V_o(\boldsymbol{\theta})| > \epsilon \right) = o(1), \textit{ for any } \epsilon > 0, \end{aligned} \tag{E.5}$$

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

$$\begin{aligned} &n^{q}\mathrm{P}\left(\sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)} n^{-1} \sum_{t=1}^{n} \varepsilon_{t,h}^2(\boldsymbol{\theta}) > \bar{M} \right) = O(1), \\ &n^{q}\mathrm{P}\left(\sup_{\boldsymbol{\theta}\in B_\delta(\boldsymbol{\theta}^*)} n^{-1} \sum_{t=1}^{n} \|D_j g_{t,h}(\boldsymbol{\theta})\|_F^2 > \bar{M} \right) = O(1), j = 1,2,3. \end{aligned} \tag{E.6}$$

Then, (E.1) *and* (E.2) *hold true.*

PROOF. We begin by proving (E.1). There is $C^* > 0$ such that on the set $\{\hat{\boldsymbol{\theta}}_n \in B_\delta(\boldsymbol{\theta}^*)\}$,

$$\begin{aligned} &\left\| (2n)^{-1} \int_0^1 D_2 S_n(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)) - D_2 S_n(\boldsymbol{\theta}^*) dr \right\| \\ &\leq C^* \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \Lambda_n, \end{aligned} \tag{E.7}$$

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

$$\begin{aligned} &\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n} \\ &\leq 3^q \|(\mathbf{R}^* - \mathbf{A}^*)^{-1}\|^q \Bigg[\left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{N} D_1 g_{t,h}(\boldsymbol{\theta}^*) \varepsilon_{t,h}(\boldsymbol{\theta}^*) \right\|^q \\ &+ \left(\|\sqrt{n}(\tilde{\mathbf{R}}_n^* - \mathbf{R}^*)\| + \|\sqrt{n}(\tilde{\mathbf{A}}_n^* - \mathbf{A}^*)\| \right)^q \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|^q \mathbf{1}_{H_n} \\ &+ \left\| (2n)^{-1} \int_0^1 D_2 S_n(\boldsymbol{\theta}^* + r(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)) - D_2 S_n(\boldsymbol{\theta}^*) dr \right\|^q \|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n} \Bigg]. \end{aligned} \tag{E.8}$$

By (E.7), (E.8) and the hypotheses associated with (E1), (E4) and (E7), we obtain

$$\begin{aligned} &\mathrm{E}\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n} \leq O(1) + 3^q \|(\mathbf{R}^* - \mathbf{A}^*)^{-1}\|^q (2\bar{M}C^*\delta^*)^q \|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\|^q \mathbf{1}_{H_n} \\ &+ 3^q \|(\mathbf{R}^* - \mathbf{A}^*)^{-1}\|^q (C^*)^q (\delta^*)^{2q} \mathrm{E}(n^{q/2} \Lambda_n^q \mathbf{1}_{Q_n}) \end{aligned} \tag{E.9}$$

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 (E.4) and (E.6). Combining these with (E.9) gives

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

Moreover, it follows from (E.5), 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 (E.1). Expressing $D_1 S_n(\hat{\boldsymbol{\theta}}_n)$ as the sum of

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

and a remainder term using the mean value theorem for vector-valued functions, we can prove (E.2) in a fashion similar to the prof of (E.1). The details are omitted.

Remark 7. Being an extension of Theorem 2.2 of Chan and Ing (2011), Theorem 6 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 4 and 5. Note that in the special case of $q = 3$, where (E.1) and (E.2) correspond to condition (E5), (E.4) above is weaker than condition (E3). On the other hand, (E.5) is a strengthened version of (E.3), which ensures the consistency of $\hat{\boldsymbol{\theta}}_n$, and the role of (E.6) in the proof of Theorem 6 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, (E.5) and (E.6) 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 (18) on the regression function and its first-order derivative 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.

References

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

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, **44**, 62–91.

Brockwell, P. J. and Davis, R. A. (1987) *Time series: theory and methods.* (1st ed.), Springer.

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. *The Annals of Statistics*, **39**, 1526–1550.

Cleaskens, G., Croux, C., and Kerckhoven, J. V. (2007). Perdiction-focused model selection for autoregressive models. *Australian & New Zealand Journal of Statistics*, **49**, 359–379.

Davies, P.L. (2008). Approximating data (with discussion). *Journal of the Korean Statistical Society*, **37**, 191–240.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 849–911.

Findley, D. F. (1991). Counterexamples to parsimony and BIC. *Annals of the Institute of Statistical Mathematics*, **43**, 505–514.

Findley, D. F., and Wei, C. Z. (1993). Moment bounds for deriving time series CLT's and model selection procedures. *Statistica 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. *Journal of Multivariate Analysis*, **83**, 415–450.

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

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

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

Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Annals of Statistics*, **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. *Statistica Sinica*, **21**, 1473–1513.

Ing, C.-K., Taniguchi, M., Hsiao, W.-C. and Hsu, H.-L. (2016). Model selection for high-dimensional misspecified time series, *Technical Report*.

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

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

Inoue, A. and Kilian, L. (2006). On the selection of forecasting models. *Journal of 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. *The Annals of Statistics*, **15**, 958–975.

Liu, W. and Yang, Y. (2011). Parametric or nonparametric? a parametricness index for model selection. *The Annals of Statistics*, **39**, 2074–2102.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **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. *The Annals of Statistics*, **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. *The Annals of Statistics*, **6**, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica 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. *The Annals of Statistics*, **8**, 147–164.

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

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

Shumway, R. H., Azari, A. S. and Pawitan, Y. (1988). Modeling mortality uctuations in Los Angeles as functions of pollution and weather effects. *Environmental Research*, **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. *Journal of Econometrics*, **71**, 207–225.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, **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).

van Erven, T., Grunwald, 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. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**, 361–417.

Wei, C. Z. (1992). On predictive least squares principles. *The Annals of Statistics*, **20**, 1–42.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, **76**, 419–433.

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

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

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

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