Bridging AIC and BIC: A New Criterion for Autoregression

Jie Ding, Vahid Tarokh, and Yuhong Yang

***Abstract*—To address order selection for an autoregressive model fitted to time series data, we propose a new information criterion. It has the benefits of the two well-known model selection techniques, the Akaike information criterion and the Bayesian information criterion. When the data is generated from a finite order autoregression, the Bayesian information criterion is known to be consistent, and so is the new criterion. When the true order is infinity or suitably high with respect to the sample size, the Akaike information criterion is known to be efficient in the sense that its predictive performance is asymptotically equivalent to the best offered by the candidate models; in this case, the new criterion behaves in a similar manner. Different from the two classical criteria, the proposed criterion adaptively achieves either consistency or efficiency depending on the underlying true model. In practice where the observed time series is given without any prior information about the model specification, the proposed order selection criterion is more flexible and reliable compared with classical approaches. Numerical results are presented demonstrating the adaptivity of the proposed technique when applied to various datasets.**

***Index Terms*—Adaptivity; Akaike information criterion; Asymptotic efficiency; Bayesian information criterion; Bridge criterion; Consistency; Information criterion; Model selection; Parametricness index.**

I. Introduction

In a practical situation of the autoregressive model fitting, the order of the model is generally unknown. Many order selection methods have been proposed, and each follows a different philosophy. Anderson's multiple decision procedure [1] sequentially tests when the partial autocorrelations of the time series become zero. The final prediction error criterion proposed by Akaike [2] aims to minimize the one-step prediction error when the estimates are applied to another independently generated dataset. Bhansali and Downham [3] generalized the final prediction error criterion by replacing 2 with a parameter α in its formula, and proved that the asymptotic probability of choosing the correct order increases as α increases. The well-known Akaike Information Criterion, AIC [4], was derived by minimizing the Kullback-Leibler divergence between the true distribution and the estimate of a candidate model. Some variants of AIC, for example the modified Akaike information criterion that replaces the constant 2 by a different positive number, have also been considered [5]. Nevertheless, Akaike [6] argued that within a Bayesian framework, the original AIC is more appropriate than its variants for practical applications. Hurvich and Tsai [7] proposed the corrected AIC for the case where the sample size is small. Another popular method is the Bayesian information criterion, BIC, proposed by Schwarz [8] that aims at selecting a model that maximizes the posterior model probability. Hannan and Quinn [9] proposed a criterion, HQ, that replaces the $\log n$ term in BIC by $c \log\log n$ $(c > 1)$, where n is the sample size. They showed that this is the smallest penalty term that guarantees strong consistency of the selected order. The focused information criterion is another approach that takes into account the specific purpose of the statistical analysis, by estimating the risk quantity of interest for each candidate model [10], [11]. Other methods for autoregressive order selection include the criterion autoregressive transfer function method [12], the predictive minimum description length criterion [13], the predictive least-squares principle [14], [15], and the combined information criterion [5]. More references can be found in [16], [17]. Despite the rich literature on autoregressive models, the most common order selection criteria remain AIC and BIC.

In this paper, the specified model class for fitting is the set of autoregressions with orders $L = 1, \ldots, L_{\max}^{(n)}$ for some prescribed natural number $L_{\max}^{(n)}$. In relation to the true data generating process, the model class is referred to as *well-specified* (or parametric) if the data is generated from a finite order autoregression and the true order is no larger than $L_{\max}^{(n)}$, and *mis-specified* (or nonparametric) otherwise. It is well known that BIC is

This research was funded by the Defense Advanced Research Projects Agency (DARPA) under grant number W911NF-14-1-0508.

J. Ding and V. Tarokh are with the School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States. Y. Yang is with the School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, United States.

consistent in order selection in the well-specified setting. In other words, the probability of choosing the true order tends to one as the sample size tends to infinity. The Akaike information criterion is not consistent and has a non-vanishing probability when the sample size tends to infinity [18]. However, AIC is shown to be efficient in the mis-specified setting, while BIC is not [19]. Here we call an order selection procedure (asymptotically) efficient if its prediction performance (in terms of the squared difference between the prediction and its target conditional mean) is asymptotically equivalent to the best offered by the candidate autoregressive models. A rigorous definition of efficiency is given in Section IV-C. In other words, AIC typically produces less modeling error than BIC when the data is not generated from a finite order autoregressive process. Furthermore, asymptotic efficiency of AIC for order selection in terms of the same-realization predictions for infinite order autoregressive or integrated autoregressive processes has also been well established [20], [21]. We note that $L_{\max}^{(n)}$ is usually allowed to grow with n at an appropriate rate for two purposes. Firstly, if data is generated from a finite dimensional model, a fixed $L_{\max}^{(n)}$ (not depending on n) may impose an unnecessary upper bound to prevent the true model from being selected. Secondly, if data is generated from an infinite dimensional model, a growing $L_{\max}^{(n)}$ makes it possible to achieve the optimal bias-variance (or underfitting-overfitting) tradeoff for each n.

In real applications, one usually does not know whether or not the model class is well-specified. The task of adaptively achieving the better performance of AIC and BIC is theoretically intriguing and practically useful. There have been several efforts towards this direction. Yang [22] considered the possibility of sharing the strengths of AIC and BIC in the regression context. It has been shown under mild assumptions that any consistent model selection criterion behaves suboptimally for estimating the regression function in terms of the minimax rate of convergence [22]. In other words, the conflict between AIC and BIC in terms of achieving model selection consistency and minimax-rate optimality in estimating the regression function cannot be resolved. But this does not indicate that there exists no criterion achieving the pointwise asymptotic efficiency in both well-specified and mis-specified scenarios, because the minimaxity (uniformity over the linear coefficients) is intrinsically different from the (pointwise) efficiency. In the remarkable work of Ing [23], a hybrid selection procedure combining AIC and BIC-like criteria was proposed. Loosely speaking, if a BIC-like criterion selects the same model at sample sizes n^ℓ $(0 < \ell < 1)$ and n, then with high probability (for large n) the model class is well-specified and the true model has been converged to, and thus a BIC-like criterion is used; otherwise AIC is used. Under some conditions, the hybrid approach was proved to achieve the pointwise asymptotic efficiency in both well-specified and mis-specified scenarios. In estimating regression functions with independent observations, Yang [24] proposed a similar approach to adaptively achieve asymptotic efficiency for both parametric and nonparametric situations, by examining whether BIC selects the same model again and again at different sample sizes (instead of only two sample sizes used by [23]). Liu and Yang [25] proposed a method to adaptively choose between AIC and BIC based on a measure called parametricness index. In the context of sequential Bayesian model averaging, Erven et al. [26] and van der Pas and Grünwald [27] used a switching distribution to encourage early switch to a better model. Zhang and Yang [28] proposed cross-validation as a general solution to choosing between AIC and BIC. The proposed approach was shown to behave like the better one of AIC and BIC for both the AIC and BIC territories asymptotically, with a suitably chosen data splitting ratio.

In this paper, we introduce a new information criterion which we refer to as the bridge criterion (BC). As we shall explain in the paper, the philosophy behind the proposed criterion is fundamentally different with the classical information criteria and their hybrid approaches. The idea may well be applicable to a broad range of statistical models, but we focus on autoregressive models in this paper, and rigorously analyze its asymptotic performance. The bridge criterion is able to address the following two issues: First, given a realistic time series data, an analyst is usually unaware of whether the model class is well-specified or not; Second, even if the model class is known to be correct, the order (dimension) is not known, so that any prescribed finite candidate set suffers the risk of missing the true model. We show that BC achieves both consistency when the model class is well-specified and asymptotic efficiency when the model class is mis-specified under mild conditions. Recall that the penalty terms of AIC and BIC (and their variants) are proportional to L for an autoregressive model of order L. In contrast, a key element of BC is the expression $1 + 2^{-1} + \cdots + L^{-1}$ employed in its penalty term. As we shall see, the harmonic number is what "bridges" the features of AIC and BIC automatically. Another key element is to let $L_{\max}^{(n)}$ grow with sample size at an appropriate rate. We emphasize that for the well-specified case, once the true order is selected with prob-

ability close to one, the resulting predictive performance is also asymptotically optimal/efficient. From this angle, the criterion achieves the asymptotic efficiency for both well-specified and mis-specified cases.

The outline of this paper is given below. In Section II, we formulate the problem and introduce some notation. In Section III, we propose a simplified version of bridge criterion and give an intuitive interpretation of it. In Section IV, we establish the consistency and the asymptotic efficiency property. We formulate an adjusted bridge criterion that adaptively chooses the number of candidate models, in order to further improve the scope of applications. Furthermore, we propose a concept called parametricness index, in order to measure the extent to which the specified model class is adequate in explaining the observed data. Some extensions of the theoretical analysis are briefly discussed. Numerical results are given in Section V validating the performance of our approach (even under small data size). We conclude this paper in Section VI. Finally, the appendix consists of three parts. The first part reviews how we originally derived the simplified bridge criterion, by making a connection between sequential hypothesis tests and information criteria. It serves as another perspective alternative to what we present in the main body of the paper. The second part of the appendix includes technical lemmas and proofs to all our theoretical results. In the third part, we propose a formula to bound the overfitting probability of the bridge criterion under finite sample size.

II. Background

A. Problem formulation

Given observations $\{x_t : t = 1, \ldots, n\}$, we consider the following autoregressive model of order L ($L \in \mathbb{N}$)

$$x_t + \sum_{\ell=1}^{L} \psi_{L,\ell} x_{t-\ell} = \epsilon_t, \tag{1}$$

where $\psi_{L,\ell} \in \mathbb{R}$ ($\ell = 1, \ldots, L$), $\psi_{L,L} \neq 0$, the roots of the polynomial $z^L + \sum_{\ell=1}^{L} \psi_{L,\ell} z^{L-\ell}$ have modulus less than 1, and ε_t's are independent random noises with zero mean and variance σ^2. The autoregressive model is referred to as an AR(L) model, and $[\psi_{L,1}, \ldots, \psi_{L,L}]^{\mathrm{T}}$ is referred to as the stable autoregressive filter Ψ_L. Let L_0 denote the true order, which is considered to be finite for now. In other words, the data is generated in the way described by (1) with $L = L_0$. When L_0 is unknown, we assume that $\{1, \ldots, L_{\max}^{(n)}\}$ is the candidate set of orders. Due to the reasons mentioned in the introduction, we allow the largest order $L_{\max}^{(n)}$ to increase to infinity with n in order to reduce prediction errors. We define

$$N = n - L_{\max}^{(n)}. \tag{2}$$

The sample autocovariance vector and matrix are respectively $\hat{\gamma}_L = [\hat{\gamma}_{1,0}, \ldots, \hat{\gamma}_{L,0}]^{\mathrm{T}}, \hat{\Gamma}_L = [\hat{\gamma}_{i,j}]_{i,j=1}^{L}$, where $\hat{\gamma}_{i,j} = N^{-1} \sum_{t=L_{\max}^{(n)}+1}^{n} x_{t-i} x_{t-j}$ $(0 \leq i, j \leq L_{\max}^{(n)})$. The filter of the autoregressive model of order L can be estimated by

$$\hat{\Psi}_L = -\hat{\Gamma}_L^{-1} \hat{\gamma}_L, \tag{3}$$

which yields consistent estimates [29, Appendix 7.5]. The one-step prediction error is $\hat{e}_L = \sum_{t=L_{\max}^{(n)}+1}^{n} (x_t + \hat{\psi}_{L,1} x_{t-1} + \cdots + \hat{\psi}_{L,L} x_{t-L})^2 / N$. For convenience, we define $\hat{e}_0 = \hat{\gamma}_{0,0}$. The error of the AR($L$) model can be calculated by

$$\hat{e}_L = \hat{e}_0 - \hat{\gamma}_L^{\mathrm{T}} \hat{\Gamma}_L^{-1} \hat{\gamma}_L. \tag{4}$$

Let $\gamma_{i-j} = E\{x_{t-i} x_{t-j}\}$ $(i, j \in \mathbb{Z})$ be the autocovariances and $\Psi_L = [\psi_{L,1}, \ldots, \psi_{L,L}]^{\mathrm{T}}$ be the best linear predictor of order L. In other words, Ψ_L ($L \geq 1$) is the minimum of

$$e_L = \min_{\psi_{L,1}^*, \ldots, \psi_{L,L}^* \in \mathbb{R}} E\left\{ \left((x_t + \sum_{\ell=1}^{L} \psi_{L,\ell}^* x_{t-\ell} \right)^2 \right\}, \tag{5}$$

where the expectation is taken with respect to the stationary process $\{X_n\}$. In addition, we define $e_0 = \gamma_0$. The values of Ψ_L and e_L can be calculated from a set of equations similar to (3)–(4), by removing the hat symbol from parameters.

Given an observed time series, one critical problem is the identification of the (unknown) order of the autoregressive model fitted to the data. Suppose that we generate a time series to simulate an AR(L_0) process using (1). Clearly, $\hat{e}_1 \geq \cdots \geq \hat{e}_{L_0-1} \geq \hat{e}_{L_0} \geq \hat{e}_{L_0+1} \geq \cdots \geq \hat{e}_{L_{\max}}$. Because of (4) and the consistency of $\hat{\psi}_L$, generally $\hat{e}_L$ is large for $L < L_0$ and is small for $L \geq L_0$. If we plot $\hat{e}_L$ against L for $L = 1, \ldots, L_{\max}$, the curve is usually decreasing for $L < L_0$ and becomes almost flat for $L > L_0$. Intuitively, the order $\hat{L}$ may be selected such that $\hat{e}_{L-1}/\hat{e}_L$ becomes "less significant" than its predecessors for $L > \hat{L}$. For later convenience, we define the empirical and theoretical gain of goodness of fit using AR(L) over AR($L-1$), respectively, as

$$\hat{g}_L = \log\left(\frac{\hat{e}_{L-1}}{\hat{e}_L} \right), \quad g_L = \log\left(\frac{e_{L-1}}{e_L} \right). \tag{6}$$

For time series data that are not permutable, a penalized method is usually adopted which selects $\hat{L}$ by minimizing $\log \hat{e}_L$ plus some penalty term. AIC and BIC are the state-of-art for autoregressive order selection. They select $\hat{L}$ ($1 \leq \hat{L} \leq L_{\max}^{(n)}$) that respectively minimizes

the quantities $\text{AIC}(n,L) = \log \hat{e}_L + 2L/n$, $\text{BIC}(n,L) = \log \hat{e}_L + L\log(n)/n$. The above forms of AIC and BIC were originally derived by assuming Gaussian noises, but they have been widely adopted in a broader context. Note that AIC and BIC also serve as representatives of two types of criteria in terms of asymptotic performance [16], [17]. For example, it was shown that the final prediction error criterion behaves asymptotically similar to AIC [19], and the predictive least-squares principle (when all the past data is considered) behaves like BIC asymptotically [14], [15]. As we shall see, the watershed of AIC and BIC (and their variants) in terms of asymptotic performance is whether the penalty is a fixed well-chosen constant or goes to infinity as a function of n. A detailed discussion in the context of linear regressions can be found in [17]. As was mentioned in the introduction, we are interested in developing one criterion that simultaneously achieves the advantages of AIC and BIC.

B. Notation

We write $h_n = \Theta(g_n)$ if $c < h_n/g_n < 1/c$ for some positive constant c for all sufficiently large n, and $h_n = O(g_n)$ if $|h_n| < cg_n$ for some positive constant c for all sufficiently large n. If $\lim_{n\to\infty} f_n/g_n = 0$, we write $f = o_n(g)$, or for brevity, $f = o(g)$. Let $\lfloor x \rfloor$ denote the largest integer less than or equal to x. Let $\mathcal{N}(\mu,\sigma^2), \mathcal{B}(a,b), \chi^2_k$ respectively denote the normal distribution with density function $f(x) = \exp\{-(x-\mu)^2/(2\sigma^2)\}/(\sqrt{2\pi}\sigma)$, the Beta distribution with density function $f(x) = x^{a-1}(1-x)^{b-1}/B(a,b)$, where $B(\cdot,\cdot)$ is the beta function, and the chi-square distribution with k degrees of freedom.

The matrix norm $\|\cdot\|$ is defined by $\|M\| = \sup_{\|y\|_2=1}\|My\|_2$, where $\|\cdot\|_2$ denotes the Euclidean norm of a column vector. For a positive definite matrix A, the norm $\|\cdot\|_A$ is defined by $\|y\|_A = (y^{\mathrm{T}}Ay)^{1/2}$. If two vectors $y_1 = [y_{1,1},\ldots,y_{1,L_1}]^{\mathrm{T}}$ and $y_2 = [y_{2,1},\ldots,y_{2,L_2}]^{\mathrm{T}}$ are of different sizes, then we allow subtraction of those vectors by modifying the definition in the following way. Given y_1, y_2, define y_1', y_2' as vectors of size $L' = \max\{L_1,L_2\}$ by appending $\max\{L_1,L_2\} - \min\{L_1,L_2\}$ zeros to the tail of y_1 or y_2. We define subtraction of y_1, y_2 in this case as $y_1' - y_2'$. Similarly, if the size of a vector y is smaller than the dimension of a positive definite matrix A of size $k \times k$, $\|y\|_A$ is the same as $\|y'\|_A$ where y' is of size k by appending zeros to the tail of y. For any positive integers t and m ($t > m$), we sometimes use $x_{t:t-m}$ to represent $[x_t,\ldots,x_{t-m}]^{\mathrm{T}}$.

We are usually interested in the one-step prediction error if a mismatch filter, as defined below, is specified [2], [30]. Assume that the data is generated from a filter Ψ_{L_0} as in (1). The average one-step prediction error of using filter Λ_L minus that of using the true filter is referred to as mismatch error

$$E\big\{[x_t,\ldots,x_{t-L'+1}](\Psi_{L_0} - \Lambda_L)\big\}^2 = \|\Lambda_L - \Psi_{L_0}\|^2_{\Gamma_{L'}}. \tag{7}$$

where $L' = \max\{L_0, L\}$ and $\Gamma_{L'}$ is the $L' \times L'$ covariance matrix of the true autoregression, namely its (i,j)th element is γ_{i-j}.

III. Bridge Criterion

For easy interpretation, we start by proposing the simplified bridge criterion, and giving an intuitive explanation of it. We study its theoretical performance and propose the adjusted bridge criterion in the next section.

The estimated order $\hat{L}$ by bridge criterion is

$$\hat{L} = \operatorname*{arg\,min}_{1\le L\le L^{(n)}_{\max}} \text{BC}(n,L) \triangleq \log \hat{e}_L + \frac{2L^{(n)}_{\max}}{N}\sum_{k=1}^{L}\frac{1}{k}, \tag{8}$$

where $L^{(n)}_{\max}$ is the largest candidate order. $L^{(n)}_{\max}$ must be selected such that $\lim_{n\to\infty} L^{(n)}_{\max} = \infty$, and its rate of growth will be studied in Section IV. It is well known that $\sum_{k=1}^{L} 1/k = \log L + c_E + o(1)$ for large L, where c_E is the Euler-Mascheroni constant. Fig. 1 illustrates the penalty term given by (8) for different n and $L^{(n)}_{\max} = \lfloor n^{1/3} \rfloor$. Without loss of generality, we can shift the curves to be at the same position at $L = 1$.

Fig. 2 illustrates the penalty curves for the Akaike information criterion, the Bayesian information criterion, and the simplified bridge criterion, respectively denoted by

$$J_{\text{AIC}}(L) = \frac{2}{n}L, \quad J_{\text{BIC}}(L) = \frac{\log(n)}{n}L,$$
$$J_{\text{BC}}(L) = \frac{2L^{(n)}_{\max}}{n}\sum_{k=1}^{L}\frac{1}{k},$$

with $n = 1000$. Any of the above penalty curves can be written in the form of $\sum_{k=1}^{L} t_k$, and only the slopes t_k ($k = 1,\ldots,L_{\max}$) matter to the performance of order selection. For example, suppose that L_2 is selected instead of L_1 ($L_2 > L_1$) by some criterion. This implies that the gain of goodness of fit $\log \hat{e}_{L_1} - \log \hat{e}_{L_2}$ is greater than the sum of slopes $\sum_{k=L_1+1}^{L_2} t_k$. Thus, we have shifted the curves of the latter three criteria to be tangent to the bent curve of the bridge criterion in order to highlight their differences and connections. Here, two curves are referred to as tangent to each other if one is above the other and they intersect at one one point, the tangent point. The tangent points (marked by circles) of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2715599, IEEE Transactions on Information Theory

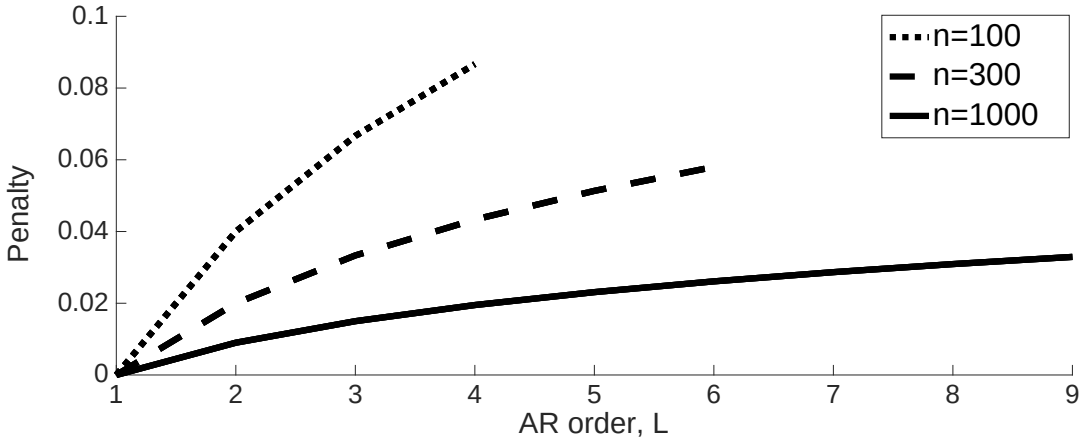


Fig. 1. A graph showing the penalty term in (8) for sample sizes 100 (small-dash), 300 (dash), and 1000 (solid).

J_{AIC} and J_{BIC} are respectively 9 and 2. Take the curve J_{BIC} as an example. The meaning of the tangent point is that BC penalizes more than BIC for $k \leq 2$ and otherwise for $k > 2$.

Given a sample size n, the tangent point between J_{BC} and J_{BIC} curves is at $T_{\text{BC:BIC}} = 2L_{\max}^{(n)}/\log n$. Consider the example $L_{\max}^{(n)} = \lfloor n^{1/3} \rfloor$. If the true order L_0 is finite, $T_{\text{BC:BIC}}$ will be larger than L_0 for all sufficiently large n. In other words, there will be an infinitely large region as n tends to infinity, namely $1 \leq L \leq T_{\text{BC:BIC}}$, where L_0 falls into and where BC penalizes more than BIC. As a result, asymptotically the bridge criterion does not overfit. On the other hand, the bridge criterion will not underfit because the largest penalty preventing from selecting $L+1$ versus L is $2L_{\max}^{(n)}/n$, which will be less than any fixed positive constant g_{L_0} and hence $\hat{g}_{L_0}$ (both defined in (6)) with high probability for large n. This reasoning suggests that the bridge criterion is consistent.

The inequality $(2L_{\max}^{(n)}/n)/k \geq 2/n$ for any $1 \leq k \leq L_{\max}^{(n)}$ guarantees that BC penalizes more than AIC. Since BC penalizes less for larger orders and finally becomes similar to AIC, it is able to share the asymptotic optimality of AIC under suitable conditions. To further illustrate why the bridge criterion is expected to work well in general, we make the following intuitive argument about the model selection procedure. As we shall see, the bent curve of BC well connects BIC and AIC so that a good balance between the underfitting and overfitting risks is achieved. The rigorous theory will be established in Section IV.

A heuristic understanding of the "bridge":

To gain further intuition, we consider an insect who is climbing a slope that is determined by a particular penalty curve $J(L)$ from the starting point $L = 1$ to the maximal possible end $L = L_{\max}^{(n)}$ (Fig. 3). Fig. 3(a) illustrates $J_{\text{AIC}}(L)$ (blue dash) and $J_{\text{BIC}}(L)$ (black small dash).

The climbing scheme and the goal: At each step L, the insect moves to step $L+1$ if its gain is larger than its loss, and it will not move any more once it stops. The gain refers to the increased goodness of fit to the data (which is $\hat{g}_{L+1}$ in our autoregressive model), the loss refers to the penalty of increased model complexity (which is $J(L+1) - J(L)$), and the last step where the insect stops is denoted by $\hat{L}$. The goal is to design a proper slope such that the insect stops at a "desired destination" that will be elaborated on below.

The tangent points of two slopes: A slope can be written as $\sum_{k=1}^{L} t_k$. The performance of the insect is determined by each increment t_k, and is not affected if the slope is shifted by any constant that does not depend on L. We thus shift the curves $J_{\text{AIC}}(L)$ and $J_{\text{BIC}}(L)$ to be tangent to the bent curve of $J_{\text{BC}}(L)$ if possible. By our design of $J_{\text{BC}}(L)$, the tangent points between $J_{\text{BC}}(L)$ and $J_{\text{AIC}}(L), J_{\text{BIC}}(L)$ curves are respectively at steps $T_{\text{BC:AIC}} = L_{\max}^{(n)}$, $T_{\text{BC:BIC}} = 2L_{\max}^{(n)}/\log n$. Before step $T_{\text{BC:BIC}}$, the insect on BC slope suffers more loss than on BIC slope in each move, while the other way around after step $T_{\text{BC:BIC}}$. Now we categorize two distinct scenarios.

The well-specified scenario: The first scenario is where the desired destination is within finitely many steps, and the second is where the desired destination is beyond finitely many steps. In the former case, there is a clear target step L_0. A good slope should be designed such that the insect stops at step L_0. It is already known in this case that BIC slope is good while AIC slope is not. In fact, it can be illustrated by Fig. 3(a), in which the gain after L_0 is $O_p(1)/n$, smaller than $\Theta(\log n)/n$ while larger than $O(1)/n$ with a positive probability for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2717599, IEEE Transactions on Information Theory

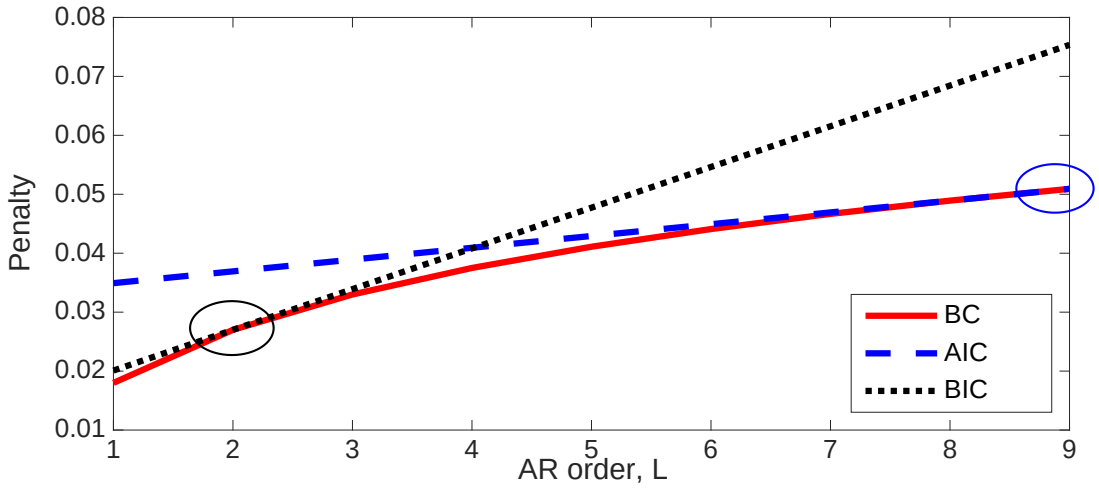


Fig. 2. A graph showing the penalty curves of the bridge criterion (red solid) together with the Akaike information criterion (blue dash), the Bayesian information criterion (black small dash), and the tangent points (circled) for $n = 1000$, $L_{\max}^{(n)} = \lfloor n^{1/3} \rfloor$.

sufficiently large n. How does BC compare? It is worth mentioning that our argument for the insect is implicitly built upon n, and the concept of consistency is about large n asymptotics. Suppose that n keeps increasing, the aforementioned tangent step $T_{\text{BC:BIC}}$ will be not only larger than L_0 but also diverging to infinity given that $\log n = o\big(L_{\max}^{(n)}\big)$. In other words, there is the "blackhole" region $[0, T_{\text{BC:BIC}}]$ (Fig. 3(b) and (c)), in which BC slope is steeper than BIC slope, and which grows to be infinitely large. It results in two consequences: First, the insect will find it more and more difficult to escape from the region because the increased loss from moving each step needs to be compensated by its gain. Take the autoregressive models as an example. After moving each step the gain is approximately independent χ_1^2/n, the expectation of which is less than the AIC penalty increment $2/n$. Therefore, the probability of the cumulated sum of gains being larger than that of loss decreases to zero rapidly as the number of steps increases. Second, once the insect is trapped in the blackhole, it encounters more difficulty to move forward on a BC slope than on a BIC one. Since on the BIC slope the insect will not move beyond step L_0 (due to the strong consistency of BIC), on a BC slope it will not, either.

On the other hand, the insect will not stop before step L_0. This occurs because the largest penalty preventing from moving forward is $J_{\text{BC}}(1) = o(1)$, but the gain of the insect moving from step L to $L+1$ when $L < L_0$ is usually at least $\Theta(1) + o_p(1)$ (which is true when $\psi_{L+1,L+1} \neq 0$ in autoregressive models). Therefore, the insect stops at step L_0 on a BC slope.

The mis-specified scenario: The fact that $T_{\text{BC:AIC}} = L_{\max}^{(n)}$ guarantees that BC slope is always steeper than AIC slope so that the insect does not move too far. Because the BC slope is in a concave shape, the insect moves easier and easier for larger steps. In the case where the appropriate destination tends to infinity, the insect will soon move to the tail part of the slope. As one can see from Fig. 3(c), in the tail part the slope is designed to be similar to AIC (and it becomes exactly AIC at the end step $L = L_{\max}^{(n)}$), it is possible to share the asymptotic optimality of AIC.

In summary, the bent curve of the BC well connects AIC and BIC so that a good balance between the underfitting and overfitting risks can be achieved. We emphasize that the above argument does not match exactly to the rigorous proof, since the decision making of the insect is carried out sequentially, while the aforementioned criteria select $\hat{L}$ via global optimum. Nevertheless, the argument for the insect does shed some light on why BC is likely to perform in the way we desire: to automatically behave like a consistent one while the underlying model is well-specified, and an efficient one otherwise, alleviating the risk caused by an analyst's initial prejudice. Besides this, the above argument does not assume any concrete probabilistic model, and thus it seems to be a promising criterion for other statistical inference tasks as well.

IV. Performance of The Bridge Criterion

In this section, we establish a rigorous theory on the asymptotic performance of the proposal in (8) and its adjusted version. We start with basic assumptions in Subsection IV-A, and prove the consistency and asymptotic efficiency of the simplified bridge criterion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2717599, IEEE Transactions on Information Theory

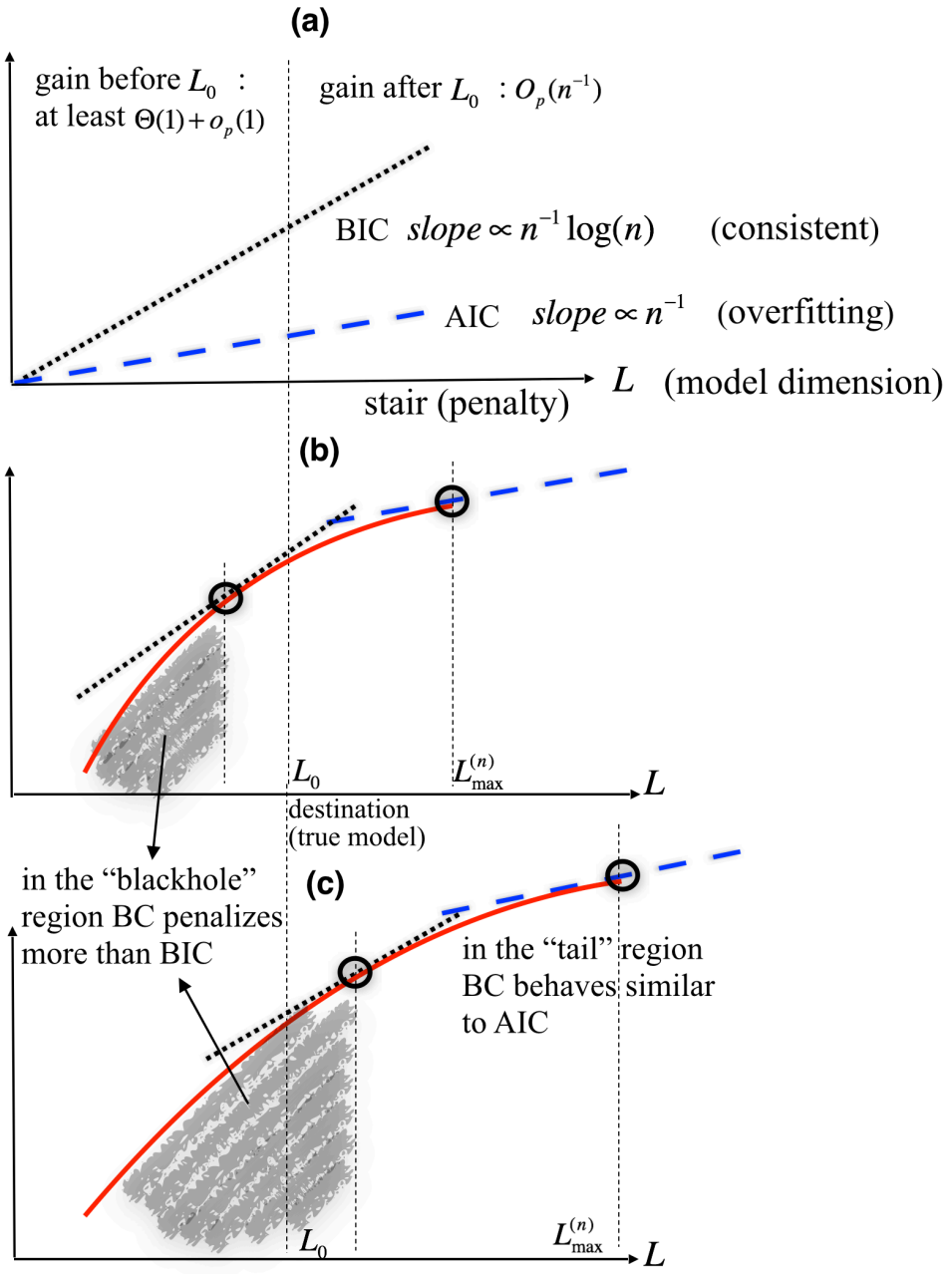


Fig. 3. (a) Curve J_{AIC} (blue dash) and J_{BIC} (black small dash), (b) the joint plot of J_{BC} (red thick line) and $J_{\text{AIC}}, J_{\text{BIC}}$, by shifting the latter two to be tangent to J_{BC} at tangent points $T_{\text{BC:AIC}}$, $T_{\text{BC:BIC}}$ (circled), in which $T_{\text{BC:AIC}} < L_0$, and (c) the evolution of plot (b) to the scenario $T_{\text{BC:BIC}} \geq L_0$ as n increases

in Subsections IV-B and IV-C, respectively. In Subsection IV-D, we propose an (adjusted) bridge criterion and its associated two-step strategy, in order to further improve the scope of applications. In Subsection IV-E, we propose a concept which we refer to as the parametricness index, in order to assess the confidence that the selected model can be practically treated as the data-generating model. In Subsection IV-F, we briefly discuss the extension of our established results to non-Gaussian and non-identically distributed noises. In view of the above intuitive argument, the adjusted criterion works in the following way.

Let the insect clime on the AIC slope, and record its ending point $\hat{L}_{\text{AIC}}$; modify the BC increment $J_{\text{BC}}(L) - J_{\text{BC}}(L-1)$ from $2L_{max}^{(n)}/(nL)$ to $2M_n/(nL)$, where M_n is slightly smaller than $L_{max}^{(n)}$; let the insect move again on the BC slope with boundary $\hat{L}_{\text{AIC}}$. In this way, the insect can still stop at L_0 if it is finite, and otherwise moves faster towards the end $\hat{L}_{\text{AIC}}$ as if it were on the AIC slope.

A. Notation and assumptions

We will need the following assumptions. Alternative assumptions that do not require Gaussian noises will be discussed in Subsection IV-F.

Assumption 1: $\{x_t : t = 1, \ldots, n\}$ is a stationary Gaussian process that satisfies $x_t + \psi_{\infty,1}x_{t-1} + \psi_{\infty,2}x_{t-2} + \cdots = \varepsilon_t$, where $\psi_{\infty,t} \in \mathbb{R}$, $\sum_{t=1}^{\infty} |\psi_{\infty,t}| < \infty$, ε_t's are independent and identically distributed according to $\mathcal{N}(0, \sigma^2)$, and the associated power series $\Psi(z) = 1 + \psi_{\infty,1}z^{-1} + \psi_{\infty,2}z^{-2} + \cdots$ converges and is not zero for $|z| \geq 1$.

Assumption 2: $\{L_{\max}^{(n)}\}$ is a sequence of positive integers such that $L_{\max}^{(n)} \to \infty$ and $L_{\max}^{(n)} = o(n^{1/2})$ as n tends to infinity.

Remark 1: Under Assumption 1, we have

$$0 < \gamma_0 = \|\Gamma_1\| \leq \|\Gamma_2\| \leq \cdots \leq \|\Gamma\| < \infty, \quad (9)$$

where $\Gamma = [\gamma_{i-j}]_{i,j=1}^{\infty}$ is the infinite dimensional covariance matrix with norm

$$\|\Gamma\| = \sup_{\|y\|_2=1} \left[\sum_{i=1}^{\infty} \left\{ \sum_{j=1}^{\infty} \gamma_{i-j} y_j \right\}^2 \right]^{1/2}.$$

Here, for a sequence $y_1, y_2, \ldots$, $\|y\|_2 = 1$ means $y_1^2 + y_2^2 + \cdots = 1$.

There can be either finitely or infinitely many nonzero elements in $\{\psi_{\infty,1}, \psi_{\infty,2}, \ldots\}$. Both cases are addressed in the following two sections.

B. Consistency

In this section, we show that the proposed order selection criterion is consistent in the well-specified case (see the assumption below).

Assumption 3: There are only finitely many k such that $\psi_{\infty,k} \neq 0$.

Under Assumption 3, we also say that the order of the autoregressive process is finite, or it is well-specified.

Theorem 1: Suppose that Assumptions 1, 2, 3 hold, then the bridge criterion is consistent.

Moreover, if $\hat{L}$ is selected from any finite set of integers that does not depend on n and that contains the true order L_0, and

$$\lim_{n\to\infty} \frac{L_{\max}^{(n)}}{\log\log n} = \infty, \quad (10)$$

then $\hat{L}$ converges not only in probability but also almost surely to L_0.

Remark 2: Theorem 1 proves the consistency of bridge criterion for a wide range of $L_{\max}^{(n)}$. In the intuitive arguments in the previous section, we let $L_{\max}^{(n)}/\log n \to \infty$ as $n \to \infty$. But as our proof indicates, it is not necessary for proving consistency. For strong consistency, however, $L_{\max}^{(n)}/\log\log n \to \infty$ is needed to apply the law of the iterated logarithm. Note that the upper limit of $L_{\max}^{(n)}$, $n^{1/2}$, is a common bound to prevent excessive estimation variances [19], [20]. The proof of Theorem 1 is given in the appendix.

C. Asymptotic efficiency

In this section, we show that the proposed order selection criterion can asymptotically minimize the mismatch error in the mis-specified case (see the assumption below).

Assumption 4: There are infinitely many k such that $\psi_{\infty,k} \neq 0$.

Under Assumption 4, we also say that the order of the autoregressive process is infinite, or it is mis-specified.

Remark 3: Assumption 4 has been assumed in some technical lemmas in [19] that we are going to introduce. For those lemmas and the scope of this paper, Assumption 4 can be generalized to allow for the case where the order of the autoregressive process, denoted by $L_0(n)$, is finite but increases to infinity as n tends to infinity. Thus, the data generating process varies with n. In that case, our technical proofs (in the appendix) are still applicable as long as we assume the divergence of L_n^* (introduced below) as n tends to infinity.

We define the cost function $C_n(L) = L\sigma^2/N + \|\Psi_L - \Psi_\infty\|_\Gamma^2$ (where N has been defined in (2)). It can be regarded as the expected mismatch error if an estimated filter of order L is used for prediction. In fact, under Assumptions 1, 2, 4, it holds that [19, Proposition 3.2]

$$\lim_{n\to\infty} \max_{1\leq L\leq L_{\max}^{(n)}} \left| \frac{\|\hat{\Psi}_L - \Psi_\infty\|_\Gamma^2}{C_n(L)} - 1 \right| = 0 \quad \text{in probability.} \quad (11)$$

Moreover, if we use $\{L_n^*\}$ to denote a sequence of positive integers that achieves the minimum of $C_n(L)$ for each n, namely $L_n^* = \arg\min_{1\leq L\leq L_{\max}^{(n)}} C_n(L)$, then for any random variable $\tilde{L}$ possibly depending on $\{x_t : t = 1, \ldots, n\}$, and for any $\epsilon > 0$, it holds that $\lim_{n\to\infty} \mathrm{pr}\{\|\hat{\Psi}_{\tilde{L}} - \Psi_\infty\|_\Gamma^2 / C_n(L_n^*) \geq 1 - \epsilon\} = 1$ [19, Theorem 3.2]. The result shows that the cost of the estimate $\hat{\Psi}(\tilde{L})$ is no less than $C_n(L_n^*)$ in probability for any order selection $\tilde{L}$. An order selection $\tilde{L}$ is called asymptotically efficient if

$$\lim_{n\to\infty} \frac{\|\hat{\Psi}_{\tilde{L}} - \Psi_\infty\|_\Gamma^2}{C_n(L_n^*)} = 1 \quad \text{in probability.} \quad (12)$$

Equality (12) can be equivalently written as $\lim_{n\to\infty} C_n(\tilde{L})/C_n(L_n^*) = 1$ in probability in view

of Equality (11). The following result establishes the asymptotic efficiency of bridge criterion in two common scenarios, where the mismatch error $\|\Psi_L - \Psi_\infty\|_\Gamma^2$ decays algebraically or exponentially in L. The two cases cover a wide range of linear processes as we point out in Remark 4. Its proof is given in the appendix.

Proposition 1: Suppose that Assumptions 1, 2, 4 hold.

1) Suppose that the mismatch error $\|\Psi_L - \Psi_\infty\|_\Gamma^2$ satisfies

$$\log\|\Psi_L - \Psi_\infty\|_\Gamma^2 = -\gamma \log L + \log c_L \qquad (13)$$

where $\gamma \geq 1$ is a constant, and the series $\{c_L : L = 1, 2, \ldots\}$ is lower bounded by a positive constant and $c_{L+1}/c_L < 1 + \gamma/(L+1)$. If

$$L_{\max}^{(n)} = O\left(n^{\frac{1}{1+\gamma}-\varepsilon}\right) \qquad (14)$$

holds for a fixed constant $0 < \varepsilon < 1/(1+\gamma)$, then the bridge order selection criterion is asymptotically efficient.

2) Suppose that the mismatch error satisfies the equality

$$\log\|\Psi_L - \Psi_\infty\|_\Gamma^2 = -\gamma L + \log c_L \qquad (15)$$

where $\gamma > 0$ is a constant, and the series $\{c_L : L = 1, 2, \ldots\}$ is lower bounded by a positive constant and $c_{L+1}/c_L \leq q$ for some constant $q < \exp(\gamma)$. If

$$L_{\max}^{(n)} \leq \frac{1-\varepsilon}{\gamma} \log n \qquad (16)$$

holds for a fixed constant $0 < \varepsilon < 1$, then the bridge order selection criterion is asymptotically efficient.

Remark 4: To provide an intuition of condition (13), in view of Remark 3 we prove that if the order of autoregressive process is not infinity but $L_0(n)$ (which grows with n) instead, and if $\Psi_{L_0(n)}$ is uniformly distributed in the space of all stable filters of order $L_0(n)$ (for any given n), then for large L ($1 \leq L \leq L_0(n)$)

$$\begin{aligned} &E\big\{\log\|\Psi_L - \Psi_{L_0(n)}\|_{\Gamma_{L_0(n)}}^2\big\} \\ &= -\log L + \log L_0(n) + o(1) \end{aligned} \qquad (17)$$

as L tends to infinity. More discussions on the above uniform distribution and the proof of (17) are given in the appendix.

Furthermore, it is known that condition (15) holds (with constant series c_L) when the data is generated from a finite order moving-average process [19].

We note that the proposed bridge criterion in (8) is not fully satisfactory in terms of asymptotic efficiency. For BC to achieve efficiency, our Proposition 1 requires $L_{\max}^{(n)}$ to satisfy (14) or (16) depending on the underlying mismatch error. This poses two concerns: first, the mismatch error as a function of L is usually unknown in advance, and it can be more complex than those characterized by (13) and (15); second, the chosen $L_{\max}^{(n)}$ is not large enough to incorporate all possible competitive models into the candidate set; this is because $L_{\max}^{(n)}$ is ε-away (in terms of the order) to the minimum of $C_n(L)$ over all positive integers $L \in \mathbb{N}$ (though for an arbitrarily small ε). This has motivated us to adjust the bridge criterion in such a way that 1) it relaxes the conditions required by (13) and (15), and 2) it selects the optimal order from a broad candidate set, and 3) it still achieves either consistency in well-specified cases or efficiency in mis-specified cases.

D. Adaptive selection of $L_{max}^{(n)}$

To achieve the aforementioned goal, we propose a general strategy that consists of two steps.

1. choose any $L_{\max}^{(n)} = o(\sqrt{n})$ and apply AIC to obtain $\hat{L}_{\text{AIC}}$;
2. within the range $1, 2, \ldots, \hat{L}_{\text{AIC}}$, select the optimal order (denoted by $\hat{L}_{\text{BC}}$) by minimizing the BC penalty

$$\text{BC}(n, L) = \log \hat{e}_L + \frac{2M_n}{n} \sum_{k=1}^{L} \frac{1}{k} \qquad (18)$$

where $\{M_n\}$ is a deterministic sequence to be chosen.

We note that $M_n = L_{\max}^{(n)}$ was chosen in the previous sections, but it may not be the ideal choice in our two-stage approach, as we shall see later. An intuition of the above bridge criterion in view of the "tale of an insect" has been made at the beginning of this section. The range of admissible $L_{\max}^{(n)}$ is rather wide, as we pointed out in Remark 2.

We define $L_0^{(n)} = \arg\min_{L \in \mathbb{N}} C_n(L)$ to be the the truly optimal order without restrictions on L. In the rest of this section, we consider that the $L_{\max}^{(n)}$ is chosen large enough so that $L_0^{(n)}$ is included, i.e., $L_0^{(n)} \leq L_{\max}^{(n)}$.

Assumption 5: In the mis-specified scenario, it holds that $L_0^{(n)} \leq L_{\max}^{(n)}$. In addition, $C_n(L)$ is regular in the sense that if $\lim_{n\to\infty} C_n(L_n)/C_n(L_0^{(n)}) = 1$ holds for a sequence L_n, then $\lim_{n\to\infty} L_n/L_0^{(n)} = 1$.

Remark 5: The efficiency of AIC under mis-specified model implies that $\lim_{n\to\infty} C_n(\hat{L}_{\text{AIC}})/C_n(L_0^{(n)}) = 1$ in probability. Given Assumption 5, it further implies that

$$\lim_{n\to\infty} \frac{\hat{L}_{\text{AIC}}}{L_0^{(n)}} = 1 \quad \text{in probability.} \qquad (19)$$

Assumption 5 typically holds for familiar nonparametric cases. For example, we consider two common scenarios that were also described in Proposition 1:

the mismatch error has an algebraic decay $\|\Psi_L - \Psi_\infty\|_\Gamma^2 = cL^{-\gamma}$, or an exponential decay $\|\Psi_L - \Psi_\infty\|_\Gamma^2 = c\exp(-\gamma L)$. We let $q_n = L_n/L_0^{(n)}$. Via straightforward calculation, $\lim_{n\to\infty} C_n(L_n)/C_n(L_0^{(n)}) = 1$ can be rewritten as $\lim_{n\to\infty} q_n^{-\gamma}(1+\gamma q_n^{\gamma+1})/(1+\gamma) = 1$ in the case of algebraic decay, and it can be rewritten as $\lim_{n\to\infty} \exp\{-\gamma(L_n - L_0^{(n)})\}/(1+\gamma L_0^{(n)}) + \gamma L_n/(1+\gamma L_0^{(n)}) = 1$ in the case of exponential decay. In both cases, it follows that $\lim_{n\to\infty} q_n = 1$ in probability.

The following theorem establishes the consistency and efficiency of the two-stage strategy.

Theorem 2: Suppose that $\hat{L}_{\text{AIC}}$ is obtained from the first step of the two-step strategy, and Assumptions 1, 2 hold, $M_n \to \infty$ as $n \to \infty$. Suppose that under a mis-specified model class, Assumption 5 holds, and for all sufficiently large n

$$M_n \le \frac{qL_0^{(n)}}{\log L_0^{(n)}}, \tag{20}$$

where $0 < q < 1$ is some constant. Then the bridge criterion in (18) is consistent in the well-specified case and efficient in the mis-specified case.

Moreover, if in the well-specified case $\hat{L}$ is selected from a finite candidate set that does not depend on n and that contains the true order L_0, and $\lim_{n\to\infty} M_n/\log\log n = \infty$, then $\hat{L}$ converges almost surely to L_0.

Remark 6: Recall that if the bridge criterion is consistent in the well-specified case and efficient in the mis-specified case, then it is efficient in both cases. We note that restrictions on M_n are fairly weak. For instance, $L_0^{(n)}$ is respectively $\Theta(n^r)$ $(0 < r < 1)$ and $\Theta(\log n)$ in the two cases described in Proposition 1, so we may choose $M_n = (\log n)^\tau$ with any $0 < \tau < 1$.

Remark 7: We provide an intuitive reasoning here. In the well-specified scenario, $M_n \to \infty$ guarantees consistency due to Theorem 1. In the mis-specified scenario, (19) and (20) imply that $M_n < \hat{L}_{\text{AIC}}/\log\hat{L}_{\text{AIC}}$. Such M_n produces penalty increments $J_{\text{BC}}(L+1) - J_{\text{BC}}(L)$ that are lighter than AIC for large L (recall that the candidate set in the second step is $1, \ldots, \hat{L}_{\text{AIC}}$). In view of that, BC produces $\hat{L}$ that is close to the boundary $\hat{L}_{\text{AIC}}$.

Remark 8: Another form of the bridge criterion is written as

$$\text{BC}(n, L) = \frac{2M_n}{n}\sum_{k=1}^{L} k^{-\zeta} \tag{21}$$

where $\zeta > 0, \zeta \neq 1$. By a similar proof, it can be shown that Theorem 2 can be modified to the case $0 < \zeta < 1$ by requiring the following changes: replace $L_0^{(n)}/\log L_0^{(n)}$ with $(L_0^{(n)})^\zeta$ in (20), and require $q < 1 - \zeta$. Moreover, Theorem 2 can be modified to the case $\zeta > 1$ by replacing $L_0^{(n)}/\log L_0^{(n)}$ with $L_0^{(n)}/a(\zeta)$ in (20), where $a(\zeta) = \sum_{k=1}^{\infty} k^{-\zeta}$. As a possible future work, it would be interesting to compare the performance of $\zeta = 1$ and $\zeta \neq 1$.

E. Parametricness index

Building upon the proposed bridge criterion, we define the following parametricness index (PI):

$$\text{PI}_n = \begin{cases} \dfrac{|\hat{L}_{\text{BC}} - \hat{L}_{\text{AIC}}|}{|\hat{L}_{\text{BC}} - \hat{L}_{\text{AIC}}| + |\hat{L}_{\text{BC}} - \hat{L}_{\text{BIC}}|} & \text{if } \hat{L}_{\text{AIC}} \neq \hat{L}_{\text{BIC}} \\ 1 & \text{otherwise.} \end{cases} \tag{22}$$

Following the definition, $\text{PI}_n \in [0, 1]$. Intuitively, PI_n is close to one in the well-specified model class where $\hat{L}_{\text{BC}}, \hat{L}_{\text{BIC}}$ do not differ much, while close to zero in a mis-specified one where $\hat{L}_{\text{BC}}, \hat{L}_{\text{AIC}}$ are close and much larger than $\hat{L}_{\text{BIC}}$. The goal of PI is to measure the extent to which the specified model class is adequate in explaining the observed data, namely to assess the confidence that the selected model can be practically treated as the data-generating model. The larger PI_n, the more confidence. Similar concept has been introduced in [25] for the goal of estimating the regression function. The following proposition shows that PI_n converges in probability to one for the well-specified case. Though we cannot prove that PI_n converges in probability to zero for various mis-specified cases in general, for illustration purpose we prove for some typical mis-specified cases. Experiments on various synthetic data in Section V have shown that PI_n performs in the way we expected.

Proposition 2: Under the same conditions of Theorem 2, if the model class is well-specified, PI_n converges in probability to one as n goes to infinity; If the model class is mis-specified, and we further assume that $C_n(L) + (\log n - 2)L\sigma^2/N$ achieves its minimum at $L_*^{(n)}$ and $\lim_{n\to\infty} L_*^{(n)}/L_0^{(n)} = 0$, then PI_n converges in probability to zero as n goes to infinity. In particular, the above condition holds if the mismatch error satisfies $\|\Psi_L - \Psi_\infty\|_\Gamma^2 = cL^{-\gamma}$, where γ and c are positive constants.

F. Relaxation of the assumption on noises

In the aforementioned theoretical results, we have used Assumptions 1 and 2 that assume i.i.d. Gaussian noises. In this subsection, we discuss a set of alternative assumptions that does not require normality or identical distribution, but at the cost of a slightly stronger condition on the AR coefficients.

Assumption 1′: $\{x_t : t = 1, \ldots, n\}$ is a stationary process that satisfies $x_t + \psi_{\infty,1}x_{t-1} + \psi_{\infty,2}x_{t-2} + \cdots = \varepsilon_t$, where $\psi_{\infty,t} \in \mathbb{R}$, $\sum_{t=1}^{\infty} |t^{1/2}\psi_{\infty,t}| < \infty$, and the associated power series $\Psi(z) = 1 + \psi_{\infty,1}z^{-1} + \psi_{\infty,2}z^{-2} + \cdots$ converges and is not zero for $|z| \geq 1$. The noises ε_t's are independent with zero mean and variance σ^2, and satisfy

$$\sup_{-\infty<t<\infty} E|e_t|^8 < \infty. \tag{23}$$

Let F_t denote the cumulative distribution function of e_t. There exist constants $c_0, \delta_0 > 0$ and $a \in (0, 1]$ such that for all $|x - y| < \delta_0$,

$$\sup_{-\infty<t<\infty} |F_t(x) - F_t(y)| \leq c_0|x - y|^a. \tag{24}$$

Assumption 2′: $\{L_{\max}^{(n)}\}$ is a sequence of positive integers such that $c_1 \leq L_{\max}^{(n)}/n^{1/(2+\delta)} \leq c_2$ for some constants $c_1, c_2, \delta > 0$.

Remark 9: In Assumption 1, the condition on $\psi_{\infty,t}$ implies that x_t admits an infinite moving-average representation $x_t = \sum_{j=0}^{\infty} \phi_j e_{t-j}$, where $\phi_0 = 1$ and $\{\phi_j\}$ are absolutely summable. The stronger Assumption 1′ implies that $\{j^{1/2}\phi_j\}$ are absolutely summable. We refer to [20] for more discussions. The choices of $L_{\max}^{(n)}$ in Assumption 2′ is slightly more restrictive than before. However, we now require much weaker conditions on the noises. Sufficient conditions to guarantee (23) and (24) are: e_i's are identically distributed, the moment generating function exists and the density function is bounded.

It has been shown by Ing and Wei [20, Lemmas 3&5] that Assumptions 1′ and 2′ lead to counterparts of Lemma 3.4 and Proposition 3.2 in [19] under Assumptions 1 and 2, which we have used in the proof of previous results.

In the sequel, we provide a counterpart of Theorem 2 under the alternative set of assumptions. Other theorems and propositions can be extended accordingly, but we do not elaborate.

Theorem 3: Under the same assumptions of Theorem 2, except that Assumptions 1, 2 are replaced with Assumptions 1′, 2′, the bridge criterion in (18) is consistent in the well-specified case and efficient in the mis-specified case.

V. Numerical Results

In this section, we present experimental results to demonstrate the theoretical results and the advantages of bridge criterion on both synthetic and real-world datasets. Throughout the experiments, we use the two-step bridge criterion defined in (18), and we adopt

$$L_{\max}^{(n)} = \lfloor n^{1/3} \rfloor, \quad M_n = (\log n)^{0.9} \tag{25}$$

due to Theorem 2 and Remark 6, where n is the sample size.

A. Synthetic data experiment: consistency in finitely dimensional model

The purpose of this experiment is to show the consistency of BC and BIC. The performance of BC, AIC, and BIC in terms of order selection for the well-specified model class is summarized in Table I. In Table I, the data is simulated using autoregressive filters $\Psi_2 = [\alpha, \alpha^2]^{\mathrm{T}}$ for $\alpha = 0.3, -0.3, 0.8, -0.8$. For each α, the estimated orders are tabulated for 1000 independent realizations of AR(2) processes $x_t + \alpha x_{t-1} + \alpha^2 x_{t-2} = \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, 1)$. The experiment is repeated for different sample sizes $n = 100, 500, 1000, 10000$. As was expected, the performance of the bridge criterion lies in between AIC and BIC, and it is consistent when n tends to infinity. Moreover, the convergence for $\alpha = 0.3, -0.3$ is slightly slower compared with $\alpha = 0.8, -0.8$, because of their smaller signal to noise ratios.

B. Synthetic data experiment: efficiency in finitely and infinitely dimensional models

The purpose of this experiment is to show that the proposed criterion achieves the asymptotic efficiency for both the well-specified and the mis-specified cases. The performance of BC in terms of mismatch error is compared with those of AIC and BIC in Table II. Recall that the mismatch error defined in (7) is the expected one-step ahead prediction error minus the variance of noise, when an estimated filter is applied to an independent and identically generated dataset. We consider three different data generating processes below. In Table II, for each case and sample size $n = 100, 500, 1000, 10000$, the tabulated mismatch error produced by each criteria were the mean of 1000 repeated independent experiments. The mean parametricness index defined in Subsection IV-E (denoted by PI_n) in each case was also tabulated.

Case 1: The first case is AR(1) with $\Psi_1 = [0.9]$, namely $x_t + 0.9x_{t-1} = \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, 1)$. This is a well-specified model. As we can see, once the true order is selected with probability close to one, the resulting predictive performance is also asymptotically optimal.

Here, we briefly explain how to calculate the exact mismatch error in (7) for any estimated filter of size L that may or may not equal to L_0. If suffices to express the covariance matrix $\Gamma_{L'}$ or its elements $\gamma_0, \ldots, \gamma_{L'-1}$ in terms of the known Ψ_{L_0}, where $L' = \max\{L_0, L\}$.

		n =100			n =500			n =1000			n=10000		
α	$\hat{L}$	BC	AIC	BIC	BC	AIC	BIC	BC	AIC	BIC	BC	AIC	BIC
0.3	1	784	548	851	558	213	661	298	51	405	0	0	0
	2	**151**	**292**	**135**	**372**	**558**	**333**	**619**	**677**	**589**	**949**	**720**	**999**
	3	36	98	13	37	113	5	38	125	5	21	97	1
	> 3	29	62	1	33	116	1	45	147	1	30	183	0
-0.3	1	777	566	845	535	208	628	297	45	375	0	0	0
	2	**166**	**301**	**145**	**392**	**536**	**365**	**624**	**688**	**617**	**958**	**719**	**997**
	3	28	64	8	32	110	6	32	112	7	22	122	3
	> 3	29	69	2	41	146	1	47	155	1	20	159	0
0.8	1	0	0	0	0	0	0	0	0	0	0	0	0
	2	**823**	**749**	**957**	**891**	**734**	**988**	**906**	**715**	**992**	**944**	**726**	**998**
	3	102	148	36	44	125	11	41	118	8	24	102	2
	> 3	75	103	7	65	141	1	53	167	0	32	172	0
-0.8	1	0	0	0	0	0	0	0	0	0	0	0	0
	2	**860**	**783**	**968**	**876**	**738**	**980**	**878**	**709**	**994**	**949**	**703**	**999**
	3	82	127	29	54	112	18	55	133	5	23	115	1
	> 3	58	90	3	70	150	2	67	158	1	28	182	0

TABLE I: Selected orders for AR(2) processes, computed from 1000 realizations for each α and n (with $\hat{L} = 2$ line in bold)

We define the correlation vector and matrix by $\rho_{L_0} = [\gamma_1/\gamma_0, \ldots, \gamma_{L_0}/\gamma_0]^{\mathrm{T}}, P_{L_0} = \Gamma_{L_0}/\gamma_0$. By rewriting the Yule-Walker equation $P_{L_0}\Psi_{L_0} = -\rho_{L_0}$, we obtain $(I + \Phi_{L_0})\ \rho_{L_0} = -\Psi_{L_0}$ where

$$\Phi_{L_0} = \begin{bmatrix} \psi_{L_0,2} & \psi_{L_0,3} & \cdots & \psi_{L_0,L_0-1} & \psi_{L_0,L_0} & 0 \\ \psi_{L_0,3} & \psi_{L_0,4} & \cdots & \psi_{L_0,L_0} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \psi_{L_0,L_0} & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & 0 \\ \psi_{L_0,1} & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \psi_{L_0,L_0-2} & \psi_{L_0,L_0-3} & \cdots & \psi_{L_0,1} & 0 & 0 \\ \psi_{L_0,L_0-1} & \psi_{L_0,L_0-2} & \cdots & \psi_{L_0,2} & \psi_{L_0,1} & 0 \end{bmatrix}.$$

We thus obtain $\rho_{L_0} = -(I+\Phi_{L_0})^{-1}\Psi_{L_0}$, $\gamma_0 = \sigma^2/(1+\rho_{L_0}^{\mathrm{T}}\Psi_{L_0})$, and $\gamma_\ell = \gamma_0\rho_{L_0,\ell}$ $(\ell = 1, \ldots, L_0)$. Furthermore, for each $\ell > L_0$, γ_ℓ equals to $-\sum_{k=1}^{L_0}\Psi_{L_0,k}\gamma_{\ell-k}$.

Case 2: The second case is AR($L_0(n)$) with $L_0(n) = \lfloor n^{0.4} \rfloor$ and $\Psi_{L_0(n)} = [0.7^k]_{k=1}^{L_0(n)}$, namely $x_t + \psi_{L_0(n),1}x_{t-1} + \cdots + \psi_{L_0(n),L_0(n)}x_{t-L_0(n)} = \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$. This is the case where the true order is large in terms of sample size, and thus it can be treated as the infinite dimensional model (see Remark 3). Note that all the roots of each characteristic polynomial have modulus 0.7. For each sample size $n = 100, 500, 1000, 10000$, the true order that generated the autoregression is $6, 12, 15, 39$, respectively.

Case 3: The third case is the first order moving average process $x_t = \epsilon_t - 0.8\epsilon_{t-1}, \epsilon_t \sim \mathcal{N}(0,1)$. It is an autoregression with infinite order. The exact mismatch error of an estimated filter Λ_L could be calculated in the following way: $\|\Lambda_L - \Psi_\infty\|^2_{\Gamma_\infty} = E\{x_{t+1} + [x_t, \ldots, x_{t-L+1}]\Lambda_L\}^2\} - \sigma^2 = 1.64(1 + \|\Lambda_L\|_2^2) - 2 \cdot 0.8\ (\Lambda_{L,1} + \sum_{k=1}^{L-1}\Lambda_{L,k}\Lambda_{L,k+1}) - 1$, where we have used $E(x_t^2) = 1.64$, $E(x_t x_{t-1}) = -0.8$, and $E(x_t x_{t-k}) = 0$ for $k > 1$.

In summary, Table I and II show that BC achieves the performance that we had expected: it is consistent when the model class is well-specified, and its predictive performance is always close to the optimum of AIC and BIC in both well-specified and mis-specified cases. In practice when no prior knowledge about the model specification is available, the proposed method is more flexible and reliable than AIC and BIC in selecting the most appropriate dimension.

C. Real data experiment: the El Nino data from 1935 to 2015

As the largest climate pattern, El Nino serves as the most dominant factor of oceanic influence on climate. The NINO3 index, defined as the area averaged sea surface temperature from 5°S-5°N and 150°W-90°W, is calculated from HadISST1 within the range of January 1935 to May 2015 [31]. The monthly data with overall 965 points is shown in Fig. 4(a). The data seems to be highly dependent from its sample partial autocorrelations shown in Fig. 4(b).

Case	n =100				n =500			
	BC	AIC	BIC	PI_n	BC	AIC	BIC	PI_n
1	19.7 (1.13)	28.6 (1.28)	16.6 (1.01)	0.96 (0.0061)	2.9 (0.18)	5.7 (0.26)	2.4 (0.13)	0.97 (0.0050)
2	76.7 (1.24)	71.9 (1.08)	94.2 (1.33)	0.58 (0.016)	17.6 (0.25)	17.5 (0.24)	25.2 (0.33)	0.29 (0.014)
3	97.8 (1.28)	94.7 (1.12)	122.8 (1.55)	0.58 (0.016)	26.6 (0.27)	26.6 (0.27)	38.0 (0.41)	0.32 (0.015)

Case	n =1000				n =10000			
	BC	AIC	BIC	PI_n	BC	AIC	BIC	PI_n
1	1.6 (0.11)	3.4 (0.15)	1.3 (0.065)	0.98 (0.0047)	0.11 (0.012)	0.39 (0.020)	0.10 0.0049	0.99 (0.0033)
2	9.9 (0.13)	9.9 (0.13)	14.6 (0.18)	0.18 0.012	1.4 (0.019)	1.4 (0.019)	2.1 (0.025)	0.11 (0.0097)
3	14.6 (0.15)	14.6 (0.15)	22.1 (0.24)	0.21 (0.013)	2.02 (0.021)	2.02 (0.021)	3.19 (0.032)	0.032 (0.0056)

TABLE II: Mismatch errors (and their standard errors) of autoregressive models selected by BC, AIC, and BIC, along with the parametricness index, in three different cases (values except PI_n and its standard errors were rescaled by 10^3)

To evaluate the predictive power of BC, AIC, and BIC, ideally we would apply each estimated filter to independent and identically generated datasets as we have done in the synthetic data experiments. But it is not realistic to apply this cross-validation to a single real-world time series data. As an alternative, we adopt a prequential perspective [20], [32], and evaluate the criteria in terms of the one-step prediction errors conditioning only on the past data at each time. Specifically, we start from an initial time step, say $n_0 = 200$, and obtain an estimated AR filter $\hat{\psi}_L(\mathcal{C})$ from the first 200 points under each criterion $\mathcal{C}$. Upon the arrival of $(n = n_0 + 1)$th point, the one-step prediction error is revealed to be $\hat{e}_n(\mathcal{C}) = (x_n - [x_{n-1}, \ldots, x_{n-L}]\hat{\psi}_L)^2$. This procedure is repeated for $n = n_0 + 2, \ldots, 965$, each time the AR filter being estimated from the observed $n-1$ data points and the tuning parameters being $L_{\max}^{(n)} = \lfloor n^{1/3} \rfloor, M_n = (\log n)^{0.9}$ (note that the n in (25) is the available sample size). The cumulated average prediction error at each n is computed to be $\overline{e}_n(\mathcal{C}) = \sum_{t=n_0+1}^{n} \hat{e}_t(\mathcal{C})/(n - n_0)$. To highlight the differences of $\overline{e}_n(\mathcal{C})$ for $\mathcal{C}$ = BC, AIC, BIC, we have plotted the normalized curve $\overline{e}_n(\mathcal{C}) - \overline{e}_n(opt)$ in Fig. 4(c), where $\overline{e}_n(opt) = \min\{\overline{e}_n(\text{AIC}), \overline{e}_n(\text{BIC})\}$ for each $n = n_0 + 1, n_0 + 2, \ldots$. In order to show predictive power that may vary at different time epochs, We have also plotted in Fig. 4(d) the (normalized) average prediction errors over a moving window of size 100, namely $\overline{e_0}_n(\mathcal{C}) = \sum_{t=s+1}^{n} \hat{e}_t(\mathcal{C})/(n - s)$ where $s = \max\{n_0, n - 100\}$.

Moreover, in order to capture potential dynamics during different time epochs, we have also considered the estimation from a moving window of size n_0. Specifically, we start from the same initial time step $n_0 = 200$, and for each $n = n_0 + 1, n_0 + 2, \ldots$, the AR filters are estimated from only $n - n_0, \ldots, n - 1$ with $L_{\max}^{(n)} = \lfloor n_0^{1/3} \rfloor, M_n = (\log n_0)^{0.9}$ (note that the n in (25) was replaced by the available sample size n_0). Similarly, we computed the one-step prediction errors, the normalized cumulated average prediction errors (plotted in Fig. 4(e)), and the normalized windowed average prediction errors (plotted in Fig. 4(f)). Fig. 4(c)-(f) show that the performance of BC is close to AIC and outperforms BIC in general.

D. Real data experiment: the English temperature data from 1659 to 2014

In this experiment, we study the monthly English temperature data from 1659 to 2014 used by [33], which is perhaps the longest recorded environmental data in human history. We have pre-processed the raw data by subtracting each month by the average of that month over the 356 years. The de-seasoned data (with overall 4272 points) is plotted in Fig. 5(a). Its sample partial autocorrelations are shown in Fig. 5(b). In order to capture potential dynamics during such a long period, we adopt the prequential approach that was used to draw Fig. 4(f), and omit the counterpart of Fig. 4(c)(d)(e). Specifically, we started from $n_0 = 500$, and for each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2717599, IEEE Transactions on Information Theory

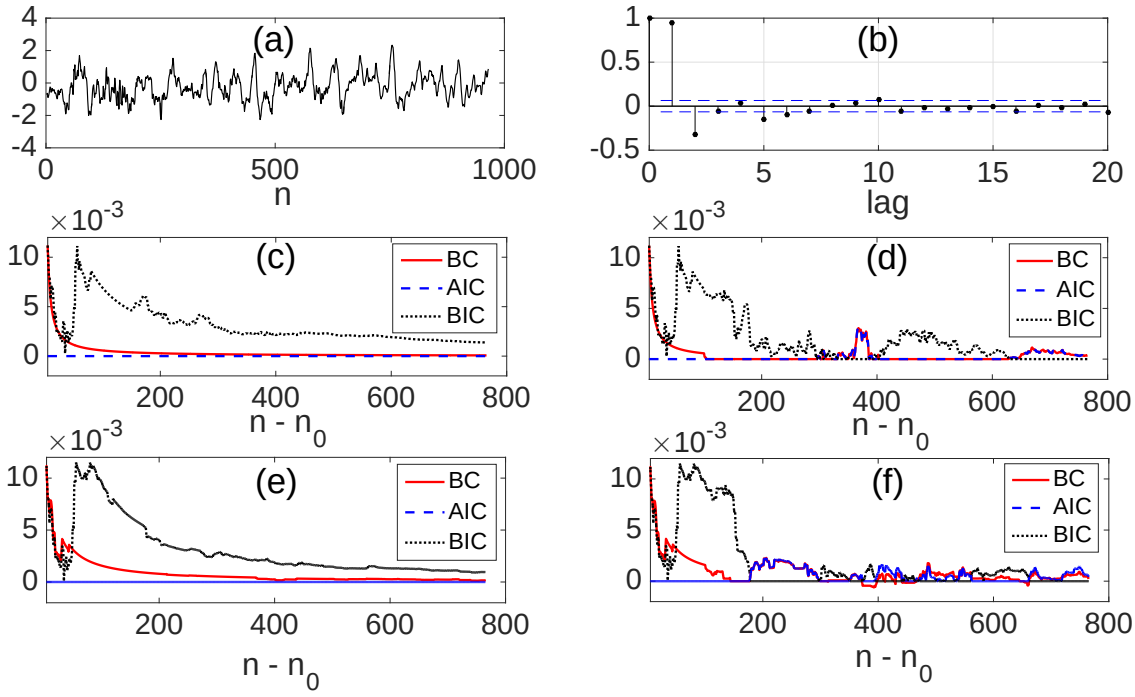


Fig. 4. (a) The monthly NINO3 index from January 1935 to May 2015; (b) the sample partial autocorrelations of the complete data with 95% confidence bounds; (c) the normalized cumulated average prediction error at each time step (using all the current observations); (d) the normalized average prediction error over the recent window of size 100 (using all the current observations); (e) the normalized cumulated average prediction error (using the recent n_0 observations); (f) the normalized average prediction error over the recent window of size 100 (using the recent n_0 observations). In subfigures (c)-(f), BC, AIC, and BIC are respectively marked in red, blue, and black, and the curves have been normalized by subtracting the minimum of AIC curve and BIC curve.

$n = n_0 + 1, \ldots, 4272$ the one-step ahead prediction was made by an AR filter produced from the recent window of n_0 observations. The prediction errors $\hat{e}_n$ were averaged over a fixed window of size 100, namely $\overline{e_0}_n(\mathcal{C}) = \sum_{t=s+1}^{n} \hat{e}_t(\mathcal{C})/(n-s)$ where $s = \max\{n_0, n-100\}$. We have plotted in Fig. 5(c) the normalized average prediction errors, which is $\overline{e_0}_n(\mathcal{C}) - \overline{e_0}_n(opt)$ where $\overline{e_0}_n(opt) = \min\{\overline{e_0}_n(\text{AIC}), \overline{e_0}_n(\text{BIC})\}$ (similar as before). We highlight the normalized average prediction errors within the range $n = n_0 + 500, \ldots, n_0 + 1500$ in Fig. 5(d). In this experiment, AIC is not constantly superior to BIC, and BC adaptively chooses to be close to the optimum of AIC and BIC. Furthermore, BC achieves the best predictive performance in some regions. The results show that BC is more flexible and reliable than AIC and BIC in practical applications. Note that we have adopted a specific choice of $L_{\max}^{(n)}$ and M_n (see (25)) throughout all the synthetic and real-world data experiments. In practice, an analyst may achieve much better predictive performance of BC, by fine tuning $L_{\max}^{(n)}$ and M_n for any particular real dataset.

VI. Concluding Remarks

There have been many debates on which of AIC and BIC should be used. A practitioner who supports AIC may argue that all models are wrong, and thus it is safe to choose AIC that generally performs better in mis-specified situations. In contrast, a practitioner who supports BIC is usually in favor of the mathematically appealing "consistency" property and is quite confident that the candidate set of models contains the true (or practically a very good) model, or simply has a strong preference of parsimony in modeling. However, the debate is aroused due to the underlying assumption that tends to be overlooked: a practitioner should choose either AIC or BIC before even looking at the observed data—if some model specification test were done, the practitioner might have changed his/her prejudice. In a certain sense, the bent curve of bridge criterion, different from straight lines, was designed to mimic a sequence of model specification test which continuously check "whether there exists a finite dimension L_0 underlying the observed data". For practical situations where there is no prior information, bridge criterion provides a prac-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2017.2717599, IEEE Transactions on Information Theory

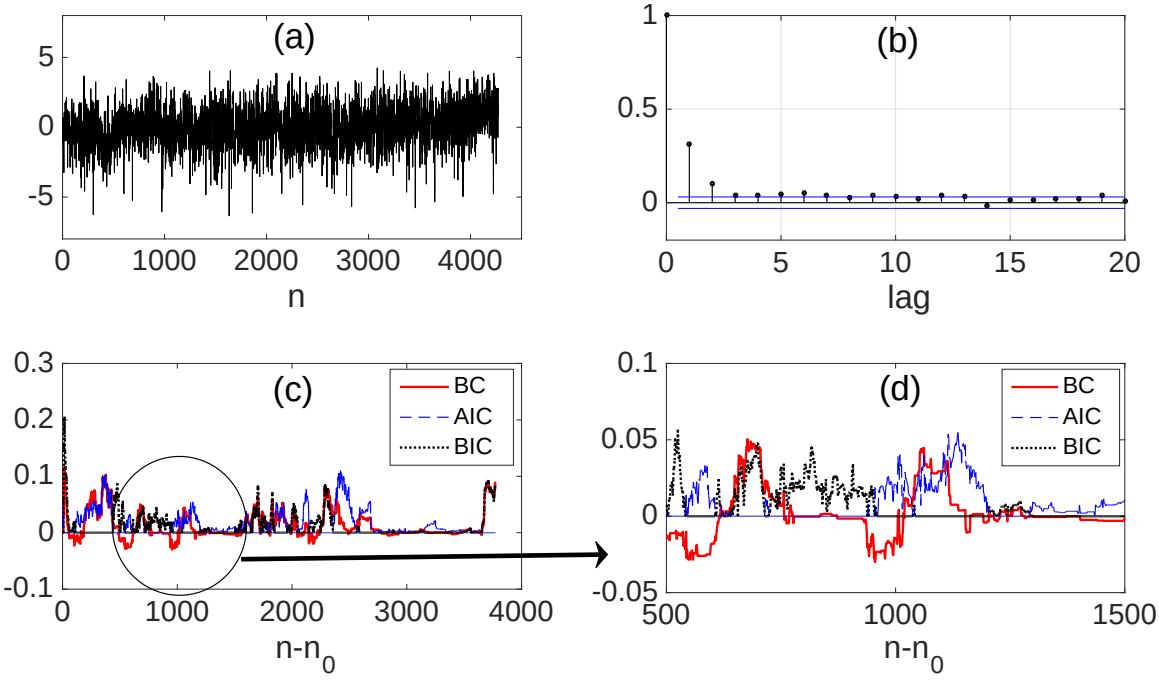


Fig. 5. (a) The de-seasoned data; (b) the sample partial autocorrelations of the complete data with 95% confidence bounds; (c) the normalized cumulated average prediction error at each time step (using the recent n_0 observations); (d) the normalized average prediction error over the recent window of size 100 (using the recent n_0 observations). In subfigures (c)-(d), BC, AIC, and BIC are respectively marked in red, blue, and black, and the curves have been normalized by subtracting the minimum of AIC curve and BIC curve.

titioner with opportunities to change or reinforce his/her belief in the model specification.

Based on the new criterion, we also proposed a parametricness index to measure the confidence that the selected model can be practically treated as the data-generating model. We established a rigorous theory for the application of the bridge criterion to autoregressive order selection. But the related ideas do not depend on specific model classes.

As a possible future work, it would be interesting to see in what extent the bridge criterion can be extended to other model selection problems, for instance the vector autoregressive model, autoregressive-moving-average model, autoregressive conditional heteroskedasticity model, and generalized linear model.

ACKNOWLEDGEMENT

The authors thank Peng Ding, Kathryn Heal, and Jiannan Lu for their suggestions in improving the presentation of this paper. The authors also thank Associate Editor Negar Kiyavash and two anonymous reviewers for their reviewing the paper and providing insightful comments.

APPENDIX

In the appendix, we begin by discussing an alternative perspective on where the harmonic "bridge" comes from. To that purpose, a new theorem and its proof will be given. Then, we provide some necessary technical lemmas, and prove Theorem 1, Proposition 1, Remark 4, Theorem 2, Proposition 2, and Theorem 3. Finally, we propose a formula to bound the overfitting probability of the bridge criterion under finite sample size.

APPENDIX A
A DERIVATION OF THE BRIDGE CRITERION: AN ALTERNATIVE PERSPECTIVE

In Section III and the beginning of Section IV, we made an intuitive tale about how the "bridge" works. In this section, we include a derivation of the bridge, which provides an interesting alternative perspective and which originally motivated us.

A. Motivation

Distinct from AIC or BIC, the new criterion was initially derived from some perspectives unique to autoregressions. The BC was initially motivated by postulating that nature randomly draws the coefficients of

true autoregressions from a non-informative uniform distribution and by fixing the type I error in a sequence of hypothesis tests on the order.

Recall our definitions of $\hat{g}_L$ and g_L in Section II. Suppose that the data is generated by a stable filter Ψ_{L_0} of order L_0. For any positive integer L that is greater than L_0 and does not depend on n, it was shown by [34, Theorem 5.6.2 and 5.6.3] that $\sqrt{n}[\hat{\psi}_{L_0+1,L_0+1}, \ldots, \hat{\psi}_{L,L}]^{\mathrm{T}}$ has a limiting joint-normal distribution $\mathcal{N}(0, I)$ as n tends to infinity, where I denotes the identity matrix. Moreover, the random variables $n\hat{g}_L$ $(L = L_0 + 1, \ldots, L_{\max})$ are asymptotically independent and distributed according to χ^2_1, where $L_{\max} > L_0$ is a constant that does not depend on n [18]. Next, we revisit AIC and BIC by associating them with a sequence of hypothesis tests. The purpose of the argument below is to motivate our new criterion.

Test: We choose a fixed number $0 < q < 1$ as the significance level (or the type I error), and thresholds s such that $q = \mathrm{pr}(W > s)$, where $W \sim \chi^2_1$. Consider the hypothesis test

$$H_0 : L_0 = L - 1 \quad H_1 : L_0 \geq L. \tag{26}$$

If $n\hat{g}_L > s$ (or equivalently $s/n - \hat{g}_L < 0$), we reject H_0 and replace $L - 1$ by L, for $L = 2, 3, \ldots$ until $L = L_{\max}$ or H_0 is not rejected. One limitation of this hypothesis test technique is that it may produce extreme values, as was pointed out in [30]. A straightforward alternative solution would be to select the L such that the aggregation of $s/n - \hat{g}_1, \ldots, s/n - \hat{g}_L$ is minimized, i.e., to select the global minimum:

$$\begin{aligned}\hat{L} &= \operatorname*{arg\,min}_{1\leq L\leq L_{\max}} \sum_{k=1}^{L} \left(\frac{s}{n} - \hat{g}_k\right) \\ &= \operatorname*{arg\,min}_{1\leq L\leq L_{\max}} \left(\log \hat{e}_L + \frac{sL}{n} - \log \hat{e}_0\right),\end{aligned} \tag{27}$$

the objective function of which can be regarded as the goodness of fit $\hat{e}_L$ plus the penalty of the model complexity. The penalty term is a sum of thresholds s and $-\log \hat{e}_0$. The term $-\log \hat{e}_0$ does not depend on L, so it has no effect on the produced result and is negligible. The Akaike information criterion has penalty term $2L/n$. It corresponds to the above hypothesis tests with q =0.1573 . The Bayesian information criterion has penalty term $L \log(n)/n$. It corresponds to the hypothesis tests with varying q. As an illustration, the significance levels q of BIC under different sample sizes are tabulated in Table III.

To motivate our new criterion, suppose that nature generates the data from an AR(L_0) process, which is in turn randomly generated from the uniform distribution $\mathcal{U}_{L_0}$. Here, $\mathcal{U}_L$ is defined over the space of all the stable

n	100	500	1000	2000	10000
q	0.0319	0.0127	0.0086	0.0058	0.0024

TABLE III: Significance level q of the Bayesian information criterion at different sample sizes

AR filters of order L whose roots have modulus no larger than r $(0 < r \leq 1)$:

$$\begin{aligned}S_L(r) = \Bigg\{ &\Psi_L : z^L + \sum_{\ell=1}^{L} \psi_{L,\ell} z^{L-\ell} = \prod_{\ell=1}^{L} (z - a_\ell), \\ &\psi_{L,\ell} \in \mathbb{R}, |a_\ell| \leq r,\, \ell = 1, \ldots, L \Bigg\}.\end{aligned} \tag{28}$$

Under this data generating procedure, g_L is a random variable with distribution described by the following theorem. For the sake of continuity, we postpone a detailed discussion on $\mathcal{U}_{L_0}$ to Subsection A-C.

Theorem 4: Suppose that Ψ_{L_0} is uniformly distributed in $S_{L_0}(1)$. Then, $\psi_{1,1}, \ldots, \psi_{L_0,L_0}$ are independently distributed according to $(\psi_{L,L}+1)/2 \sim \mathcal{B}(\lfloor L/2+1 \rfloor, \lfloor (L+1)/2 \rfloor)$ $(L = 1, \ldots, L_0)$. Furthermore, $L\psi^2_{L,L}$ and Lg_L converge in distribution to χ^2_1 as L tends to infinity.

Similarly, we postulate hypothesis tests in the opposite direction (for a given $L_{\max}$):

$$H_0 : L_0 = L \quad H_1 : L_0 \leq L - 1. \tag{29}$$

Under the null hypothesis, $g_L \neq 0$ almost surely, and we approximate the distribution of $\hat{g}_L$ by that of g_L. We choose a fixed number $0 < p < 1$ as the significance level, and the associated thresholds h_L at order L such that $p = \mathrm{pr}(g_L < h_L)$, or equivalently

$$h_L = F^{-1}_{g_L}(p) \tag{30}$$

where $F^{-1}_{g_L}(\cdot)$ denotes the inverse function of the cumulative distribution function of g_L. If $\hat{g}_L < h_L$ (or equivalently $\hat{g}_L - h_L < 0$), we reject H_0 and replace L by $L - 1$, for $L = L_{\max}, L_{\max} - 1, \ldots$ until $L = 2$ or H_0 is not rejected. Likewise, the L that minimizes the following objective function can be chosen as the optimal order

$$\begin{aligned}\hat{L} &= \operatorname*{arg\,min}_{1\leq L\leq L_{\max}} \sum_{k=L+1}^{L_{\max}} (\hat{g}_k - h_k) \\ &= \operatorname*{arg\,min}_{1\leq L\leq L_{\max}} \left(\log \hat{e}_L + \sum_{k=1}^{L} h_k + c\right)\end{aligned} \tag{31}$$

where $c = -(\log \hat{e}_{L_{\max}} + \sum_{k=1}^{L_{\max}} h_k)$ does not depend on L. The next subsection introduces the proposed criterion motivated by (31).

B. Proposed criterion

Building on the idea of (31), we adopt the penalty term $\sum_{k=1}^{L} h_k(p)$ where $h_k(p)$ is defined in (30), and p is further determined by

$$h_{L_{\max}}(p) = \frac{2}{n}. \tag{32}$$

Theorem 4 implies that $h_k(p) \approx F_{\chi_1^2}^{-1}(p)/k$ for large k, where $F_{\chi_1^2}^{-1}(\cdot)$ denotes the inverse function of the cumulative distribution function of χ_1^2. From (32) we have $F_{\chi_1^2}^{-1}(p) \approx 2L_{\max}/n$, and thus $h_k(p) \approx 2L_{\max}/(nk)$. We therefore propose the following criterion: select the $L \in \{1, \ldots, L_{\max}\}$ that minimizes $\log \hat{e}_L + (2L_{\max}/n)\sum_{k=1}^{L} 1/k$.

We have seen that given a fixed type I error, the threshold for hypothesis test (26) is a constant, while the threshold for (29) decreases in L leading to the $1/k$ term. Intuitively speaking, the uniform distribution on $S_L(r)$ concentrates more around the boundary of the space, and the loss of underfitting, $e_{L-1}/e_L = 1/(1-\psi_{L,L}^2)$, becomes more negligible, as L increases. To some extent, this observation suggests an interesting idea that the penalization for different models is not necessarily linear in model dimension; one may start with a BIC-type heavy penalty, but relax it more and more to an AIC-type light penalty as the candidate model grows, offering the possibility of changing/reinforcing one's belief in the model specification.

C. Uniform distribution of stable autoregressive filters

We have defined $S_L(r)$ in (28), the space of all the stable autoregressive filters whose roots have modulus no larger than r ($0 < r \leq 1$). Ding et al. [35] proposed a simple algorithm to generate a filter Ψ_L uniformly distributed on $S_L(r)$. The algorithm for $r = 1$ is given by the following pseudocode.

The formula is similar to the Levinson-Durbin recursion except in the way where $\psi_{L,L}$ is obtained. From Algorithm 1, the generation of a uniformly distributed filter of order L breaks down to the generation of L independent but not identically distributed random variables. This agrees with the fact that a stable filter can be uniquely identified by $\psi_{k,k}$ $(k = 1, \ldots, L)$ [36]. Theorem 4 is based upon Algorithm 1.

Proof of Theorem 4:

Proof: The distributions and independence of $\psi_{1,1}, \ldots, \psi_{L_0,L_0}$ follow from the procedure of Algorithm 1, and the continuously differentiable bijective mapping from $\psi_{1,1}, \ldots, \psi_{L_0,L_0}$ to $\psi_{L_0,1}, \ldots, \psi_{L_0,L_0}$ [37]. For any two positive integers a and b, if $Y \sim \mathcal{B}(a,b)$, it is well known that the distribution of Y is the same as the distribution of $(Z_1 + \cdots + Z_a)/(Z_1 + \cdots + Z_{a+b})$, where $Z_1, \ldots, Z_{a+b}$ are independent exponential random variables with density function $f_Z(z) = \exp(-z)$ $(z \geq 0)$. Suppose that $a/(a+b)$ tends to a constant $0 < c < 1$ as a and b tend to infinity. From the central limit theorem,

$$Y_a = \frac{b}{a+b}\frac{\sqrt{a}}{\sqrt{a+b}}\frac{Z_1 + \cdots + Z_a - a}{\sqrt{a}},$$
$$Y_b = -\frac{a}{a+b}\frac{\sqrt{b}}{\sqrt{a+b}}\frac{Z_{a+1} + \cdots + Z_{a+b} - b}{\sqrt{b}}$$

are asymptotically $\mathcal{N}(0, (1-c)^2 c)$ and $\mathcal{N}(0, c^2(1-c))$, respectively. Since Y_a and Y_b are independent, $Y_a + Y_b$ is asymptotically $\mathcal{N}(0, (1-c)c)$. Furthermore, from law of large numbers and Slutsky's theorem,

$$\begin{aligned}
&\sqrt{a+b}\left(\frac{Z_1 + \cdots + Z_a}{Z_1 + \cdots + Z_{a+b}} - \frac{a}{a+b}\right) \\
&= \sqrt{a+b}\left\{\frac{b(Z_1 + \cdots + Z_a) - a(Z_{a+1} + \cdots + Z_{a+b})}{(a+b)(Z_1 + \cdots + Z_{a+b})}\right\} \\
&= (Y_a + Y_b)\left(\frac{Z_1 + \cdots + Z_{a+b}}{a+b}\right)^{-1}
\end{aligned}$$

converges in distribution to $\mathcal{N}(0, (1-c)c)$. Since $(\psi_{L,L}+1)/2 \sim \mathcal{B}(\lfloor L/2+1 \rfloor, \lfloor (L+1)/2 \rfloor)$, we conclude that $\sqrt{L}\{(\psi_{L,L}+1)/2 - 1/2\}$ converges in distribution to $\mathcal{N}(0, 1/4)$, $\sqrt{L}\psi_{L,L}$ converges in distribution to $\mathcal{N}(0,1)$, and $L\psi_{L,L}^2$ converges in distribution to χ_1^2. Finally, it can be verified by the delta method and the identity $g_L = -\log(1-\psi_{L,L}^2)$ that Lg_L converges in distribution to χ_1^2. ■

APPENDIX B
TECHNICAL LEMMAS

Lemma 1: [19, Lemma 3.4] Let L be a positive integer and suppose that Assumption 1 holds. Then

$$\begin{aligned}
&E\left(N\left\|\frac{1}{N}\sum_{t=L_{\max}^{(n)}+1}^{n} x_{t-1:t-L}\,\epsilon_t\right\|_{\Gamma_L^{-1}}^2 - L\sigma^2\right)^4 \\
&= (48L + 12L^2)\sigma^8 + O(N^{-1})L^4\ .
\end{aligned}$$

Lemma 2: [19, Lemma 3.3] Under Assumptions 1 and 2,

$$\lim_{n\to\infty}\left(\max_{1\leq L\leq L_{\max}^{(n)}} \|\hat{\Gamma}_L^{-1} - \Gamma_L^{-1}\|\right) = 0,$$
$$\lim_{n\to\infty}\left(\max_{1\leq L\leq L_{\max}^{(n)}} \|\hat{\Gamma}_L - \Gamma_L\|\right) = 0 \quad \text{in probability.}$$

We will need the following lemmas for the proof of Theorem 1.

Lemma 3: Suppose that the model class is well-specified and that the true lag order is L_0. Under Assumptions 1 and 2, for any constant $\eta > 0$, there

Algorithm 1 Generate a uniform sample in $S_L(1)$

Randomly draw $\psi_{1,1}$ from the uniform distribution on $[-1,1]$
For $k=2$ to $k=L$
  Randomly draw β_k according to the beta distribution $\beta_k \sim \mathcal{B}(\lfloor k/2+1 \rfloor, \lfloor (k+1)/2 \rfloor)$
  Let $\psi_{k,k}=2\beta_k-1, \quad \psi_{k,\ell}=\psi_{k-1,\ell}+\psi_{k,k}\psi_{k-1,k-\ell} \quad (\ell=1,\ldots,k-1)$
Output $\psi_{L,1},\ldots,\psi_{L,L}$

exists a constant $c_\eta > 0$ (only depending on η), such that

$$\mathrm{pr}\left(\left|\frac{N}{L\sigma^2}\|\hat{\Psi}_L-\Psi_{L_0}\|_\Gamma^2-1\right|>\eta\right)<c_\eta L^{-2}$$

for all integer L satisfying $L_0 \le L \le L_{\max}^{(n)}$.

Proof: It follows from Lemma 1 and Markov's inequality that

$$\begin{aligned}&\mathrm{pr}\left(\left|\frac{N}{L\sigma^2}\left\|\frac{1}{N}\sum_{t=L_{\max}^{(n)}+1}^{n} x_{t-1:t-L}\ \epsilon_t\right\|_{\Gamma_L^{-1}}^2-1\right|>\eta\right)\\ &<48\eta^{-4}L^{-3}+12\eta^{-4}L^{-2}+O(N^{-1})\\ &\le c'_\eta L^{-2} \qquad (33)\end{aligned}$$

for some constant $c'_\eta > 0$, where (33) is due to $\hat{L} \le L_{\max}^{(n)}$ and $L_{\max}^{(n)}=o(n^{1/2})$. Since

$$\begin{aligned}\hat{\Psi}_L &= -(Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}}X_{n:L_{\max}^{(n)}+1}\\ &= -(Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}}(-Y\Psi_L+\epsilon_{n:L_{\max}^{(n)}+1})\\ &= \Psi_L-(Y^{\mathrm{T}}Y)^{-1}Y^{\mathrm{T}}\epsilon_{n:L_{\max}^{(n)}+1}\\ &= \Psi_L-\hat{\Gamma}_L^{-1}\ \frac{1}{N}\sum_{t=L_{\max}^{(n)}+1}^{n} x_{t-1:t-L}\ \epsilon_t\ ,\end{aligned}$$

where

$$Y=\begin{pmatrix} x_{n-1} & \cdots & x_{n-L}\\ \vdots & \ddots & \vdots\\ x_{L_{\max}^{(n)}} & \cdots & x_{L_{\max}^{(n)}-L+1}\end{pmatrix},$$

we have

$$\|\hat{\Psi}_L-\Psi_L\|_{\hat{\Gamma}_L}^2=\left\|\frac{1}{N}\sum_{t=L_{\max}^{(n)}+1}^{n} x_{t-1:t-L}\ \epsilon_t\right\|_{\hat{\Gamma}_L^{-1}}^2. \qquad (34)$$

Applying inequality (9), Lemma 2, inequality (33) and equality (34), we have the desired result. ∎

Lemma 4: For any positive integer L such that $L \ge L_0$, we have $\hat{e}_{L_0}-\hat{e}_L=\|\hat{\Psi}_L-\Psi_{L_0}\|_{\hat{\Gamma}_L}^2-\|\hat{\Psi}_{L_0}-\Psi_{L_0}\|_{\hat{\Gamma}_L}^2$.

Proof: It follows from Equations (3) and (4) that

$$\begin{aligned}\hat{e}_{L_0}-\hat{e}_L &= \|\hat{\gamma}_L\|_{\hat{\Gamma}_L^{-1}}^2-\|\hat{\gamma}_{L_0}\|_{\hat{\Gamma}_{L_0}^{-1}}^2=\|\hat{\Psi}_L\|_{\hat{\Gamma}_L}^2-\|\hat{\Psi}_{L_0}\|_{\hat{\Gamma}_{L_0}}^2\\ &= \|\hat{\Psi}_L-\Psi_{L_0}\|_{\hat{\Gamma}_L}^2-\|\hat{\Psi}_{L_0}-\Psi_{L_0}\|_{\hat{\Gamma}_L}^2\\ &\quad +2\Psi_{L_0}^{\mathrm{T}}\hat{\Gamma}_L(\hat{\Psi}_L-\hat{\Psi}_{L_0})\\ &= \|\hat{\Psi}_L-\Psi_{L_0}\|_{\hat{\Gamma}_L}^2-\|\hat{\Psi}_{L_0}-\Psi_{L_0}\|_{\hat{\Gamma}_L}^2\ .\end{aligned}$$

∎

We also need the following lemma for the proof of Proposition 1.

Lemma 5: [19, Theorem 4.2] Suppose that Assumptions 1, 2, 4 hold. Suppose that an order selection criterion is to select the $\hat{L}$ that minimizes the statistics $\{N+2L+\delta_n(L)\}\hat{e}_L$, where $\delta_n(L)$ is a real-valued random or non-random function of L, $1 \le L \le L_{\max}^{(n)}$. If

$$\lim_{n\to\infty}\max_{1\le L\le L_{\max}^{(n)}}\frac{|\delta_n(L)|}{N}=0 \quad \text{in probability,} \qquad (35)$$

$$\lim_{n\to\infty}\max_{1\le L\le L_{\max}^{(n)}}\frac{|\delta_n(L)-\delta_n(L_n^*)|}{NC_n(L)}=0 \quad \text{in probability,} \qquad (36)$$

then the selection $\hat{L}$ is asymptotically efficient.

APPENDIX C
PROOF OF THEOREM 1

Proof: We first prove that the probability of underfitting is zero when n tends to infinity. The probability of choosing $\hat{L}=L$ $(L<L_0)$ is bounded by

$$\begin{aligned}\mathrm{pr}(\hat{L}=L) &\le \mathrm{pr}\Bigg\{\log\hat{e}_L-\log\hat{e}_m<\frac{2L_{\max}^{(n)}}{n}\sum_{k=L+1}^{m}\frac{1}{k}\\ &\qquad (m=L+1,\cdots,L_0)\Bigg\}\\ &\le \mathrm{pr}\left(\log\hat{e}_L-\log\hat{e}_{L_0}<\frac{2L_{\max}^{(n)}}{n}\sum_{k=L+1}^{L_0}\frac{1}{k}\right)\\ &\le \mathrm{pr}\left(\log\hat{e}_{L_0-1}-\log\hat{e}_{L_0}<\frac{2L_{\max}^{(n)}L_0}{n}\right)\end{aligned}$$

Choose any $\varepsilon > 0$. Since $\psi_{L_0,L_0} \neq 0$, the consistency of $\hat{\Psi}_{L_0}$ and Assumption 2 imply that there exists a positive

number λ such that for all sufficiently large n

$$\mathrm{pr}\bigg(\log\hat{e}_{L_0-1}-\log\hat{e}_{L_0}>\lambda\bigg)=\mathrm{pr}\bigg(\log\frac{1+o_p(1)}{1-\hat{\psi}^2_{L_0,L_0}}>\lambda\bigg)>1-\varepsilon,$$

$$\text{and}\quad \frac{2L^{(n)}_{\max}L_0}{n}<\lambda.$$

It follows that $\lim_{n\to\infty}P\big(\hat{L}<L_0\big)=0$.

We next prove that the probability of overfitting tends to zero when n tends to infinity. It suffices to show that given any $\varepsilon>0$, $\mathrm{pr}(\hat{L}>L_0)<\varepsilon$ for all sufficiently large n. For any positive integer L' greater than L_0, the probability of choosing $\hat{L}\geq L'$ is bounded by

$$\begin{aligned}\mathrm{pr}(\hat{L}\geq L')&=\mathrm{pr}\bigg\{\bigcup_{L=L'}^{L^{(n)}_{\max}}(\hat{L}=L)\bigg\}\\&\leq\mathrm{pr}\bigg[\bigcup_{L=L'}^{L^{(n)}_{\max}}\bigg\{\log\hat{e}_m-\log\hat{e}_L>\frac{2L^{(n)}_{\max}}{n}\sum_{k=m+1}^{L}\frac{1}{k}\\&\qquad(m=L_0,\cdots,L-1)\bigg\}\bigg]\\&\leq\mathrm{pr}\bigg\{\bigcup_{L=L'}^{L^{(n)}_{\max}}\bigg(\log\hat{e}_{L_0}-\log\hat{e}_L>\frac{2L^{(n)}_{\max}}{n}\sum_{k=L_0+1}^{L}\frac{1}{k}\bigg)\bigg\}.\end{aligned}\tag{37}$$

Because of $1/k\geq 1/L^{(n)}_{\max}$, we obtain

$$\begin{aligned}\mathrm{pr}(\hat{L}\geq L')&\leq\mathrm{pr}\bigg\{\bigcup_{L=L'}^{L^{(n)}_{\max}}\bigg(\log\frac{\hat{e}_{L_0}}{\hat{e}_L}>\sum_{k=L_0+1}^{L}\frac{2}{n}\bigg)\bigg\}\\&\leq\sum_{L=L'}^{L^{(n)}_{\max}}\mathrm{pr}\bigg\{\log\frac{\hat{e}_{L_0}}{\hat{e}_L}>\frac{2(L-L_0)}{n}\bigg\}\\&\leq\sum_{L=L'}^{L^{(n)}_{\max}}\mathrm{pr}\bigg[\frac{\hat{e}_{L_0}-\hat{e}_L}{\hat{e}_{L_0}}>1-\exp\bigg\{-\frac{2(L-L_0)}{n}\bigg\}\\&\qquad=\frac{2(L-L_0)}{N}\{1+o(1)\}\bigg]\\&\leq\sum_{L=L'}^{L^{(n)}_{\max}}\mathrm{pr}\bigg\{\frac{N}{L-L_0}\frac{\hat{e}_{L_0}-\hat{e}_L}{\hat{e}_{L_0}}-1>1+o(1)\bigg\}.\end{aligned}\tag{38}$$

It follows from Lemmas 2, 3, and 4, and the consistency of $\hat{\Psi}_{L_0}$ and $\hat{e}_{L_0}$, that there exists a positive integer L_1 ($L_1>L_0$) such that with probability at least $1-\varepsilon/3$, for all L ($L_1<L<L^{(n)}_{\max}$), the event $\big\{N/(L-L_0)\cdot(\hat{e}_{L_0}-\hat{e}_L)/\hat{e}_{L_0}-1>1+o(1)\big\}$ is contained in the event $\big\{N/(L\sigma^2)\cdot\|\hat{\Psi}_L-\Psi_{L_0}\|^2_\Gamma-1>2^{-1/4}\big\}$, implying that

$$\begin{aligned}&\mathrm{pr}\bigg\{\frac{N}{L-L_0}\frac{\hat{e}_{L_0}-\hat{e}_L}{\hat{e}_{L_0}}-1>1+o(1)\bigg\}\\&\leq\mathrm{pr}\bigg(\bigg|\frac{N}{L\sigma^2}\|\hat{\Psi}_L-\Psi_{L_0}\|^2_\Gamma-1\bigg|>2^{-1/4}\bigg)<cL^{-2}\end{aligned}$$

for some positive constant c. Combining this result and Inequality (38), we obtain

$$\mathrm{pr}(\hat{L}\geq L_1)\leq\frac{\varepsilon}{3}+\sum_{L=L_1}^{L^{(n)}_{\max}}cL^{-2}\leq\frac{\varepsilon}{3}+c(L_1-1)^{-1}.$$

Furthermore, there exists a positive integer L_2 ($L_2\geq L_1$) such that

$$\mathrm{pr}(\hat{L}\geq L_2)\leq\frac{\varepsilon}{3}+\frac{\varepsilon}{3}=\frac{2\varepsilon}{3}\ .\tag{39}$$

It remains to prove that $\mathrm{pr}(\hat{L}<L_2)\leq\varepsilon/3$ for all sufficiently large n. Similar to (37), we have

$$\begin{aligned}&\mathrm{pr}(L_0<\hat{L}<L_2)\\&\leq\mathrm{pr}\bigg\{\bigcup_{L=L_0+1}^{L_2-1}\bigg(\log\hat{e}_{L_0}-\log\hat{e}_L>\frac{2L^{(n)}_{\max}}{n}\sum_{k=L_0+1}^{L}\frac{1}{k}\bigg)\bigg\}\\&\leq\mathrm{pr}\bigg\{\log\hat{e}_{L_0}-\log\hat{e}_{L_2-1}>\frac{2L^{(n)}_{\max}}{n}\frac{1}{L_0+1}\bigg\}.\end{aligned}\tag{40}$$

It has been proved in [18] that the random variables $n(\hat{e}_{L-1}-\hat{e}_L)$ ($L=L_0+1,\ldots,L_2-1$) are asymptotically independent and distributed according to χ^2_1 (for the L_2 that does not depend on n). Therefore, $\log\hat{e}_{L_0}-\log\hat{e}_{L_2-1}=O_p(n^{-1})$, and the value in (40) is less than $\varepsilon/3$ for all sufficiently larger n. In sum, we obtain $\mathrm{pr}(\hat{L}>L_0)\leq\varepsilon$ for all sufficiently larger n, and thus $\lim_{n\to\infty}\mathrm{pr}(\hat{L}>L_0)=0$.

Finally, we prove that $\hat{L}$ converges almost surely to L_0, when $\hat{L}$ is restricted to a finite candidate set that does not depend on n and that contains L_0. Without loss of generality, we suppose that the candidate set is $\{1,\ldots,L_{\max}\}$, where $L_{\max}>L_0$ is a constant integer. Let $c>0$ be any fixed constant. Condition (10) implies that there exists a positive integer n_1 such that for all $n>n_1$, $cL_{\max}<L^{(n)}_{\max}/(\log\log n)$. This means that for each L ($L_0<L<L_{\max}$) the penalty increment of BC is larger than that of HQ criterion, i.e.,

$$\frac{2L^{(n)}_{\max}}{n}\frac{1}{L}>\frac{2c\log\log n}{n}.$$

Therefore, the event $E_1=\{\hat{L}=L\ (L_0<L<L_2)\}$ implies the event $E_2=\{\log\hat{e}_{L-1}-\log\hat{e}_L>2L^{(n)}_{\max}/(nL)\}$, which further implies $E_3=\{\log\hat{e}_{L-1}-\log\hat{e}_L>(2c\log\log n)/n\}$. On the other hand, it can be proved by the law of the iterated logarithm that the event E_3

is eventually null with probability one (see for example [9]).

■

APPENDIX D
PROOF OF PROPOSITION 1

Proof:

We prove for the case where $c_L = c$ is a constant series. Similar proof can be applied to the general case.

Case 1) Recall that the bridge criterion is to select L $(1 \le L \le L_{\max}^{(n)})$ that minimizes

$$\log \hat{e}_L + \frac{2L_{\max}^{(n)}}{n} \sum_{k=1}^{L} \frac{1}{k}.$$

By adding a constant $\log N$ that does not depend on L, it is equivalent to minimizing

$$\log N + \log \hat{e}_L + \frac{2L_{\max}^{(n)}}{n} \sum_{k=1}^{L} \frac{1}{k},$$

which is further equivalent to minimizing $(N + 2L + \delta_n(L))\hat{e}_L \quad (1 \le L \le L_{\max}^{(n)})$ where

$$\delta_n(L) = N \exp\left[\frac{2L_{\max}^{(n)}}{n} \sum_{k=1}^{L} \frac{1}{k}\right] - (N + 2L). \tag{41}$$

Due to Lemma 5, it suffices to prove that $\delta_n(L)$ satisfies the conditions (35) and (36). Using Taylor series expansion, equality (41), and the fact that $\lim_{n\to\infty}(L_{\max}^{(n)} \log L_{\max}^{(n)})^2/n = 0$ holds for both (14) and (16), we obtain

$$\begin{aligned}\delta_n(L) &= N\left[1 + \frac{2L_{\max}^{(n)}}{n} \sum_{k=1}^{L} \frac{1}{k} + o(n^{-1})\right] - (N + 2L) \\ &= 2L_{\max}^{(n)} \frac{N}{n} \sum_{k=1}^{L} \frac{1}{k} - 2L + o(1) = o(N),\end{aligned} \tag{42}$$

which guarantees (35).

The denominator in (36) is

$$\begin{aligned}G(L) = NC_n(L) &= L\sigma^2 + N\|\Psi_L - \Psi_\infty\|_\Gamma^2 \\ &= L\sigma^2 + NcL^{-\gamma}.\end{aligned}$$

The derivative of the function $G(y)$ satisfies $dG(y)/dy < 0$ when $1 \le y \le \Theta(n^{1/(1+\gamma)-\varepsilon})$ for sufficiently large n. Therefore, (14) implies that $G(L) \ge G(L_{\max}^{(n)})$ and

$$\begin{aligned}\frac{|\delta_n(L) - \delta_n(L_n^*)|}{NC_n(L)} &\le \frac{(2L_{\max}^{(n)} \log L_{\max}^{(n)})\{1 + o(1)\}}{L_{\max}^{(n)}\sigma^2 + Nc(L_{\max}^{(n)})^{-\gamma}} \\ &= \Theta\left\{\frac{(2L_{\max}^{(n)} \log L_{\max}^{(n)})(L_{\max}^{(n)})^\gamma}{Nc}\right\}.\end{aligned}$$

Furthermore, the desired result (36) follows from

$$\lim_{n\to\infty} \frac{(2L_{\max}^{(n)} \log L_{\max}^{(n)})(L_{\max}^{(n)})^\gamma}{Nc} = 0.$$

Case 2) Using a similar reasoning we obtain

$$\begin{aligned}\frac{|\delta_n(L) - \delta_n(L_n^*)|}{NC_n(L)} &\le \frac{(2L_{\max}^{(n)} \log L_{\max}^{(n)})\{1 + o(1)\}}{L_{\max}^{(n)}\sigma^2 + Nc\exp(-\gamma L_{\max}^{(n)})} \\ &= \Theta\left(\frac{2L_{\max}^{(n)} \log L_{\max}^{(n)}}{n^\varepsilon}\right),\end{aligned}$$

which goes to zero as n tends to infinity.

■

APPENDIX E
PROOF OF REMARK 4

Proof: With a slight abuse of notation, we write $L_0(n)$ as L_0 for brevity. For $L < L_0$ we have

$$\begin{aligned}\log\|\Psi_L - \Psi_{L_0}\|_{\Gamma_{L_0}}^2 = \log e_L &= \sum_{k=L+1}^{L_0} g_k \\ &= -\sum_{k=L+1}^{L_0} \log(1 - \psi_{k,k}^2)\ .\end{aligned} \tag{43}$$

From Theorem 4,

$$(\psi_{k,k} + 1)/2 \sim \mathcal{B}(\lfloor k/2 + 1\rfloor, \lfloor (k+1)/2\rfloor).$$

Let $b_k = (\psi_{k,k} + 1)/2$. Straightforward calculation using the central moments of the beta distribution [38, Chapter 4] gives

$$\begin{aligned}E(\psi_{k,k}^2) = E\big(2b_k - 1\big)^2 &= 4E\left(b_k - \frac{1}{2}\right)^2 \\ &= \frac{1}{k} + \Theta_k\left(\frac{1}{k^2}\right),\end{aligned} \tag{44}$$

and

$$\begin{aligned}E\left\{\frac{\psi_{k,k}^4}{2(1-\psi_{k,k}^2)^2}\right\} &= \frac{B(\lfloor k/2 - 1\rfloor, \lfloor (k-3)/2\rfloor)}{2^5 B(\lfloor k/2 + 1\rfloor, \lfloor (k+1)/2\rfloor)} \\ &\quad \times E\big(2b_{k-2} - 1\big)^4 \\ &= \Theta_k\left(\frac{1}{k^2}\right),\end{aligned} \tag{45}$$

where $k > 4$ and $b_{k-2} \sim \mathcal{B}(\lfloor k/2 - 1\rfloor, \lfloor (k-3)/2\rfloor)$. For any $0 \le x < 1$, Taylor's theorem gives

$$\begin{aligned}x < -\log(1 - x) &= x + \frac{1}{2(1-\hat{x})^2}x^2 \\ &< x + \frac{x^2}{2(1-x)^2},\end{aligned} \tag{46}$$

where $0 < \hat{x} < x$. Combining (44), (45), and (46), we

have

$$E\{-\log(1-\psi_{k,k}^2)\} = \frac{1}{k} + \Theta_k\left(\frac{1}{k^2}\right). \quad (47)$$

Taking (47) into (43), we further obtain

$$\begin{aligned} E\big(\log\|\Psi_L - \Psi_{L_0}\|^2_{\Gamma_{L_0}}\big) &= \sum_{k=L+1}^{L_0}\left\{\frac{1}{k} + \Theta_k\left(\frac{1}{k^2}\right)\right\} \\ &= \sum_{k=1}^{L_0}\frac{1}{k} - \sum_{k=1}^{L}\frac{1}{k} + o(1) \\ &= \log L_0 - \log L + o(1). \end{aligned}$$

■

APPENDIX F
PROOF OF THEOREM 2

Proof of consistency:

By a similar argument as in the proof of Theorem 1, $\hat{L} \geq L_0$ with probability approaching one as n tends to infinity. Next, we prove that there is no overfitting. Suppose that $\varepsilon > 0$ is any fixed constant. It can be seen from the proof towards (38) and (39) that $\hat{L}_{\text{AIC}}$ is stochastically bounded. There exists a constant L_3 such that for all sufficiently large n, $\text{pr}(\hat{L}_{\text{AIC}} > L_3) < \varepsilon/6$. Similar to (38), we obtain for any fixed L' (to be chosen)

$$\begin{aligned} &\text{pr}(\hat{L} \geq L') \leq \text{pr}(\hat{L} \geq L', \hat{L}_{\text{AIC}} \leq L_3) + \text{pr}(\hat{L}_{\text{AIC}} > L_3) \\ &\leq \sum_{L=L'}^{L_3} \text{pr}\left\{\frac{N}{L-L_0}\frac{\hat{e}_{L_0} - \hat{e}_L}{\hat{e}_{L_0}} - 1 > \frac{M_n}{L_3} + o(1)\right\} \\ &\quad + \varepsilon/6. \end{aligned} \quad (48)$$

Following from $\lim_{N\to\infty} M_n = \infty$ and similar arguments as in the proof of Theorem 1, there exists a constant L_1 ($L_1 > L_0$) such that with probability at least $1 - \varepsilon/6$, it holds for all L satisfying $L > L_1$ that

$$\begin{aligned} &\text{pr}\left\{\frac{N}{L-L_0}\frac{\hat{e}_{L_0} - \hat{e}_L}{\hat{e}_{L_0}} - 1 > \frac{M_n}{L_3} + o(1)\right\} \\ &\leq \text{pr}\left(\left|\frac{N}{L\sigma^2}\|\hat{\Psi}_L - \Psi_{L_0}\|^2_{\Gamma} - 1\right| > 1\right) < cL^{-2} \end{aligned}$$

for some positive constant c. The remaining proof of consistency follows similar proof of Theorem 1.

Proof of efficiency:

Similar to the proof of Proposition 1, minimizing $\text{BC}(n, L)$ is equivalent to minimizing $\overline{S}_n(L) = (N + 2L + \delta_n(L))\hat{e}_L$, where

$$\delta_n(L) = 2M_n\frac{N}{n}\sum_{k=1}^{L}\frac{1}{k} - 2L + o(1) \quad (49)$$

We first prove

$$\lim_{n\to\infty}\max_{1\leq L\leq \hat{L}_{\text{AIC}}}\frac{|\delta_n(L)|}{N} = 0 \quad \text{in probability.} \quad (50)$$

This immediately follows from $\hat{L}_{\text{AIC}} \leq L^{(n)}_{\max} = o(n)$ and $M_n < L_0^{(n)} = O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.

We define $S_n(L) = (N+2L)\hat{e}_L$. Using Equation (4.1), Proposition 3.2, Lemma 4.1, Proposition 4.1 of [19], we obtain

$$\begin{aligned} S_n(L) - S_n(\hat{L}_{\text{AIC}}) =& NC_n(L) - NC_n(\hat{L}_{\text{AIC}}) \\ &+ \tau_L + \tau_{\hat{L}_{\text{AIC}}} \end{aligned} \quad (51)$$

where τ_L is used to denote a generic quantity that is negligible compared with $NC_n(L)$ uniformly in $1 \leq L \leq L^{(n)}_{\max}$, namely

$$\lim_{n\to\infty}\max_{1\leq L\leq L^{(n)}_{\max}}\frac{\tau_L}{NC_n(L)} = 0 \quad \text{in probability.}$$

From the proof of Theorem 4.1 in [19],

$$\overline{S}_n(L) = S_n(L) + \delta_n(L)\sigma^2(1 - 2L/N) + \tau_L. \quad (52)$$

We further obtain from (51) and (52) that

$$\begin{aligned} &\overline{S}_n(L) - \overline{S}_n(\hat{L}_{\text{AIC}}) \\ &= NC_n(L) - NC_n(\hat{L}_{\text{AIC}}) + \delta_n(L)\sigma^2 - \delta_n(\hat{L}_{\text{AIC}})\sigma^2 + \\ &\qquad \tau_L + \tau_{\hat{L}_{\text{AIC}}} - \frac{2L}{N}\sigma^2\delta_n(L) + \frac{2\hat{L}_{\text{AIC}}}{N}\sigma^2\delta_n(\hat{L}_{\text{AIC}}) \end{aligned} \quad (53)$$

$$\begin{aligned} &= NC_n(L) - NC_n(\hat{L}_{\text{AIC}}) + \delta_n(L)\sigma^2 - \delta_n(\hat{L}_{\text{AIC}})\sigma^2 \\ &\qquad + \tau_L + \tau_{\hat{L}_{\text{AIC}}}, \end{aligned} \quad (54)$$

where (54) is because the last four terms of (53) can be written as $\tau_L + \tau_{\hat{L}_{\text{AIC}}}$ due to (50) and $NC_n(L) > L\sigma^2$.

Recall that BC selects $\hat{L}$ that minimizes $\overline{S}_n(L)$ over $L = 1, \ldots, \hat{L}_{\text{AIC}}$. Thus, $\overline{S}_n(\hat{L}) - \overline{S}_n(\hat{L}_{\text{AIC}}) \leq 0$ almost surely. Suppose that we have further proved

$$\begin{aligned} \lim_{n\to\infty}\text{pr}\Big(&\delta_n(L) \geq \delta_n(\hat{L}_{\text{AIC}}), \\ &L = 1, \ldots, \hat{L}_{\text{AIC}}\Big) = 1 \end{aligned} \quad (55)$$

(which we will prove later on), then

$$\begin{aligned} \lim_{n\to\infty}\text{pr}\Big(&NC_n(\hat{L}) - NC_n(\hat{L}_{\text{AIC}}) \\ &+ \tau_{\hat{L}} + \tau_{\hat{L}_{\text{AIC}}} \leq 0\Big) = 1. \end{aligned} \quad (56)$$

Dividing both sides of the inequality in (56) by $NC_n(\hat{L})$,

we obtain that for any fixed $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathrm{pr}\left(\frac{C_n(\hat{L})}{C_n(\hat{L}_{\mathrm{AIC}})} < 1+\varepsilon\right) = 1 \tag{57}$$

Since

$$\lim_{n\to\infty} \frac{C_n(\hat{L}_{\mathrm{AIC}})}{C_n(L_0^{(n)})} = 1 \quad \text{in probability,}$$

for any fixed $\varepsilon > 0$ we obtain

$$\lim_{n\to\infty} \mathrm{pr}\left(\frac{C_n(\hat{L})}{C_n(L_0^{(n)})} < 1+\varepsilon\right) = 1. \tag{58}$$

On the other hand, $C_n(\hat{L})/C_n(L_0^{(n)}) \geq 1$ by the definition of $L_0^{(n)}$. Therefore,

$$\lim_{n\to\infty} \frac{C_n(\hat{L})}{C_n(L_0^{(n)})} = 1 \quad \text{in probability} \tag{59}$$

which ensures the efficiency of the bridge criterion.

Due to the above arguments, it suffices to prove (55). It trivially holds for $L = \hat{L}_{\mathrm{AIC}}$, so we only need to prove for $L < \hat{L}_{\mathrm{AIC}}$. It follows from (49) that

$$\delta_n(L) - \delta_n(\hat{L}_{\mathrm{AIC}}) = g(L) + o(1) \tag{60}$$

where $g(L)$ $(1 \leq L < \hat{L}_{\mathrm{AIC}})$ is defined as

$$g(L) = -2M_n \frac{N}{n} \sum_{k=L+1}^{\hat{L}_{\mathrm{AIC}}} \frac{1}{k} - 2(L - \hat{L}_{\mathrm{AIC}}). \tag{61}$$

for a given $\hat{L}_{\mathrm{AIC}}$. It is easy to verify that $g(L)$ achieves its minimum over $L = 1, \ldots, \hat{L}_{\mathrm{AIC}} - 1$ at either $L = 1$ or $L = \hat{L}_{\mathrm{AIC}} - 1$. For $L = 1$, we obtain from (19), (20), and the inequality $\sum_{k=1}^{L} k^{-1} < \log L + 1$ $(\forall L \in \mathbb{N})$ that

$$\begin{aligned} g(1) &= -2M_n \frac{N}{n} \sum_{k=2}^{\hat{L}_{\mathrm{AIC}}} \frac{1}{k} - 2(1 - \hat{L}_{\mathrm{AIC}}) \\ &> -2M_n \log \hat{L}_{\mathrm{AIC}} + 2(\hat{L}_{\mathrm{AIC}} - 1) \\ &> 2q_1 M_n \log \hat{L}_{\mathrm{AIC}}. \end{aligned} \tag{62}$$

for some fixed positive constant q_1, with probability approaching one as n tends to infinity. For $L = \hat{L}_{\mathrm{AIC}} - 1$, we get

$$\begin{aligned} g(\hat{L}_{\mathrm{AIC}} - 1) &= -2\frac{M_n}{\hat{L}_{\mathrm{AIC}}} \frac{N}{n} + 2 \\ &> -\frac{2}{\log \hat{L}_{\mathrm{AIC}}} + 2 > 1 \end{aligned} \tag{63}$$

with probability approaching one as n tends to infinity. Finally, from (62) and (63) it is easy to verify that (60) is no less than zero, with probability approaching one as n tends to infinity.

Remark 10: Lemma 5 seems not directly applicable to proving the efficiency of the adjusted bridge criterion. In the above proof of efficiency, we have weakened the second condition (i.e., equality (36)) of Lemma 5.

APPENDIX G
PROOF OF PROPOSITION 2

In the well-specified scenario, conditioning on $\hat{L}_{\mathrm{AIC}} \neq L_0$, it follows from $\hat{L}_{\mathrm{BC}}, \hat{L}_{\mathrm{BIC}} \to L_0$ in probability (due to the consistency of BC and BIC) that $\mathrm{PI}_n \to 1$ in probability, as n tends to infinity; conditioning on $\hat{L}_{\mathrm{AIC}} = L_0$, it follows from the definition and $\hat{L}_{\mathrm{BIC}} = L_0$ with probability tending to one that $\lim_{n\to\infty} \mathrm{PI}_n = 1$ in probability.

In the mis-specified scenario, dividing both the nominator and denominator by $\hat{L}_{\mathrm{AIC}}$ in (22) for $\hat{L}_{\mathrm{AIC}} \neq \hat{L}_{\mathrm{BIC}}$, PI_n becomes

$$\frac{|\hat{L}_{\mathrm{BC}}/\hat{L}_{\mathrm{AIC}} - 1|}{|\hat{L}_{\mathrm{BC}}/\hat{L}_{\mathrm{AIC}} - 1| + |(\hat{L}_{\mathrm{BC}}/\hat{L}_{\mathrm{AIC}} - 1) - (\hat{L}_{\mathrm{BIC}}/\hat{L}_{\mathrm{AIC}} - 1)|}.$$

To prove $\lim_{n\to\infty} \mathrm{PI}_n = 0$, it suffices to prove that

$$\lim_{n\to\infty} \mathrm{pr}\{\hat{L}_{\mathrm{BIC}} = \hat{L}_{\mathrm{AIC}}\} = 0, \tag{64}$$

and

$$\lim_{n\to\infty} \frac{|\hat{L}_{\mathrm{BC}}/\hat{L}_{\mathrm{AIC}} - 1|}{|\hat{L}_{\mathrm{BIC}}/\hat{L}_{\mathrm{AIC}} - 1|} = 0 \quad \text{in probability.} \tag{65}$$

Similar to the proof of Proposition 1, minimizing $\mathrm{BIC}(n, L)$ is equivalent to minimizing $\overline{S}_n(L) = (N + 2L + \delta_n(L))\hat{e}_L$, where $\delta_n(L) = (\log n - 2)L + o(1)$. Similar to the proof of (54), we obtain

$$\begin{aligned} \overline{S}_n(L) - \overline{S}_n(L') = NC_n(L) - NC_n(L') + \delta_n(L)\sigma^2 \\ - \delta_n(L')\sigma^2 + \tau_L + \tau_{L'} \end{aligned} \tag{66}$$

where τ_L is used to denote a generic quantity that is negligible compared with $NC_n(L)$ uniformly in $1 \leq L \leq L_{\max}^{(n)}$, and L, L' are any given integers in between 1 and $\hat{L}_{\mathrm{AIC}}$. By fixing L' to any given number, say $L' = 1$, we can see that $\hat{L}_{\mathrm{BIC}}$ minimizes $NC_n(L) + \delta_n(L)\sigma^2 + \tau_L = N\overline{C}_n(L)\{1 + o_p(1)\}$, or equivalently $\overline{C}_n(L)\{1 + o_p(1)\}$, where

$$\overline{C}_n(L) = C_n(L) + \frac{\log n - 2}{N} L\sigma^2. \tag{67}$$

Suppose that $\overline{C}_n(L)$ achieves its minimum at $L_*^{(n)}$. Clearly, $\lim_{n\to\infty} \overline{C}_N(L_*^{(n)})/\overline{C}_N(\hat{L}_{\mathrm{BIC}}) = 1$ in probability. From $C_n(L) = L\sigma^2/N + \|\Psi_L - \Psi_\infty\|_\Gamma^2$, $\overline{C}_n(L)$ can be rewritten as $\overline{C}_n(L) = L\sigma^2/\{N/(\log n - 1)\} + \|\Psi_L - \Psi_\infty\|_\Gamma^2$. Since Assumption 5 guarantees that $C_n(L)$ is regular, so is $\overline{C}_n(L)$ (by recognizing $N/(\log n - 1)$ as N). Therefore, $\lim_{n\to\infty} L_*^{(n)}/\hat{L}_{\mathrm{BIC}} = 1$ in probability.

From the assumption $\lim_{n\to\infty} L_*^{(n)}/L_0^{(n)} = 0$, we further obtain

$$\lim_{n\to\infty} \frac{\hat{L}_{\text{BIC}}}{\hat{L}_{\text{AIC}}} = \lim_{n\to\infty} \frac{\hat{L}_{\text{BIC}}}{L_*^{(n)}} \frac{L_*^{(n)}}{L_0^{(n)}} \frac{L_0^{(n)}}{\hat{L}_{\text{AIC}}} = 1 \times 0 \times 1 = 0 \quad \text{in probability.} \tag{68}$$

It follows from (68) and $\lim_{n\to\infty} \hat{L}_{\text{BC}}/\hat{L}_{\text{AIC}} = 1$ (proved in Theorem 2) that the desired equalities (64) and (65) hold.

In the example of algebraic decay, namely $\|\Psi_L - \Psi_\infty\|_\Gamma^2 = cL^{-\gamma}$, we have

$$C_n(L) = \frac{L}{N}\sigma^2 + cL^{-\gamma},$$
$$\overline{C}_n(L) = \frac{L(\log n - 1)}{N}\sigma^2 + cL^{-\gamma},$$

with minimum achieved at $L_0^{(n)} = (c\gamma N/\sigma^2)^{1/(\gamma+1)}$ and $L_*^{(n)} = \{c\gamma N/(\sigma^2 \log n - \sigma^2)\}^{1/(\gamma+1)}$, respectively. Clearly, the assumption $\lim_{n\to\infty} L_*^{(n)}/L_0^{(n)} = 0$ is satisfied.

APPENDIX H
PROOF OF THEOREM 3

For brevity, we highlight the major changes in the proof of Theorem 2. The technical lemmas that need to be adapted from Assumptions 1, 2 to Assumptions 1′, 2′ are: Lemmas 2 and 3 in our appendix (used in the proof of consistency), and Proposition 3.2, Lemma 4.1, Proposition 4.1 in [19] (used in the proof of asymptotic efficiency). We briefly prove them under the surrogate assumptions. The proof borrows some technical results from [20].

For Lemma 2, the second identity follows from the same proof as [19, Lemma 3.3]. To prove the first identity, we use [20, Proposition 1]. It states that for any $q > 0$ there exists a constant $c > 0$, such that $E\|\hat{\Gamma}_L^{-1} - \Gamma_L^{-1}\|^q \le c(N^{-1}L^2)^{q/2}$ for all $1 \le L \le L_{\max}^{(n)}$ and all sufficiently large n. Therefore, for any $\varepsilon > 0$ we have

$$\begin{aligned}
&\text{pr}\left(\max_{1\le L\le L_{\max}^{(n)}} \|\hat{\Gamma}_L - \Gamma_L\| > \varepsilon\right) \\
&\le \sum_{1\le L\le L_{\max}^{(n)}} \text{pr}\big(\|\hat{\Gamma}_L - \Gamma_L\| > \varepsilon\big) \\
&\le \varepsilon^{-q} \sum_{1\le L\le L_{\max}^{(n)}} E\big(\|\hat{\Gamma}_L - \Gamma_L\|^q\big) \\
&\le \varepsilon^{-q} c N^{-q/2} (L_{\max}^{(n)})^{1+q} \to 0
\end{aligned}$$

as $n \to \infty$, as long as we choose $q > 2/\delta$ (with δ being defined in Assumption 2′).

For Lemma 3, we only need to replace Lemma 1 with [20, Lemma 3] (under the surrogate assumptions), and the remaining proof is the same.

Proposition 3.2 in [19] can be replaced by [20, Lemma 5].

Lemma 4.1 in [19] can be proved by combining our adapted versions of [19, Proposition 3.2] and Lemma 2, following similar proof as was in [19].

Proposition 4.1 in [19] can be replaced with [20, Remark 6] under the surrogate assumptions.

APPENDIX I
BOUNDS ON THE OVERFITTING PROBABILITY UNDER FINITE SAMPLE SIZES

To derive analytical bounds for the overfitting probability of using the bridge criterion in (18), we assume that the filter has finite size L_0. Let $Z_1, \ldots, Z_m$ denote independent χ_1^2 random variables. For any $c > E(Z_1) = 1$ and positive integer m, Chernoff's bound [39] gives

$$\begin{aligned}
\text{pr}\left(\sum_{i=1}^m Z_i > cm\right) &< \left[\min_{u>0} \exp\big\{-uc + \log E(e^{uZ_1})\big\}\right]^m \\
&= \left[c\exp\{-(c-1)\}\right]^{\frac{m}{2}} .
\end{aligned}$$

The probability of choosing L ($L > L_0$) is upper bounded by

$$\begin{aligned}
\text{pr}\{\hat{L} = L\} &\le \text{pr}\left\{\sum_{k=L_0+1}^{L} \hat{g}_k > \frac{2M_n}{n} \sum_{k=L_0+1}^{L} \frac{1}{k}\right\} \\
&\le \text{pr}\left\{\sum_{k=L_0+1}^{L} n\hat{g}_k > \frac{2M_n}{L}(L - L_0)\right\},
\end{aligned}$$

which can be further approximately upper bounded by

$$\begin{aligned}
&\text{pr}\left\{\sum_{k=1}^{L-L_0} Z_k > \frac{2M_n}{L}(L - L_0)\right\} \\
&< \left\{\frac{2M_n}{L}\exp\left(-\frac{2M_n}{L} + 1\right)\right\}^{\frac{L-L_0}{2}}
\end{aligned}$$

since $n\hat{g}_{L_0+1}, \ldots, n\hat{g}_L$ are asymptotically independent and distributed χ_1^2. Furthermore, we can derive the following tighter bound, the technical detail of which is summarized in Proposition 3 below. For a positive integer k, define $\mathfrak{A}(k) = \{[a_1, \ldots, a_k]^{\text{T}} : a_1 + \cdots + ka_k = k,\ a_1, \ldots, a_k \text{ are nonnegative integers}\}$. For $L > L_0$, we have the approximation

$$\begin{aligned}
\text{pr}(\hat{L} = L\} &\le \sum_{\mathfrak{A}(L-L_0)} \prod_{j=1}^{L-L_0} \frac{1}{a_j!}\left(\frac{\eta_j}{j}\right)^{a_j}, \\
\eta_j &= 1 - F_{\chi^2_{L-L_0}}\left\{\frac{2M_n}{L}(L - L_0)\right\},
\end{aligned} \tag{69}$$

where $F_{\chi^2_{L-L_0}}(\cdot)$ denotes the cumulative distribution function of $\chi^2_{L-L_0}$.

Proposition 3: Suppose that the true autoregressive filter has finite size L_0, and that the selected order is L $(L_0 < L \le M_n)$. Given any $0 < \varepsilon < 1$, there exists a positive integer n_1 such that for all $n > n_1$,

$$\mathrm{pr}(\hat{L} = L) \le \sum_{\mathfrak{A}(L-L_0)} \prod_{j=1}^{L-L_0} \frac{1}{a_j!}\left(\frac{\eta_j}{j}\right)^{a_j} + \varepsilon, \qquad (70)$$

$$\eta_j = 1 - F_{\chi^2_j}\left\{\frac{2M_n}{L} j\right\}$$

where $F_{\chi^2_j}(\cdot)$ denotes the cumulative distribution function of χ^2_j. In a special case where M_n does not depend on n, we have

$$\limsup_{n\to\infty} \mathrm{pr}(\hat{L} = L) \le \sum_{\mathfrak{A}(L-L_0)} \prod_{j=1}^{L-L_0} \frac{1}{a_j!}\left(\frac{\eta_j}{j}\right)^{a_j}.$$

To prove Proposition 3, we will need the following lemma which was also used in deriving the overfitting probability of AIC [18].

Lemma 6: [40, Equality 3.4] Let $Z_1, \ldots, Z_k$ denote independent and identically distributed random variables and $T_i = Z_1 + \cdots + Z_i \, (1 \le i \le k)$. Define

$$\mathfrak{A}(k) = \{[a_1, a_2, \ldots, a_k]^{\mathrm{T}} : a_1 + 2a_2 + \cdots + ka_k = k, \; a_1, \ldots, a_k \text{ are nonnegative integers}\}.$$

The identity

$$\mathrm{pr}\left\{\bigcap_{j=1}^{k}(T_j > 0)\right\} = \sum_{\mathfrak{A}(k)} \prod_{j=1}^{k} \frac{1}{a_j!}\left(\frac{\eta_j}{j}\right)^{a_j},$$

holds, where $\eta_j = \mathrm{pr}(T_j > 0)$.

Proof of Proposition 3:

Let $Z_1, \ldots, Z_{L-L_0}$ denote independent χ^2_1 random variables. The probability of choosing L $(L > L_0)$ is upper bounded by

$$\begin{aligned}
&\mathrm{pr}\{\hat{L} = L\} \\
&\le \mathrm{pr}\left\{\sum_{k=m+1}^{L} \hat{g}_k > \frac{2M_n}{n}\sum_{k=m+1}^{L}\frac{1}{k} \quad (L_0 \le m \le L-1)\right\} \\
&\le \mathrm{pr}\left\{\sum_{k=m+1}^{L} n\hat{g}_k > \frac{2M_n}{L}(L-m) \quad (L_0 \le m \le L-1)\right\} \\
&= \mathrm{pr}\left\{\sum_{k=m+1}^{L} Z_{L+1-k} > \frac{2M_n}{L}(L-m) - \sum_{k=m+1}^{L}\Lambda^{(n)}_{L+1-k} \right. \\
&\qquad \left. (L_0 \le m \le L-1)\right\},
\end{aligned}$$

where $\Lambda^{(n)}_1, \ldots, \Lambda^{(n)}_{L-L_0}$ are random variables that converge in distribution to zero as n tends to infinity. Define $\Lambda^{(n)} = \max\{|\Lambda^{(n)}_1|, \ldots, |\Lambda^{(n)}_{L-L_0}|\}$, which also converges in distribution to zero. Given any $0 < \delta < 1$ (which will be determined later), there exists a positive integer n_1 such that for all $n > n_1$, $\mathrm{pr}(\Lambda^{(n)} > \delta) < \varepsilon/2$. For any given $n > n_1$, we define $T_j = \sum_{k=1}^{j}(Z_k - 2M_n/L + \delta)$ for $j = 1, \ldots, L - L_0$. From Lemma 6, we obtain

$$\begin{aligned}
&\mathrm{pr}\{\hat{L} = L\} \\
&\le \mathrm{pr}\left\{\sum_{k=m+1}^{L} Z_{L+1-k} > \frac{2M_n}{L}(L-m) - \delta\,(L-m) \right. \\
&\qquad \left. (L_0 \le m \le L-1)\right\} + \frac{\varepsilon}{2} \\
&= \mathrm{pr}\left\{\bigcap_{j=1}^{L-L_0}(T_j > 0)\right\} + \frac{\varepsilon}{2} \\
&\le \sum_{\mathfrak{A}(L-L_0)} \prod_{j=1}^{L-L_0} \frac{1}{a_j!\; j^{a_j}}\{P(T_j > 0)\}^{a_j} + \frac{\varepsilon}{2} \\
&= \sum_{\mathfrak{A}(L-L_0)} \prod_{j=1}^{L-L_0} \frac{1}{a_j!\; j^{a_j}}\left[1 - F_{\chi^2_j}\{(c-\Delta c)j\}\right]^{a_j} + \frac{\varepsilon}{2}
\end{aligned} \qquad (71)$$

where $c = 2M_n/L \ge 2, \Delta c = \delta < 1$. Let $f_j(x) = \{2^{j/2}\Gamma(j/2)\}^{-1}x^{j/2-1}\exp(-x/2)$ denote the probability density function of χ^2_j, or the derivative of $F_{\chi^2_j}(x)$, where $\Gamma(\cdot)$ denotes the Gamma function. It is easy to verify that $f_j(x)$ is strictly decreasing for $x > j$. By the mean value theorem, there exists a $\hat{c}$ such that $1 < c - \Delta c < \hat{c} < c$, and

$$\begin{aligned}
1 - F_{\chi^2_j}\{(c-\Delta c)j\} &= 1 - F_{\chi^2_j}(cj) + f_j(\hat{c}j)\Delta c \\
&< 1 - F_{\chi^2_j}(cj) + f_j(j)\Delta c\,.
\end{aligned}$$

Applying the above inequality to (71), we obtain

$$\begin{aligned}
\mathrm{pr}\{\hat{L} = L\} \le \sum_{\mathfrak{A}(L-L_0)} \prod_{j=1}^{L-L_0} \frac{1}{a_j!\; j^{a_j}}\left[1 - F_{\chi^2_j}\{cj\}\right]^{a_j} \\
+ \delta c_L + \frac{\varepsilon}{2}
\end{aligned}$$

where $c_L > 1/2$ is some constant that depends only on L. Finally, we obtain (70) by letting $\delta = \varepsilon/(2c_L)$.

Remark 11: In an experiment, we choose the true filter $\Psi_3 = [0.8, 0.64, 0.512]^{\mathrm{T}}$ and independently generate 1000 time series of size 1000 for the cases $M_n = 6$ and $M_n = 7$. The frequencies of $\hat{L} = 4, \ldots, M_n$ for the generated time series along with the upper bounds are summarized in Table IV.

	$M_n = 7$				$M_n = 6$		
L	4	5	6	7	4	5	6
Frequency (in percentages)	6.4	1.6	1.1	0.6	7.2	2.9	2.7
Upper bounds (in percentages)	24.8	6.1	1.9	0.7	28.86	9.1	3.7

TABLE IV: The frequency of the selected orders and the approximate upper bounds from (69)

REFERENCES

[1] T. Anderson, "Determination of the order of dependence in normally distributed time series," DTIC Document, Tech. Rep., 1962.

[2] H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Statist. Math.*, vol. 21, no. 1, pp. 243–247, 1969.

[3] R. J. Bhansali and D. Y. Downham, "Some properties of the order of an autoregressive model selected by a generalization of akaike's fpe criterion," *Biometrika*, vol. 64, no. 3, pp. 547–551, 1977.

[4] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. Springer, 1998, pp. 199–213.

[5] P. M. Broersen, "Finite sample criteria for autoregressive order selection," *IEEE Trans. Signal Process.*, vol. 48, no. 12, pp. 3550–3558, 2000.

[6] H. Akaike, "A bayesian extension of the minimum aic procedure of autoregressive model fitting," *Biometrika*, vol. 66, no. 2, pp. 237–242, 1979.

[7] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[8] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[9] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Statist. Soc. Ser. B*, vol. 41, no. 2, pp. 190–195, 1979.

[10] G. Claeskens and N. L. Hjort, "The focused information criterion," *J. Am. Stat. Assoc.*, vol. 98, no. 464, pp. 900–916, 2003.

[11] G. Claeskens, C. Croux, and J. Van Kerckhoven, "Prediction-focused model selection for autoregressive models," *Aust. N. Z. J. Stat.*, vol. 49, no. 4, pp. 359–379, 2007.

[12] E. Parzen, "Some recent advances in time series modeling," *IEEE Trans. Automat. Control*, vol. 19, no. 6, pp. 723–730, 1974.

[13] J. Rissanen, "Stochastic complexity and modeling," *The annals of statistics*, pp. 1080–1100, 1986.

[14] C.-Z. Wei, "On predictive least squares principles," *The Annals of Statistics*, pp. 1–42, 1992.

[15] E. Hemerly and M. Davis, "Strong consistency of the pls criterion for order determination of autoregressive processes," *Ann. Statist.*, vol. 17, no. 2, pp. 941–946, 1989.

[16] J. G. de Gooijer, B. Abraham, A. Gould, and L. Robinson, "Methods for determining the order of an autoregressive-moving average process: A survey," *Int. Stat. Rev.*, vol. 53, no. 3, pp. 301–329, 1985.

[17] J. Shao, "An asymptotic theory for linear model selection," *Statist. Sinica*, vol. 7, no. 2, pp. 221–242, 1997.

[18] R. Shibata, "Selection of the order of an autoregressive model by akaike's information criterion," *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[19] ——, "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *Ann. Statist.*, vol. 8, no. 1, pp. 147–164, 1980.

[20] C.-K. Ing and C.-Z. Wei, "Order selection for same-realization predictions in autoregressive processes," *Ann. Statist.*, vol. 33, no. 5, pp. 2423–2474, 2005.

[21] C.-K. Ing, C.-y. Sin, and S.-H. Yu, "Model selection for integrated autoregressive processes of infinite order," *Journal of Multivariate Analysis*, vol. 100, no. 3, pp. 57–71, 2012.

[22] Y. Yang, "Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation," *Biometrika*, vol. 92, no. 4, pp. 937–950, 2005.

[23] C.-K. Ing, "Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series," *Ann. Statist.*, vol. 35, no. 3, pp. 1238–1277, 2007.

[24] Y. Yang, "Prediction/estimation with simple linear models: is it really that simple?" *Econom. Theory*, vol. 23, no. 01, pp. 1–36, 2007.

[25] W. Liu and Y. Yang, "Parametric or nonparametric? a parametricness index for model selection," *Ann. Statist.*, pp. 2074–2102, 2011.

[26] T. v. Erven, P. Grünwald, and S. De Rooij, "Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the aic–bic dilemma," *J. R. Stat. Soc. Ser. B.*, vol. 74, no. 3, pp. 361–417, 2012.

[27] S. van der Pas and P. Grünwald, "Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection," *arXiv:1408.5724*, 2014.

[28] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *J. Econometrics*, vol. 187, no. 1, pp. 95–112, 2015.

[29] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control*. John Wiley & Sons, 2011, vol. 734,.

[30] H. Akaike, "Statistical predictor identification," *Ann. Inst. Statist. Math.*, vol. 22, no. 1, pp. 203–217, 1970.

[31] N. Rayner, D. E. Parker, E. Horton, C. Folland, L. Alexander, D. Rowell, E. Kent, and A. Kaplan, "Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. D14, 2003.

[32] A. P. Dawid, "Present position and potential developments: Some personal views: Statistical theory: The prequential approach," *J. Roy. Statist. Soc. Ser. A*, pp. 278–292, 1984.

[33] B. Dieppois, A. Durand, M. Fournier, and N. Massei, "Links between multidecadal and interdecadal climatic oscillations in the north atlantic and regional climate variability of northern france and england since the 17th century," *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 10, pp. 4359–4372, 2013.

[34] T. W. Anderson, *The statistical analysis of time series*. John Wiley & Sons, 1971.

[35] J. Ding, M. Noshad, and V. Tarokh, "Data-driven learning of the number of states in multi-state autoregressive models," *Proc. of 53rd Allerton Conf. Commun., Control, Computing*, 2015.

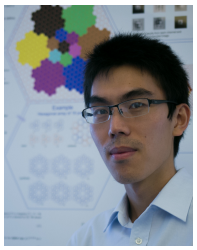
[36] F. L. Ramsey *et al.*, "Characterization of the partial autocorrelation function," *Ann. Statist.*, vol. 2, no. 6, pp. 1296–1301, 1974.

[37] O. Barndorff-Nielsen and G. Schou, "On the parametrization of autoregressive models by partial autocorrelations," *J. Multivariate Anal.*, vol. 3, no. 4, pp. 408–419, 1973.

[38] C. Walck, *Handbook on statistical distributions for experimentalists,*. University of Stockholm Internal Report SUF-PFY/96-01, 2007.

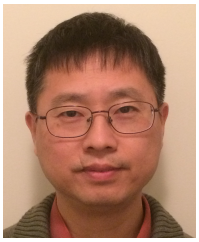
[39] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Stat.*, vol. 23, no. 4, pp. 493–507, 1952.

[40] F. Spitzer, "A combinatorial lemma and its application to probability theory," *Trans. Amer. Math. Soc.*, vol. 82, no. 2, pp. 323–339, 1956.



Jie Ding (jieding@fas.harvard.edu) received Bachelor of Science degree from Tsinghua University in May 2012, majoring in mathematics and electrical engineering. He received Master of Arts degree in statistics in May 2016, and Ph.D. degree in Engineering Sciences in March 2017, both from Harvard University. His research areas are statistical inference, machine learning, signal processing, and combinatorics. His recent goal is to establish a reliable, efficient, and widely applicable time series prediction system.

Vahid Tarokh (vahid@seas.harvard.edu) received the Ph.D. in electrical engineering from the University of Waterloo, Ontario, Canada in 1995. He then worked at AT&T Labs-Research and AT&T wireless services until August 2000 as Member, Principal Member of Technical Staff, and finally as the Head of the Department of Wireless Communications and Signal Processing. In Sept 2000, he joined MIT as an Associate Professor where he worked until June 2002. In June 2002, he joined Harvard University as a Professor of Electrical Engineering. He was named Perkins Professor and Vinton Hayes Senior Research Fellow of Electrical Engineering in 2005. His current research areas are in statistical signal processing and data analysis.



Yuhong Yang (yyang@stat.umn.edu) received his Ph.D from Yale in statistics in 1996. He then joined Department of Statistics at Iowa State University and moved to the University of Minnesota in 2004. He has been full professor there since 2007. His research interests include model selection, multi-armed bandit problems, forecasting, high-dimensional data analysis, and machine learning. He has published in journals in several fields, including Annals of Statistics, IEEE Transaction on Information Theory, Journal of Econometrics, Journal of Approximation Theory, Journal of Machine Leaning Research, and International Journal of Forecasting. He is a fellow of Institute of Mathematical Statistics.