Hindawi
Journal of Probability and Statistics
Volume 2017, Article ID 1235979, 12 pages
https://doi.org/10.1155/2017/1235979

Research Article

Bootstrap Order Determination for ARMA Models: A Comparison between Different Model Selection Criteria

Livio Fenga

University of California San Diego (UCSD), San Diego, CA, USA

Correspondence should be addressed to Livio Fenga; lfenga@ucsd.edu

Received 31 July 2016; Accepted 19 March 2017; Published 16 April 2017

Academic Editor: Ramón M. Rodríguez-Dagnino

The present paper deals with the order selection of models of the class for autoregressive moving average. A novel method—previously designed to enhance the selection capabilities of the Akaike Information Criterion and successfully tested—is now extended to the other three popular selectors commonly used by both theoretical statisticians and practitioners. They are the final prediction error, the Bayesian information criterion, and the Hannan-Quinn information criterion which are employed in conjunction with a semiparametric bootstrap scheme of the type sieve.

1. Introduction

Autoregressive moving average (ARMA) models [1] are a popular choice for the analysis of stochastic processes in many fields of applied and theoretical research. They are mathematical tools employed to model the persistence, over time and space, of a given time series. They can be used for a variety of purposes, for example, the generation of predictions of future values, to remove the autocorrelation structure from a time series (prewhitening) or to achieve a better understanding of a physical system. As it is well known, performances of an ARMA model are critically affected by the determination of its order: once properly built and tested, such models can be successfully employed to describe the reality, for example, trend patterns of economic variables and temperature oscillations in a given area, or to build futures scenarios through simulation exercises. Model order choice plays a key role not only for the validity of the inference procedures but also, from a more general point of view, for the fulfillment of the fundamental principle of parsimony [2, 3]. Ideally, the observation of this principle leads to choosing models showing simple structures on one hand but able to provide an effective description of the data set under investigation on the other hand. Less parsimonious models tend to extract idiosyncratic information and therefore are prone to introduce high variability in the estimated parameters. Such a variability determines for the model a lack of generalization capabilities (e.g., when new data become available), even though, by adding more and more parameters, an excellent fit of data is usually obtained [4]. Overfitting is more likely to occur when the system under investigation is affected by different sources of noise, for example, related to changes in survey methodologies, time evolving processes, and missing observations. These phenomena, very common and in many cases simply unavoidable in "real life" data, might have a significant impact on the quality of the data set at hand. Under noisy conditions, a too complex model is likely to fit the noise components embedded in the time series and not just the signal and therefore it is bound to yield poor future values' predictions. On the other hand, bias in the estimation process arises when underfitted models are selected, so that only a suboptimal reconstruction of the underlying Data Generating Process (DGP) can be provided. As it will be seen, bias also arises as a result of the uncertainty conveyed by the process *itself* of model selection. ARMA model order selection is a difficult step in time series analysis. This issue has attracted a lot of attention so that, according to different philosophies, theoretical and practical assumptions as well as several methods, both parametric and nonparametric, have been proposed over the years as a result. Among them, bootstrap strategies [5–9] are gaining more and more acceptance among researchers and practitioners.

In particular, in [6] bootstrap-based procedures applied to the Akaike Information Criterion (AIC) [10, 11] in the case of ARMA models, called b-MAICE (bootstrap-Minimum AIC Estimate), has proven to enhance the small sample performances of this selector. The aim of this work is to extend such a procedure to different selectors, that is, final prediction error (FPE) [12] and two information based criteria, that is, Bayesian information criterion (BIC) [13, 14] and Hannan-Quinn criterion (HQC) [15, 16]. In particular, the present paper is aimed at giving empirical evidences of the quality of the bootstrap approach in model selection, by comparing it with the standard procedure, which, as it is well known, is based on the minimization of a selection criterion. In particular, the empirical study (presented in Section 4) has been designed to contrast the performances of each of the considered selectors both in nonbootstrap and bootstrap world. The validity of the proposed method is assessed not only in the case of pure ARMA processes, but also when real life phenomena are simulated and embedded in the artificial data. In practice, the problem of order determination is considered also when the observed series is contaminated with outliers and additive Gaussian noise. The last type of contamination has been employed, for example, in [17], for testing a model selection approach driven by information criteria in the autoregressive fractionally integrated moving average (ARFIMA) and ARMA cases. Such a source of disturbance has been employed here in order to test the degree of robustness of the method proposed against overfitting. As it will be seen, computer simulations show that the addition of white noise generates a number of incorrect specifications comparable to those resulting from the contamination of the process with outliers of the type innovation. Outliers are a common phenomenon in time series, considering the fact that real life time series from many fields, for example, economic, sociology, and climatology, can be subjected and severely influenced by interruptive events, such as strikes, outbreaks of war, unexpected heat or cold waves, and natural disasters [18, 19]. The issue is absolutely nontrivial, given that outliers can impact virtually all the stages of the analysis of a given time series. In particular, model identification can be heavily affected by additive outliers, as they can induce the selection of underfitted models as a result of the bias elements introduced into the inference procedures. In the simulation study (Section 4), outliers of the type additive (i.e., added to some observations) and innovative (i.e., embedded in the innovation sequence driving the process) [19] will be considered.

The remainder of the paper is organized as follows: in Section 2, after introducing the problem of order identification for time series, the considered selectors are illustrated along with the related ARMA identification procedure. In Section 3 the employed bootstrap selection method is illustrated and the bootstrap scheme briefly recalled. Finally, small sample performances of the proposed method will be assessed via Monte Carlo simulations in Section 4.

2. Order Selection for Time Series Models

A key concept underlying the present paper is that, in general, "reality" generates complex structures, possibly ∞-dimensional, so that a model can at best capture only the main features of the system under investigation in order to reconstruct a simplified version of a given phenomenon. Models are just approximations of a given (nontrivial) phenomenon and the related identification procedures could never lead to the determination of the "true" model. In general, there is no true model in a finite world. What we can do is to find the one giving the best representation of the underlying DGP, according to a predefined rule. In this section, after highlighting the role played by model selection procedures in generating uncertainty, we briefly introduce the models belonging to the class ARMA along with the order selectors considered. Finally, the information criterion-based standard selection procedure is illustrated.

2.1. Uncertainty in Model Selection. Uncertainty is an unfortunate, pervasive, and inescapable feature characterizing real life data which has to be faced continually by both researchers and practitioners. The framework dealt with here is clearly no exception: if the true model structure is an unattainable goal, approximation strategies have to be employed. Such strategies are generally designed on iterative basis and provide an estimate of the model structure which embodies, by definition, a certain amount of uncertainty. Common sources of uncertainty are those induced by the lack of discriminating power of the employed selector and by the so-called model selection bias [20, 21], which arises when a model is specified and fitted on the *same* data set. Unfortunately, not only are these two types of uncertainty not mutually exclusive but also statistical theory provides little guidance to quantify their effect in terms of bias introduced in the model as a result [22]. Particularly dangerous is this last form of uncertainty, as it is based upon the strong and unrealistic assumption of making correct inference *as if* a model is known to be true, while its determination has been made on the same set of data. On the other hand, the first source of uncertainty is somehow less serious, given its direct relationship with the size of the competition set, which is usually included in the design of the experiment. In practice, it is related to the fact that very close SC minima can be found in the model selection process, so that even small variations in the data set can cause the identification of different model structures. In general, trying to explain only in part the complexity conveyed in the observed process by means of as simple as possible structures is a way to minimize uncertainty in the model selection, as it is likely to lead to the definition of a smaller set of candidate models. This approach can be seen as an extension of the principle of parsimony to the competition set. In the sequel, how the proposed procedure, being aimed at replicating both the original process and the related selection procedure, has a positive effect in reducing both the considered sources of uncertainty will be emphasized [23].

2.2. The Employed Identification Criteria. Perhaps the most well-known model order selection criteria (SC), among those considered, are the AIC and the FPE, whose asymptotic equivalence to the F-test has been proved in [24]. AIC has been designed on information-theoretic basis as an asymptotically unbiased estimate of the Kullback-Leibler divergence [25] of the fitted model relative to the true model.

Assuming X_T, T being the sample size, to be randomly drawn from an unknown distribution $H(x)$ with density $h(x)$, the estimation of h is done by means of a parametric family of distributions, with densities $[f(x \mid \theta;\ \theta \in \Theta)]$, θ being the unknown parameters' vector. Denoting $f(z \mid \widehat{\theta})$ as the predictive density function, f as the true model, and h as the approximating one, Kullback-Leibler discrepancy can be expressed as follows:

$$I\left(h(z); f\left(z \mid \widehat{\theta}\right)\right) = \int h(z) \log h(z)\, dz - \int h(z) \log f\left(z \mid \widehat{\theta}\right) dz. \tag{1}$$

As the first term on the right hand side of (1) does not depend on the model, it can be neglected so that we can rewrite the distance in terms of the expected log likelihood, $L(X_T; H)$; that is,

$$\begin{aligned} L(X_T; H) &= \int h(z) \log f\left(z \mid \widehat{\theta}\right) dz \\ &= \int \log f\left(z \mid \widehat{\theta}\right) dH(z). \end{aligned} \tag{2}$$

This quantity can be estimated by replacing H with its empirical distribution $\widehat{H}$, so that we have that $L(X_T; \widehat{H}) = (1/T) \sum_{\alpha=1}^{T} \log f(X_\alpha \mid \widehat{\theta})$. This is an overestimated quantity of the expected log likelihood, given that $\widehat{H}$ is closer to $\widehat{\theta}$ than H. The related bias can be written as follows:

$$b(H) = E_H\left\{L\left(X_T; \widehat{H}\right) - L(X_T; H)\right\}, \tag{3}$$

and therefore an information criterion can be derived from the bias-corrected log likelihood; that is, $(1/T) \sum_{\alpha}^{T} \log f(X_\alpha \mid \widehat{\theta}) - B(\widehat{H})$

Denoting by k and T the number of estimated parameters and the sample size, respectively, Akaike proved that $b(H)$ is asymptotically equal to k/T, so that the information based criterion takes the form $L(X_T; \widehat{H}) + k/T$. By multiplying this quantity by -2, finally AIC is defined as $-2 \log L(X_T; \widehat{H}) + 2k$. In such a theoretical framework, AIC can be seen as a way to solve the Akaike Prediction Problem [6], that is, to find a model M_0 producing estimation of density $\widehat{f}$ minimizing *Kullback-Leibler* discrepancy (1). Originally conceived for AR process, extended to the ARMA case by Soderstrom and Stoica [24], FPE was designed as the minimizer of the one-step-ahead mean square forecast error, after taking in account the inflating effect of the estimated parameter. FPE statistic is defined as $\mathrm{FPE}(k) = [(1 + k/T)/(1 - k/T)]\widehat{\sigma}_{\varepsilon}^{2}(k)$, where $\widehat{\sigma}_{\varepsilon}^{2}$ is the estimated variance of the residuals and k is the model's size. A different perspective has led to the construction of BIC-type criteria, which are grounded on the maximization of the model posterior probability [14]. In more detail, they envision the specification of the prior distribution on parameter values and the models, respectively, denoted by $P(\theta \mid k)$ and $P(k)$, and their introduction into the analysis through the joint probability function $P(\theta, k) = P(k)P(\theta \mid k)$. Posterior probabilities for (θ, k) are then obtained through Bayes theorem, so that the value of k maximizing (4), that is,

$$P(k \mid X_t) \propto P(k) \int_{\theta \in \Theta} f(X_t; \theta, k)\, P(\theta \mid k)\, d\theta, \tag{4}$$

is found. With $f(X_T; \theta, k)$ being the likelihood function associated with both the data X_T and the model M_k, the selected order will be $\widehat{k} = \arg\max_k P(k \mid X_T)$. By assuming all the models equally probable, that is, $p(k) = 1/(k_{\max} + 1)$, the BIC criterion is hence defined by $-2 \log L(\widehat{\theta}) + 2k \log(T)$. The last criterion considered—constructed from the law of iterated algorithm—is the BIC, in which the penalty function grows at a very slow rate as the samples size increases. It is defined as follows: $\mathrm{HQC} = \log L(\widehat{\theta}) + 2k \log(\log(T))$.

All these selectors can be divided into two groups: one achieving asymptotic optimality [26] and one selection consistency. AIC and FPE fall in the first group, in the sense that the selected model asymptotically tends to reach the smallest average squared error [27, 28], if the true DGP is not included in the competition set. On the other hand, BIC and HQ are dimension consistent [29], in that the probability of selection of the "true" model approaches 1 as the sample size goes to infinity. However, it should be pointed out that such an asymptotic property holds only if the true density is in the set of the candidate models. In this regard, AIC and FPE as well as the other Shibata efficient criteria (e.g., Mallows C_p [30]) fail to select the "true" model asymptotically. As pointed out earlier, ∞-dimensionality of the "truth" implies for all the models being "wrong" to some extent—except in trivial cases—so that no set of competition models will ever encompass the true DGP. As long as this approach is held true, asymptotic efficient criteria might be preferred. In this case, one may argue a lack of significance in comparing any finite list of candidate models when we rule out the existence of a true one. Such an approach is justified in that, even if no model can ever represent the truth, we can achieve the goal to find the one being approximately correct. Conversely, if one does believe that the true density belongs to the model space, hence dimension consistent selection criteria can be preferred.

2.3. ARMA Model Selection through Minimization of Selection Criteria

In what follows, it is assumed that the observed time series $\{X_t\}_{t \in \mathbb{Z}^+}$ is a realization of a real valued, 0–mean, second-order stationary process, admitting an autoregressive moving average representation of orders p and q; that is, $x_t \sim \mathrm{ARMA}(p, q)$, with $(p, q) \in \mathbb{Z}^+$. Its mathematical expression is as follows:

$$\phi(B)(x_t) = \theta(B)\, \varepsilon_t; \quad t \in \mathbb{Z}^+, \tag{5}$$

with $\phi(z) = (1 - \sum_{i=1}^{p} \phi_i z^i)$ and $\theta(z) = (1 - \sum_{i=1}^{q} \theta_i z^i)$, being $\phi_i \in \mathbb{R}$ and $\theta_i \in \mathbb{R}$, AR polynomial, and MA polynomial, respectively. With B the backward shift operator, such that $B^k Y_t = Y_{t-k}$, is denoted whereas ε_t is assumed to be sequence of centered, uncorrelated variables with common variance σ^2. The parameters vector is denoted by $\boldsymbol{\Gamma}$. Standard assumptions of stationarity and invertibility, respectively, of AR and MA polynomials, that is,

$$|\phi(z)| \neq 0 \quad |z| \leq 1, \tag{6}$$

$$|\theta(z)| \neq 0 \quad |z| \leq 1, \tag{7}$$

are supposed to be satisfied. Finally, the ARMA parameters of the true underlying DGP (5) are denoted by (p°, q°) (i.e., $\{X\}_t^T \sim \text{ARMA}(p^{\circ}, q^{\circ})$) and the related model by $M^0(\mathbf{\Gamma})$.

Identification procedures of the best approximating model for M^0 is carried out on a priori specified set $\mathbf{\Lambda}$ of plausible candidate models M_g; that is,

$$\left\{\mathbf{\Lambda} \supseteq M_g\left(\widehat{\mathbf{\Gamma}}\right) \;\; g = 1, 2, \ldots, G\right\}, \tag{8}$$

where the chosen model, say $M_0(\widehat{\mathbf{\Gamma}}) = (\widehat{p}_0, \widehat{q}_0)$, is selected from (i.e., $[M_0(\widehat{\mathbf{\Gamma}}) \equiv (\widehat{p}_0, \widehat{q}_0) \subset \mathbf{\Lambda}] \approx M^0(\mathbf{\Gamma})$). In the ARMA case, each model $M_g \in \mathbf{\Lambda}$ represents a specific combination of autoregressive and moving average parameters (p, q). The set $\mathbf{\Lambda}$ is upper bounded by the two integers P and Q for the AR and MA part, respectively; that is,

$$\mathbf{\Lambda} = \{(p, q) : 0 \leq p_0 \leq P, \; 0 \leq q_0 \leq Q\}. \tag{9}$$

This assumption is a necessary condition for the abovementioned Shibata efficiency and dimension consistency properties to hold other than for the practical implementation of the procedure (the model space needs to be bounded). From an operational point of view, the four SC considered in this work, when applied to models of the class ARMA, take the following form:

$$\text{AIC}(p, q) = T \ln \widehat{\sigma}^2_{p,q} + 2(p + q + 1), \tag{10}$$

$$\text{FPE}(p, q) = \widehat{\sigma}^2_{p,q} \left\{\frac{T - (p + q + 1)}{T + (p + q + 1)}\right\}, \tag{11}$$

$$\text{BIC}(p, q) = T \ln\left[\widehat{\sigma}^2_{p,q}\right] + \left[(p + q + 1)\ln(T)\right], \tag{12}$$

$$\begin{aligned} \text{HQC}(p, q) &= T \ln\left[\widehat{\sigma}^2_{p,q}\right] \\ &\quad + \left[2(p + q + 1)\ln(\ln(T))\right], \end{aligned} \tag{13}$$

where $\widehat{\sigma}_{p,q}$ is an estimate of the Gaussian pseudo-maximum likelihood residual variance when fitting ARMA (p, q) models; that is,

$$\begin{aligned} \widehat{\sigma}^2_{p,q} &= \frac{1}{T - (p + q + 1)} \\ &\quad \cdot \sum_{\max(p,q)}^{T} \left[y_t - \left(\sum_{j=0}^{P} \phi_j y_{t-j} - \sum_{i=0}^{Q} \theta_i y_{t-i}\right)\right]^2. \end{aligned} \tag{14}$$

Equations (10)–(13) can be synthetically expressed as follows:

$$\text{SC}(\widehat{p}, \widehat{q}) = f\left(\widehat{\sigma}^2_{p,q}, \xi_{pq}\right), \tag{15}$$

where $\widehat{\sigma}^2_{p,q}$ is defined in Section 3 and ξ is the penalty term as a function of model complexity.

The standard identification procedure, here called for convenience Minimum Selection Criterion Estimation (MSCE), is based on the minimization of the SC. In practice, the model M_0 minimizing a given SC is the winner; that is,

$$M_0 : (\widehat{p}_0, \widehat{q}_0) = \arg \min_{p<P, q<Q} \text{SC}(p, q). \tag{16}$$

3. The Bootstrap Method

As already pointed out, in [6] a bootstrap selection method has been proposed to perform AIC-based ARMA structure identification. The comparative Monte Carlo experiment with its nonbootstrap counterpart, commonly referred to as MAICE (Minimum Akaike Information Criterion Expectation) procedure, gave empirical evidences in favor of *b*-MAICE procedure. Such results motivated us to extend this approach to other selectors (see (11), (12), and (13)). For convenience, the proposed generalized version of *b*-MAICE procedure has been called bMSE (Bootstrap Minimum Selector Expectation) procedure. Finally, in order to keep the paper as self-contained as possible, and to reduce uncertainty in the experimental outcomes, AIC has also been included in the experiment.

3.1. The Bootstrapped Selection Criteria. The proposed bMSE method relies on the bootstrapped version of a given SC, obtained by bootstrapping both the residual variance term $\widehat{\sigma}^2_{\varepsilon}$ and the penalty term, so that (15) becomes

$$\text{SC}^* = f\left[\left(\widehat{\sigma}^2_{p,q}\right)^*, \xi^*\right]. \tag{17}$$

The particularization of (17) to the criteria object of this study is straightforward and yields their bootstrapped versions; that is,

$$\begin{aligned} \text{AIC}^*(p, q) &= T \ln\left(\widehat{\sigma}^2_{p,q}\right)^* + 2(p + q + 1)^*, \\ \text{FPE}^*(p, q) &= \left(\widehat{\sigma}^2_{p,q}\right)^* \left\{\frac{T - (p + q + 1)}{T + (p + q + 1)}\right\}^*, \\ \text{BIC}^*(p, q) &= T \ln\left[\widehat{\sigma}^2_{p,q}\right]^* + \left[(p + q + 1)\ln(T)\right]^*, \\ \text{HQC}^*(p, q) &= T \ln\left[\widehat{\sigma}^2_{p,q}\right]^* \\ &\quad + \left[2(p + q + 1)\ln(\ln(T))\right]^*, \end{aligned} \tag{18}$$

with T, p, q being as above defined and $\sigma^2_{p,q}$ being the residual variance of the residuals from the fitting of the bootstrapped series y_t^* with its ARMA estimate $\widehat{y}_t^*$. In symbols,

$$\begin{aligned} \left(\widehat{\sigma}^2_{p,q}\right)^* &= \left[\frac{1}{T - (p + q + 1)}\right]^* \\ &\quad \cdot \sum_{[\max(p,q)]^*}^{T} \left[y_t^* - \left(\sum_{j=0}^{P^*} \phi_j y^*_{t-j} - \sum_{i=0}^{Q^*} \theta_i y^*_{t-i}\right)\right]^2. \end{aligned} \tag{19}$$

In essence, bMSE method works as follows: MSCE procedure is applied iteratively on each X_b^* bootstrap replication $b = 1, \dots, B$ of the observed series. A winner model M_g is selected at each iteration on the basis of a given SC, which in turns works exploiting the bootstrap estimated variances of the residuals. The final model is chosen on the basis of its relative frequency over the B bootstrap replication.

3.2. The Applied Bootstrap Scheme. Sieve [31] [32, 33] is the bootstrap scheme employed here. It is an effective and conceptually simple tool to borrow randomness from white noise residuals, generated by the fitting procedure of a "long" autoregression to the observed time series. This autoregression, here supposed to be 0–*mean*, is of the type $y_t = \sum_{j=1}^{p} a_t(y_{t-j}) + \varepsilon_t$, $t \in \mathbb{Z}$, under the stationarity conditions as in (6). Its use is here motivated by the AR(∞) representation of process of type (5); that is,

$$X_t = \sum_{j=1}^{\infty} a_j \left(X_{t-j}\right) + \varepsilon_t \quad t = 1, 2, \dots, T, \tag{20}$$

with $(\varepsilon_t)_{t\in\mathbb{Z}}$ being a sequence of iid variables with $E[\varepsilon_t] = 0$ and $\sum_{j=0}^{\infty} a_j^2 < \infty$. In essence, *sieve* bootstrap approximates a given process by a finite autoregressive process, whose order $\widehat{p} = p(T)$ increases with the sample size T such that $p(T) \to \infty$, $p(T) = o(T)$, $T \to \infty$. In this regard, in the empirical study the estimation of the *p-vector* of coefficients $(\widehat{a}_1, \dots, \widehat{a}_{\widehat{p}})$ has been carried out through the Yule-Walker equations. The residuals $\widehat{\varepsilon}_t = \sum_{j=1}^{\widehat{p}} a_j X_{t-j} + \varepsilon_t$ $t = 1, 2, \dots, T$ obtained from the fitting procedure of this autoregression to the original data are then employed to build up the centered empirical distribution function, which is defined as

$$\widehat{F}_{\varepsilon}(x) = \widehat{P}\left[\varepsilon_t \le x\right] = \left(T - p\right)^{-1} \sum_{t=\widehat{p}+1}^{n} \mathbf{1}_{[S_t - \overline{S} \le x]}, \tag{21}$$

where $S_t = X_t - \sum_{j=1}^{p} \widehat{a}_j X_{t-j}$, with $\overline{S}$ being the mean value of the available residuals, that is, S_t, $t = \widehat{p} + 1, \dots, T$. From $\widehat{F}_\varepsilon$ bootstrap samples $\mathbf{X}_T^* = (X_{1-\widehat{p}}^*, \dots, X_T^*)$ are generated by the recursion

$$\sum_{j=1}^{\widehat{p}} \widehat{a}_j \left(X_{t-j}^* - \overline{X}\right) = \varepsilon_t^* \quad t \in \left(\widehat{p}, \dots, T\right), \tag{22}$$

with starting values $X_t^* = 0$, $\varepsilon_t^* = 0$ for $t \le -\max(p, q)$, $t = T + 1, \dots, 2T$.

3.3. The Proposed bMSE Procedure. Let $\{x_t\}_t^T$ be the observed time series realization of ARIMA (p, q) DGP (5), from which B bootstrap replications $\{x_{b,t}^*;\ b = 1, 2, \dots, B\}_t^T$ are generated via *sieve* method (Section 3). Our B-MSCE procedure is based on the minimization, over all the combinations of ARMA structures, of a given SC by applying MSCE procedure to each bootstrap replication $x_{t,b}^*$ of the original time series x_t.

In what follows the proposed procedure is summarized in a step-by-step fashion.

(1) A maximum ARMA order (P, Q) is arbitrarily chosen, so that exhaustive set Λ of tentative ARMA models, that is, $\{M_g,\ g = 0, 1, \dots, G\}$, with $p \le P$, $q \le Q$, of size $((P = Q) + 1)^2$, is defined.

(2) The number B of bootstrap replications is chosen.

(3) A bootstrap replication, x_b^*, of the original time series x_t is generated via *sieve* method.

(4) The competition set Λ is iteratively fitted to x_b^* so that G values (one for each of the models in Λ) of the SC* are computed and stored in the G-dimension vector v_G.

(5) Minimum SC* value is extracted from v_G so that a winner model, $M_{0,b}^*$, is selected; that is,

$$M_{0,b}^* : \left(\widehat{p}^*, \widehat{q}^*\right) = \arg \min_{p^< P,\, q<Q} \mathrm{SC}^*\left(p, q\right). \tag{23}$$

(6) By repeating B times steps (3) to (5), the final model M_0^* is chosen according to a mode-based criterion, that is, on the basis of its more frequent occurrence in the set of the bootstrap replications. In practice, the selected model is chosen according to the following rule:

$$\#\left[\mathrm{SC}^*\left(M_{0,b}\right) < \mathrm{SC}^*\left(M_{g,b}\right)\right], \qquad g = 0, \dots, G-1,\ b = 1, 2, \dots, B, \tag{24}$$

with the symbol # being used as a counter of the number of the cases satisfying the inequality condition expressed in (24).

The order p_0^{sieve} of the *sieve* autoregression is chosen by iteratively computing the Ljung-Box statistic [34] on the residuals resulting from the fitting of tentative autoregression on the original time series with sample size T_0. Further orders, say p_i^{sieve}, $i = 1, 2, \dots$, for increasing sample sizes, T_i, $i = 1, 2, \dots$, are selected according to the relation $p_i^{sieve} = c(T_i)^{1/3}$, where $c = p_0^{sieve}/T_0^{1/3}$ (in [6] p_0^{sieve} is chosen by iteratively computing the spectral density on the residuals resulting from the fitting of tentative autoregression on the original time series; the order $\widehat{p}$ for which the spectral density is approximately constant is then selected).

The presented method is exhaustive and then highly computer intensive, as for all the $(p + 1) * (q + 1)$ possible pairs (in the attempt to reduce such a burden, sometimes, see, e.g., [35], the set of the ARMA orders under investigation is restricted to $\Lambda = \{(\psi, \psi - 1) : 0 \le \psi \le \Psi\}$; i.e., the competition set is made up of ARMA $(\psi, \psi - 1)$; however, the fact that such an approach entails the obvious drawback of not being able to identify common processes, such as ARMA $(2, 0)$, has appeared to be a too strong limitation; therefore, in spite of its ability to drastically reduce the computational time, such an approach has not been followed here), the values of the given SC* must be computed for each of the B bootstrap replications.

4. Empirical Study

In this section, the outcomes of a simulation study will be reported. It has been designed with the twofold purpose of

Table 1: ARMA DGPs.

Set	Parameters	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6	DGP 7	DGP 8	DGP 9	DGP 10
$\mathscr{J}_1$	ϕ_1	−.85	−.75	−.65	−.60	−.45	.50	.65	.70	.60	.7
	ϕ_2	—	—	—	—	—	—	—	—	—	
	θ_1	.50	.40	−.40	.40	−.65	−.25	−.50	−.40	−.40	−.5
	θ_2	—	—	—	—	—	—	—	—	—	
$\mathscr{J}_2$	ϕ_1	−1.2	−1.1	−0.9	−0.8	−0.6	−.5	− 0.4	0.35	0.7	0.4
	ϕ_2	−.9	−0.9	−0.9	−0.8	−0.5	−4	−0.35	−0.6	−0.8	0.65
	θ_1	.5	.5	0.5	0.5	0.5	.35	0.35	.50	−.5	−.5
	θ_2	—	—	—	—	—	—	—	—	—	—
$\mathscr{J}_3$	ϕ_1	−0.9	−0.8	−0.7	−0.60	−0.5	−0.4	0.4	0.5	0.7	0.80
	ϕ_2	—	—	—	—	—	—	—	—	—	—
	θ_1	−0.50	−0.35	−0.35	0.5	0.	0.35	0.70	−0.9	−0.5	−0.5
	θ_2	0.35	0.50	0.35	0.35	−0.5	0.35	0.25	−0.5	− 0.5	0.35

(i) evaluating bMSE procedure's small sample performances and (ii) giving some evidences of its behavior for increasing sample sizes. As a measure of performances, the percentage frequency of selection of the true order (p^0, q^0), in the sequel denoted as g and g^* for MSCE and bMSE procedure, respectively, has been adopted; that is,

$$g = 100 * \left[\frac{\#\text{ time series correctly identified}}{S} \right], \quad (25)$$

with S denoting the number of the artificial time series employed in the experiment and # the quantifier symbol, expressing the number of times the statement "time series correctly identified" is true. Its extension to the bootstrap case g^* is straightforward.

Aspect (i) consists of a series of Monte Carlo experiments carried out on three different sets of time series, 10 for each set, detailed in Table 1, which (1) are realization of three prespecified ARMA orders, that is, $(1,1)$, $(2,1)$, and $(1,2)$ (one order for each set), and (2) differ from each other, within the same set, only for the coefficients' values, but not for the order (p, q). Two sample sizes will be considered, that is, $T = 100, 200$. Formally, these sets are, respectively, denoted as $\{(\mathscr{J}_1, \mathscr{J}_2, \mathscr{J}_3)\}$ and supposed to belong to the order subspace $\mathscr{I}$: $\{\mathscr{I} \supseteq \mathscr{J}_j;\ j = 1, 2, 3\}$. For each DGP $\in \mathscr{I}$, 10 different coefficient vectors are specified, that is, $\{\mathscr{I} \supseteq \mathscr{J} \equiv (p_1, q_1), \ldots, (p_{10}, q_{10})\}$. The validity of the presented method is assessed on comparative basis, using as benchmark the standard MSCE procedure. For the sake of concision, the values g and g^* will be computed averaging over all the DGPs belonging to either the same set $\mathscr{J}_j$ or $\mathscr{I}$. In practice, two indicators, that is, the Percentage Average Discrepancy (PAD) and the Overall Percentage Average Discrepancy (OPAD), depending on weather only one set $\mathscr{J}_j$ or the whole order subspace $\mathscr{I}_j$ is considered, will be employed. They are formalized as follows:

$$\text{PAD (SC)} = \left[g^*_{[\mathscr{J},\text{SC}]} - g_{[\mathscr{J},\text{SC}]} \right] \quad \forall \mathscr{J} \subseteq \mathscr{I}, \quad (26)$$

$$\text{OPAD (SC)} = \frac{1}{|\mathscr{I}|} \sum_{\mathscr{J}=1}^{|\mathscr{I}|} \left[g^*_{[\mathscr{J},\text{SC}]} - g_{[\mathscr{J},\text{SC}]} \right] \quad \forall \mathscr{J} \subseteq \mathscr{I}, \quad (27)$$

where with the symbol $|\cdot|$ being the cardinality of a set is denoted. In other words, the average percentage differences in the frequency of selection of the true model is used as a measure of the gain/loss generated by bMSE procedure with regard to a single $\mathscr{J}$ (26) or by averaging over the sets $\mathscr{I}$ (27). As already outlined, in analyzing aspect (ii) the attention is focused on the behavior of the proposed method for increasing sample sizes, that is, $T = 100$, 200, 500, 1000. In Table 4, the results obtained for the case of 4 DGPs—detailed in the same table—will be given. In both (i) and (ii), for each DGP $\in (\mathscr{J}_1, \mathscr{J}_2, \mathscr{J}_3)$, a set of $S = 500$ time series has been generated. Each time series s_i $(i = 1, 2, \ldots, S)$ has been artificially replicated $B = 125$ times using the bootstrap scheme outlined in Section 3.2 (the simulations have been implemented using the software R (8.1 version) and performed using the hardware resources of the University of California, San Diego; in particular, the computer server EULER (maintained by the Mathematical Department) and the supercomputer IBM-TERAGRID have been employed). The number of bootstrap replications B employed has been chosen on empirical basis, as the best compromise between performances yielded by the method and computational time.

The parameter space of all the DGPs considered always satisfies the invertibility and stationarity conditions (see (6), (7)), whereas the maximum order P and Q investigated has been kept fixed and low throughout the whole experiment ($P = Q = 3$) mainly to keep the overall computational time reasonably low. However, such an arbitrary choice seems to be able to reflect time series usually encountered in practice in a number of fields, such as economy, ecology, or hydrology. However, it should be emphasized that in many other contexts (e.g., signal processing) higher orders must be considered.

4.1. The Experiments. Other than on the pure ARMA signal, aspect (i) has been investigated in terms of the robustness shown against outliers and noisy conditions. In practice, the simulated DGPs are assumed to be

a: a pure process (no contamination),

b: contaminated with outliers of the type IO (experiments $\mathbf{b_1}, \mathbf{b_2}$) and AO (experiment $\mathbf{b_3}$),

c: contaminated with Gaussian additive noise.

The first set of simulations (experiment **a**) is designed to give empirical evidences for the case of noise-free, uncontaminated ARMA process of type (5). Experiment **b** is aimed at mimicking a situation where a given dynamic system is perturbed by shocks resulting in aberrant data, commonly referred to as outliers. As already pointed out, such abnormal observations might be generated by unpredictable phenomena (e.g., sudden events related to strikes, wars, and exceptional meteorological conditions) or noise components which have the ability to lead to an inappropriate model identification, other than to biased inference, low quality forecast performances, and, if seasonality is present in the data, poor decomposition. Without any doubt, outliers represent a serious issue in time series analysis; therefore testing the degree of robustness of any procedure against such potentially disruptive source of distortion is an important task. This topic has attracted much attention from both theoretical statisticians and practitioners. Detection of time series outliers was first studied by Fox [19], whose results have been extended to ARIMA models by Chang et al. [36]. Other references include [37–39]. In addition, more and more often outlier detection algorithms are provided in the form of stand-alone efficient routines—for example, the library TSO of the software "R," based on the procedure of Chen and Liu (1993) [37]—or included in automatic model identification procedures provided by many software packages, as in the case of the statistical program TRAMO (Time series Regression with ARIMA noise, Missing observations, and Outliers [40]) or SCA (Scientific Computing Associates [41]). Following [19], two common types of outliers, that is, additive (AO) and innovational (IO), will be considered. As it will be illustrated, unfortunately the proposed identification procedure shows sensitivity to outliers, as they are liable, even though to different extents, to noticeable deterioration of the selecting performances.

In more detail, the observed time series x_t is considered as being affected by a certain number ρ of deterministic shocks at different time $t = \tau_1, \ldots, \tau_\rho$; that is,

$$x_t = \sum_{j=1}^{n} h_j \xi_j (B) I_t^{(\tau_j)} + z_t, \qquad (28)$$

where z_t is the uncontaminated one of type (5), h_j measures the impact of the outlier at time $t = \tau_j$, and $I_t^{(\tau_j)}$ is an indicator variable taking the value 1 for $t = \tau_j$ and 0 otherwise. Outlier-induced dynamics are described by the function $\xi(B)$ which takes the form

$$\xi_j (B) = \begin{cases} 1, & \text{for AO} \\ \dfrac{\theta (B)}{\phi (B)} & \text{for IO.} \end{cases} \qquad (29)$$

As the onset of an external cause, outliers of the type IO have the ability to affect the level of the series at the time they occur until a lag τ_j, whose localization depends on the memory mechanism encoded in the ARMA model. Their effect can be even temporally unbounded, for example, under ARIMA DGPs with nonzero integrating stationary inducing constant I. Conversely, AOs affect only the level of the observations at the time of its occurrence (in this regard, typical examples are errors related to the recording process or to the measurement device employed). They are liable to corrupting the spectral representation of a process, which tends to be of the type *white noise* and in general the autocorrelations are pulled towards zero (their effect on the Autocorrelation Function (ACF) and the spectral density level has been discussed in the literature (see, e.g., [42] and the references therein)), so that meaningful conclusion based on these functions—depending on their location, magnitude, and probability of occurrence—might be severely compromised. On the other hand, the effects produced by IOs are usually less dramatic as the ACF tends to maintain the pattern of the uncontaminated process z_t and the spectral density $G_x(\omega)$, ω being the frequency, roughly shows a shape consistent with the one computed on z_t (i.e., $G_x(\omega) \propto G_z(\omega)$). The outcomes of the simulations conducted are consistent with the above.

In the present study, IOs have been randomized and introduced according to a Bernoulli (BER) distribution with parameter $\pi = .04$. In order to better assess the sensitivity of the proposed procedure to outlying observations, experiment **b** has been conducted considering two different levels of standard errors, that is, $\sigma = 3$ (experiment $\mathbf{b_1}$) and $\sigma = 4$ (experiment $\mathbf{b_2}$); in symbols, recalling (5), we have

$$\begin{gathered} \varepsilon_t = (1 - Q_t)\, \varepsilon_{1,t} + Q_t \varepsilon_{2,t} \quad Q_t \sim \text{BER}(\pi) \;\; \pi = 0.04, \\ \varepsilon_{1,t} \sim \text{NID}\left(0, \sigma^2\right) \quad \varepsilon_{2,t} \sim \text{NID}\left(0, k\sigma^2\right) \;\; k = \sqrt{3}, 2. \end{gathered} \qquad (30)$$

In $\mathbf{b_3}$, AOs have been placed according to the following scheme:

$$I_t^{(\tau_j)} = \begin{cases} 1 & j = 25, 35, 65, 75 \quad T = 100 \\ 1, & j = 125, 135, 165, 175 \quad T = 200 \\ 0, & \text{Otherwise.} \end{cases} \qquad (31)$$

The last experiment, that is, **c**, has been designed to mimic a situation characterized by low quality data, induced, for example, by phenomena like changes in survey methodologies (e.g., sampling design or data collecting procedures) or in the imputation techniques. Practically, a Gaussian-type noise ν_t is added to the output signal, so that $x_t = z_t + \nu_t$, z_t being the pure ARMA process. Using (5), we have $x_t = [\phi(B)/\theta(B)]\varepsilon_t + \nu_t$, where $\varepsilon_t \sim \text{nid}(0, \sigma^2)$ and $\nu_t \sim \text{nid}(0, w^2)$ is additive noise, independent of z_t. The variance of ν_t, say w^2, has been chosen according to the relation $w^2 = (1/10)\sigma^2(x_t)$.

4.2. Results. The empirical results pertaining to aspect (i) are summarized in Tables 2 and 3 for the sample sizes $T = 100$ and $T = 200$, respectively. By inspecting these tables it is possible to notice that, with the exception of experiment b_3, in all the other cases bMSE procedure gives no negligible

Table 2: Frequency of selection of the true model in the nonbootstrap (*nb*) and bootstrap (*b*) world for $T = 100$.

Model	Test	AIC		FPE		BIC		HQC	
		g	g^*	g	g^*	g	g^*	g	g^*
ARMA (1, 1) (SET $\mathcal{J}_1$)	a	42.3	52.4	41.5	50.2	52.4	62.7	47.7	60.2
	b_1	37.9	44.7	36.3	42.5	42.1	50.3	44.2	52.7
	b_2	36.3	42.4	32.7	38.1	38.4	45.4	42.0	49.3
	b_3	23.1	20.3	21.9	18.5	29.8	26.7	25.8	24.5
	c	31.2	38.3	26.1	32.5	39.9	48.8	43.1	52.3
ARMA (2, 1) (SET $\mathcal{J}_2$)	a	49.0	59.6	47.1	55.5	59.2	71.3	52.4	65.5
	b_1	47.5	53.4	41.0	46.6	51.8	59.4	53.5	62.7
	b_2	42.0	47.2	39.2	43.6	46.0	53.5	47.6	55.6
	b_3	28.2	26.3	24.2	21.5	27.6	30.2	31.5	33.4
	c	38.3	44.1	34.3	39.4	43.5	51.8	45.9	57.0
ARMA (1, 2) (SET $\mathcal{J}_3$)	a	47.5	57.9	43.6	53.2	55.4	67.0	48.5	61.3
	b_1	32.4	38.4	29.3	34.6	45.3	52.8	48.1	56.3
	b_2	31.6	36.7	28.1	32.4	34.7	41.3	38.5	45.7
	b_3	25.8	23.3	19.6	16.5	24.6	23.7	29.4	28.3
	c	35.1	41.4	32.2	37.6	46.4	54.2	44.7	54.4

Table 3: Frequency of selection of the true model in the nonbootstrap (g^*) and bootstrap (g^*) world for $T = 200$.

Model	Test	AIC		FPE		BIC		HQC	
		g	g^*	g	g^*	g	g^*	g	g^*
ARMA (1, 1) (SET I)	a	49.6	58.7	46.3	54.1	56.7	66.8	53.2	64.4
	b_1	45.5	51.0	39.5	44.7	47.9	54.8	51.9	59.4
	b_2	39.0	43.4	37.1	41.2	45.4	51.2	48.6	54.7
	b_3	29.5	26.4	26.8	23.6	33.4	30.7	30.9	27.4
	c	42.4	48.4	39.5	44.7	48.3	55.6	50.6	58.3
ARMA (2, 1) (SET II)	a	58.7	68.9	53.5	62.3	65.4	77.2	60.4	73.1
	b_1	51.9	57.6	48.1	53.4	55.7	62.8	57.6	66.5
	b_2	49.0	53.3	44.6	48.8	51.5	57.2	53.0	59.8
	b_3	33.8	30.7	22.7	25.3	31.4	33.6	36.3	37.4
	c	50.3	56.4	43.0	48.6	57.2	64.6	58.1	66.3
ARMA (1, 2) (SET III)	a	52.4	62.1	50.5	58.6	60.3	71.3	57.6	69.7
	b_1	48.9	54.2	42.5	47.4	51.1	57.2	54.5	61.7
	b_2	44.2	48.3	35.4	38.6	47.2	52.3	50.8	56.2
	b_3	24.7	22.6	21.0	18.4	26.1	24.2	32.0	29.8
	c	44.4	50.6	36.2	41.3	55.0	61.2	56.3	62.9

improvements over the standard procedure. In particular, it proves to perform particularly well in the pure ARMA case (experiment a) where, for $T = 100$, it brings the frequency of selection of the true model (g^*), averaging over all the three sets of orders, from 45.2% and 52.1% to 54.8% and 64.7% in the case of Shibata efficient and dimension consistent criteria, respectively. Considering the latter, the PAD values recorded are between 11.3 (BIC) and 13.8 (HQC). On the other hand, the bootstrapped version of AIC still shows good improvements (PAD over 10) whereas FPE provides the smaller gains (PAD between 8.4 for $\mathcal{J}_2$ and 9.6 for $\mathcal{J}_3$). As expected, for $T = 200$ both the methods show an increasing average frequency of selection of the correct model for all the SC: averaging over $\mathcal{J}$ and all the SC the values of 55.4% and 65.6% have been recorded for g and g^*, respectively. Regarding the gains over the standard procedure, now BIC and HQC show PAD values above 10 (with a spike of 12.7 of HQC in the case of $\mathcal{J}_2$), whereas the performances for the AIC (PAD above 9) are still good. Less satisfactory job is done by the FPE (PAD = 8.2). Finally, it is worth mentioning that the greatest gains pertain to the HQC, with PAD($\mathcal{J}_1$) = 15.5 for $T = 100$ and PAD($\mathcal{J}_2$) = 12.7 for $T = 200$.

Even though to different extents, both the procedures are affected by the presence of outliers, especially in the case of the smaller sample size. However, as long as IOs (experiments b_1 and b_2) are involved, bMSE seems to do a good job in

Table 4: Frequency of selection of the true model in the nonbootstrap (g) and bootstrap (g^*) world, for different sample sizes.

DGP	Γ	SC	$T = 100$		$T = 200$		$T = 500$		$T = 1000$	
			g	g^*	g	g^*	g	g^*	g	g^*
A	$\phi_1 = -.9$ $\phi_2 = -.9$ $\theta = .6$	AIC	59.2	72.0	71.5	80.5	79.1	84.3	81.8	85.5
		FPE	56.2	65.6	68.5	76.1	75.4	80.2	78.5	82.3
		BIC	65.5	79.7	78.7	89.8	87.3	93.8	92.6	95.5
		HQC	63.3	77.2	76.4	85.3	86.6	91.2	91.5	93.4
B	$\phi_1 = -.8$ $\phi_2 = -.8$ $\theta = .5$	AIC	52.5	61.2	65.7	73.6	74.3	80.3	79.6	82.8
		FPE	51.4	59.1	66.0	71.5	72.9	78.7	75.5	80.4
		BIC	57.3	67.5	72.4	79.7	80.7	84.6	85.2	87.1
		HQC	50.8	63.5	67.4	74.2	76.3	79.4	81.5	83.7
C	$\phi_1 = -.6$ $\phi_2 = -.5$ $\theta = .5$	AIC	45.3	55.7	60.5	69.3	70.2	75.6	73.7	77.2
		FPE	42.5	54.1	57.0	64.9	64.4	70.1	67.6	71.7
		BIC	53.1	65.8	67.3	77.6	76.5	83.1	82.7	84.9
		HQC	50.7	65.4	67.5	75.6	76.0	80.6	81.3	83.1
D	$\phi_1 = -.8$ $\phi_2 = -.6$ $\theta = .4$	AIC	47.8	58.1	61.2	68.4	68.3	72.6	69.7	73.7
		FPE	44.2	53.8	56.5	62.7	64.7	68.6	66.6	70.8
		BIC	51.9	64.2	65.3	71.8	72.5	76.6	74.7	77.5
		HQC	49.7	62.3	66.4	75.0	73.3	79.4	78.0	80.6

counteracting their adverse effects. In fact, for $T = 200$, this procedure, applied to dimension consistent criteria, selects the right model always more than 50% (experiment b_2) and approximately 55% of the times in experiment b_1. For this type of criteria, the average gain over the standard procedure is noticeable, especially in the case of experiment b_1 (OPAD = 6.7 for BIC and 7.9 for HQC). On the other hand, Shibata efficient criteria achieve less remarkable results: with PAD values ranging from 4.9 for the FPE (PAD($\mathcal{J}_3$)) to 5.7 for the AIC (PAD($\mathcal{J}_2$)). As expected, for $T = 100$ the impact of the IOs is stronger: applied to Shibata consistent criteria, bMSE procedure selects the right model in average approximately 43.4% of the time with a minimum of 34.6% recorded for FPE in the case of $\mathcal{J}_2$, whereas dimension consistent criteria show a g^* value in average equal to 55.7%. Selecting performances granted by the proposed method, even though still acceptable, tended to deteriorate to a greater extent considering the experiment b_2, especially with $T = 100$: here the frequency of selection of the true model for Shibata efficient SC is around 40.1% versus 35% of the standard procedure, for a recorded OPAD amounting to 5.5 for the AIC and 4.7 for the FPE. Slightly better results for $T = 200$ are recorded, where the correct model has been identified by dimension consistent criteria 55.2% (OPAD = 5.8%) of the times versus 49.4% of the standard procedure.

Experiment b_3 is where the proposed procedure crashes and offers little or no improvements over the standard one. The most seriously affected selector is the FPE, which shows an ability to select the correct model in average only 18.9% and 22.4% of the times, versus 21% and 25.2% recorded for the nonbootstrap counterpart, respectively, for $T = 100$ and 200. Finally the effect of the injection of a Gaussian noise to the output signal (experiment c) is commented on. Here, the performances of the method appear to be adequate: averaging over $\mathcal{J}$, the value recorded for g^* is 61.8% ($g = 54.6$) for dimension consistent criteria ($T = 200$) with particularly interesting improvements over the standard procedure yielded by HQC, which shows OPAD values amounting to 10% and 7.5% for $T = 100$ and 200, respectively. The bootstrapped version of HQC performs consistently better than the other criteria: in fact it chooses the correct model in average 63.2% and 56.6% of the times for $T = 100$ and $T = 200$, respectively. On the other hand, FPE detects the true model with the smallest probability by reaching the average frequency of selection of the true model of 39.8 ($T = 100$) and 45.1 ($T = 200$). Shibata consistent criteria show also the smallest gains over the standard procedure; for example, for $T = 100$ the maximum PAD is equal to 7.1 and 6.4 for AIC and FPE, respectively (both values' recorder for $\mathcal{J}_2$), whereas dimension consistent criteria, for the same sample sizes, show a maximum PAD of 8.9 and 11.1, in the case of BIC ($\mathcal{J}_1$) and HQC ($\mathcal{J}_2$), respectively.

In the analysis of aspect (ii), the performances yielded by the two procedures, in terms of frequency of selection of the correct model, are considered for increasing sample sizes ($T = 100, 200, 500, 1000$). The results for four different ARMA (2, 1) models, along with their details, are presented in Table 4. As possibly seen by inspecting this table, all the SC under test exhibit roughly a similar pattern: for the small sample size, remarkable disclosures in selecting performances between the two methods are noticeable whereas such discrepancies become less pronounced for $T = 500$ and very small for $T = 1000$. For example, considering all the 4 DGPs, BIC shows a PAD ranging from 12.6 (series D) to 14.7 (series C) with sample size $T = 100$, whereas for $T = 1000$, PAD is in the range 1.9–2.9 for the series B and A, respectively. For this sample size, the smallest PAD has been recorded

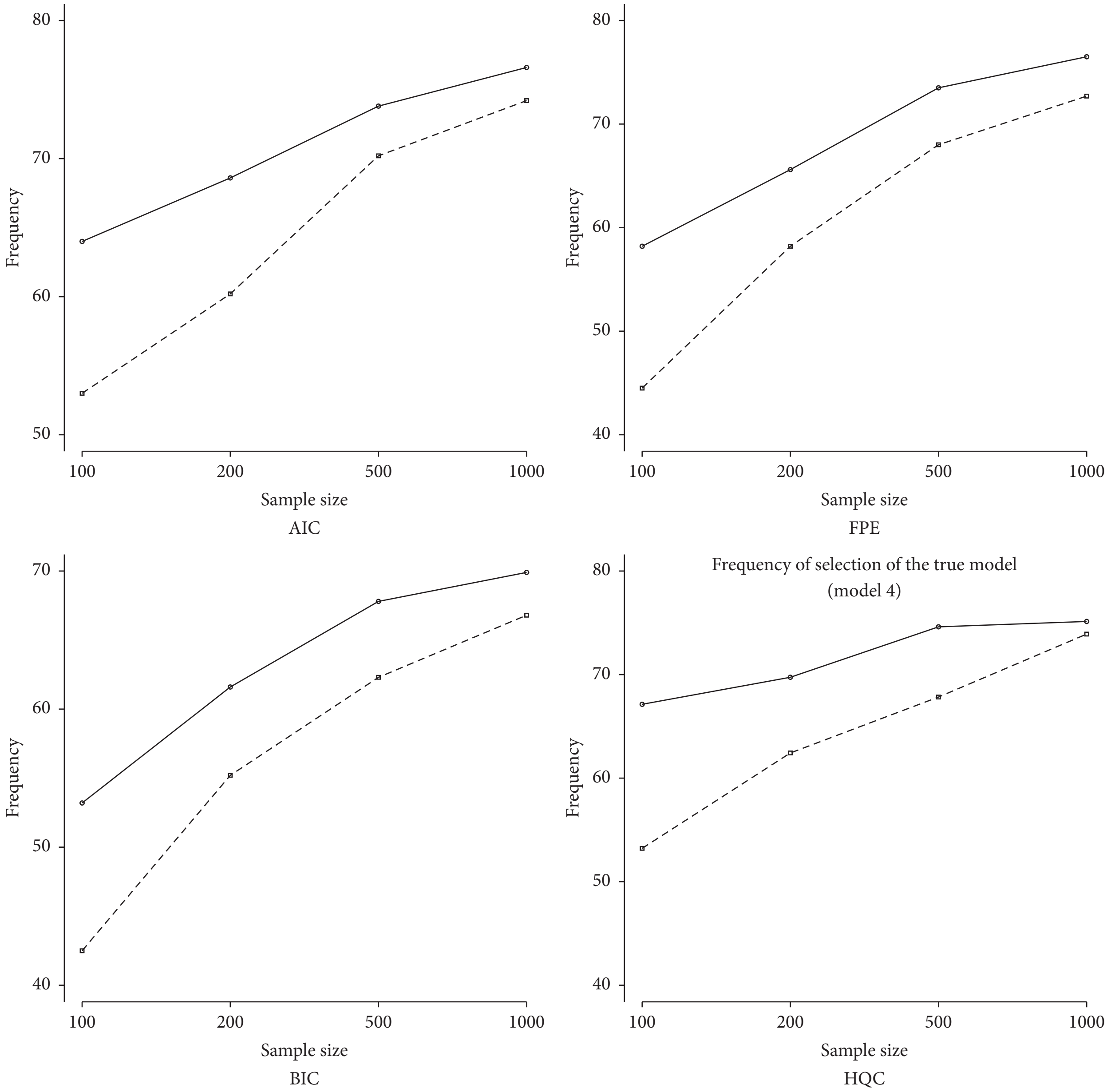


Figure 1: Frequency of selection of the true model, DGP C, Table 4: SC (continuous line) and SC* (dashed line).

for the HQC, whose values range from 1.8 (series C) to 2.6 (series D). Graphical evidence of such a behavior is finally given in Figure 1, where the results recorded for series C are considered.

Among the many factors influencing the performances of the method, the sample size is clearly one of the most crucial ones. Unreported simulations conducted with a sample size $T = 60$ confirm a still acceptable selection rate of the right model, provided that the time series is either noise- and outlier-free and generated by a robust parametric DGP structure. Under such conditions—met, for instance, in the case of DGPs 1-2 (set $\mathcal{J}_2$)—no substantial reductions in the selection performances have been noticed between the smallest sample size considered, that is, $T = 60$ and $T = 100$. For instance, for $T = 60$ and averaging between DGPs 1 and 2 (set $\mathcal{J}_2$), the bootstrap procedure applied to the dimension consistent criteria identifies the right model approx 64% and 61% of the times, versus 49.6% and 42.3% achieved by their nonbootstrap counterpart, for the BIC and HQC, receptively. However, with time series subjected to disturbances and characterized by weaker ARMA structures, as in the case of DGP 6 (set $\mathcal{J}_2$), the proposed method tends to perform poorly and to select the correct model a number of times not far from the standard MAICE procedure.

It is worth pointing out that, unlike artificial setups, in real life data set the true model order is generally unknown; therefore the optimality of the bootstrap method can be inferred on empirical basis, that is, by using the relative frequency of selection of the different tentative ARMA models. In practice, the B winning models generated at each and every bootstrap replication are ranked according to their relative frequency of selection of the true model. In this way, our confidence in the bootstrap selection procedure is linked to the difference in the relative frequency of selection of the winner model (with the highest selection rate) compared to the ones achieved by its closest competitors. Ideal situations are characterized by high rate of choices of the winner model, which drops sharply considering the rest of the competition set. In such a case, we can reasonably be sure that the selected model is closer to the true order than the one found by using the standard MAICE procedure (clearly if different models are selected). On the other hand, slight discrepancies (say 3-4%) between the winning model and the others should be regarded with suspicion and carefully evaluated on a case-by-case basis.

5. Final Remarks and Future Directions

In this paper two pairs of selectors, differing for their derivation and properties, have been brought in a bootstrap framework with the purpose of enhancing their selecting capabilities. A set of Monte Carlo-type experiments has been employed in order to assess the magnitude of the improvements achieved. These encouraging results obtained can be explained in terms of the reduction of uncertainty induced by the bootstrap approach. Identification procedures of the type MSCE, in fact, base the choice of the final model on the minimum value attained by a given SC, no matter how small the differences in the values showed by other competing models might be. When they are actually small, standard MSCE procedures are likely to introduce significant amount of uncertainty in the selection procedure; that is, different order choices can be determined by small variations in the data set. The proposed procedure accounts for such a source of uncertainty, by reestimating the competing models and recomputing the related SC value B times (one for each bootstrap replication). In doing so the identification procedure is based on B different data replications each of them embodying random variations. Also the improvements achieved by the proposed method in the case of IOs can also be explained in the light of reduction of uncertainty. Basically, what the procedure does is to reallocate these outliers B times, so that the related selection procedure can control for such anomalous observations. On the other hand, bMSE procedure breaks down in the case of AOs, probably because of the fact that the employed maximum likelihood estimation procedure is carried out on the residuals, which are severely affected by these types of outliers. Consistently with other Monte Carlo experiments, in the proposed simulations the best results are achieved by dimension consistent criteria, especially by BIC. However, two drawbacks affect this criterion: tendency in the selection of underfitted models and consistency achieved only in case of very large sample [4], under the condition that the true model *is included* in the competition set. The last assumption implies the existence of a model able to provide full explanation of the reality and the “existence” of an analyst able to include it in the competition set. Unfortunately, even assuming finite dimensionality of real life problems, reality is still very complex so that a large number of models are likely to be included in the competition set. As a result of that, selection uncertainty will rise. Superiority of BIC should also be reconsidered in the light of different empirical framework, as Monte Carlo experiment cannot capture the aforementioned problems. It is in fact characterized by the presence of the true model in the portfolio of candidate models. This appears unfair if we consider that criteria of the types AIC and FPE are designed to relax such a strong, in practice unverifiable, assumption and that they enjoy the nice Shibata efficiency property. In addition, in order to keep the computational time acceptable, in Monte Carlo experiments the true DGP is generally of low order, so that BIC underestimation tendency is likely to be masked or, at least, to appear less serious. For these reasons, from a more operational point of view, it can be advisable to consider the indications provided by both AIC* and BIC*, which are the best selectors in their respective categories, according to the simulation experiment. This is particularly true when the sample size is “small” and the information criteria, either considered in their standard or bootstrap form, tend to yield values close to each other for closer models. As a result of that, significant amount of uncertainty can be introduced in the selection process. Finally, as a future direction, it might be worth emphasizing that the purpose of a given model is built and thus identified, which can be usefully considered to assess the selector’s performances. For instance, in many cases computational time is a critical factor, so that one might be willing to accept less accurate model outcomes by reducing the number of bootstrap replications. In fact, global fitting is not necessarily the only interesting feature one wants to look at, as a model might be also evaluated on the basis of the *potential* ability in solving the specific problems it has been built for. In this regard, selection procedures optimized on a case-by-case basis and implemented in the bootstrap world might result in a more efficient tool for a better understanding of the reality.

Conflicts of Interest

The author declares that he has no conflicts of interest.

References

[1] G. E. Box and G. M. Jenkins, *Times Series Analysis. Forecasting and Control*, Holden-Day, San Francisco, Calif, USA, 1970.

[2] M. R. Forster, “Key concepts in model selection: performance and generalizability,” *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 205–231, 2000.

[3] E. E. Leamer, *Specification Searches: Ad Hoc Inference with Experimental Data*, John Wiley & Sons, New York, NY, USA, 1978.

[4] K. P. Burnham and D. Anderson, *Model Selection and Multimodal Inference: A Practical Information-Theoretic Approach*, Springer, New York, NY, USA, 2nd edition, 2002.

[5] H.-Y. Chung, K.-W. Lee, and J.-Y. Koo, “A note on bootstrap model selection criterion,” *Statistics & Probability Letters*, vol. 26, no. 1, pp. 35–41, 1996.

[6] L. Fenga and D. N. Politis, "Bootstrap-based ARMA order selection," *Journal of Statistical Computation and Simulation*, vol. 81, no. 7, pp. 799–814, 2011.

[7] D. A. Freedman, "Bootstrapping regression models," *The Annals of Statistics*, vol. 9, no. 6, pp. 1218–1228, 1981.

[8] J. S. Raho and R. Tibshirani, "Bootstrap model selection via the cost complexity parameter in regression," Tech. Rep., University of Toronto, 1993.

[9] J. Shao, "Bootstrap model selection," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 655–665, 1996.

[10] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[11] H. Akaike, "Statistical predictor identification," *Annals of the Institute of Statistical Mathematics*, vol. 22, pp. 203–217, 1970.

[12] H. Akaike, "Fitting autoregressive models for prediction," *Annals of the Institute of Statistical Mathematics*, vol. 21, pp. 243–247, 1969.

[13] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[14] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[15] E. J. Hannan, "The estimation of the order of an ARMA process," *The Annals of Statistics*, vol. 8, no. 5, pp. 1071–1081, 1980.

[16] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society, Series B*, vol. 41, pp. 190–195, 1979.

[17] K. Torre, D. Delignie, and L. c. Lemoine, "Detection of long-range dependence and estimation of fractal exponents through ARFIMA modelling," *The British Journal of Mathematical and Statistical Psychology*, vol. 60, no. 1, pp. 85–106, 2007.

[18] A. D. McQuarrie and C.-L. Tsai, "Outlier detections in autoregressive models," *Journal of Computational and Graphical Statistics*, vol. 12, no. 2, pp. 450–471, 2003.

[19] A. J. Fox, "Outliers in time series," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 34, pp. 350–363, 1972.

[20] P. Zhang, "Inference after variable selection in linear regression models," *Biometrika*, vol. 79, no. 4, pp. 741–746, 1992.

[21] L. Breiman, "The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error," *Journal of the American Statistical Association*, vol. 87, no. 419, pp. 738–754, 1992.

[22] C. Chatfield, "Model uncertainty, data mining and statistical inference," *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, vol. 158, no. 3, pp. 419–466, 1995.

[23] J. S. Hjorth, *Computer Intensive Statistical Methods—Validation Model Selection and Bootstrap*, Chapman & Hall, London, UK, 1994.

[24] T. Soderstrom and P. Stoica, *System Identification*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[25] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[26] R. Shibata, "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *The Annals of Statistics*, vol. 8, no. 1, pp. 147–164, 1980.

[27] J. Shao, "An asymptotic theory for linear model selection," *Statistica Sinica*, vol. 7, no. 2, pp. 221–264, 1997.

[28] 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.

[29] H. Bozdogan, "Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[30] C. L. Mallows, "Some comments on Cp," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

[31] P. Bühlmann, "Bootstraps for time series," *Statistical Science*, vol. 17, no. 1, pp. 52–72, 2002.

[32] P. Bühlmann, "Sieve bootstrap for time series," *Bernoulli*, vol. 3, no. 2, pp. 123–148, 1997.

[33] J. P. Kreiss, "Bootstrap procedures for AR (∞)—processes," in *Bootstrapping and Related Techniques*, K. H. Jockel, G. Rothe, and W. Sendler, Eds., vol. 376 of *Lecture Notes in Economics and Mathematical Systems*, pp. 107–113, Springer, Berlin, Germany, 1992.

[34] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978.

[35] P. M. T. Broersen, "Automatic spectral analysis with time series models," *IEEE Transactions on Instrumentation and Measurement*, vol. 51, no. 2, pp. 211–216, 2002.

[36] I. Chang, G. C. Tiao, and C. Chen, "Estimation of time series parameters in the presence of outliers," *Technometrics*, vol. 30, no. 2, pp. 193–204, 1988.

[37] C. Chen and L. Liu, "Joint estimation of model parameters and outlier effects in time series," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 284–297, 1993.

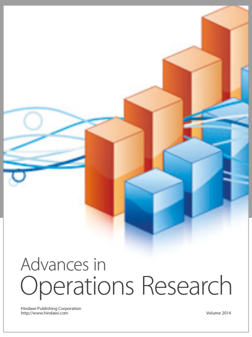
[38] D. Peña, "Influential observations in time series," *Journal of Business and Economic Statistics*, vol. 8, no. 2, pp. 235–241, 1990.

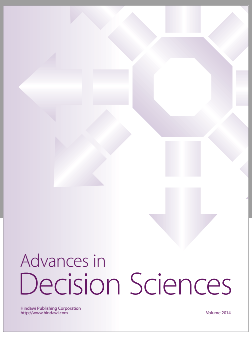
[39] A. G. Bruce and D. Martin, "Leave-k-out diagnostics for time series (with discussion)," *Journal of the Royal Statistical Society, Series B*, vol. 51, no. 3, pp. 363–424, 1989.

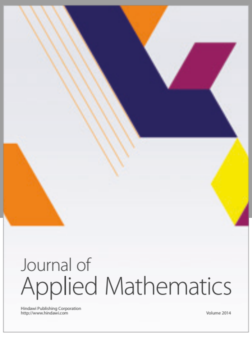
[40] V. c. Gomez and A. n. Maravall, "Estimation, prediction, and interpolation for nonstationary series with the Kalman filter," *Journal of the American Statistical Association*, vol. 89, no. 426, pp. 611–624, 1994.

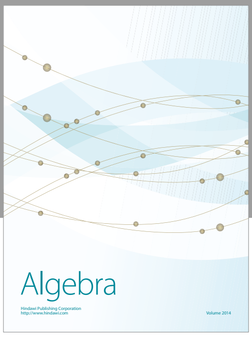
[41] L. M. Liu, G. Hudak, G. E. P. Box, M. E. Muller, and G. C. Tiao, *The SCA Statistical System: Reference Manual for Forecasting and Time Series Analysis*, Scientific Computing Associates, DeKalb, Ill, USA, 1986.

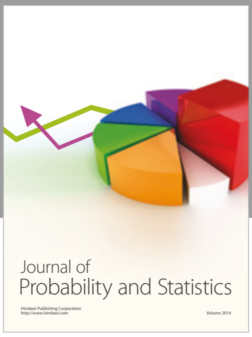
[42] M. A. Carnero, D. Peña, and E. Ruiz, "Outliers and conditional autoregressive heteroscedasticity in time series," *Estadística*, vol. 53, pp. 143–213, 2001.


Advances in Operations Research

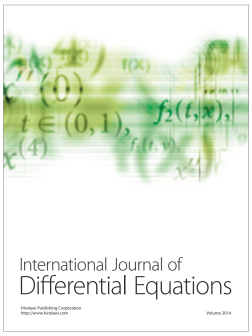

Advances in Decision Sciences

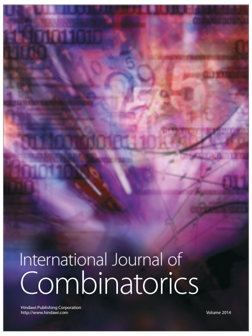

Journal of Applied Mathematics

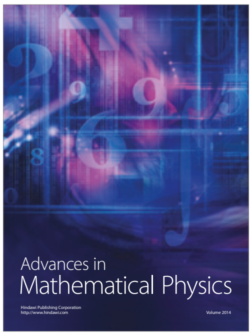

Algebra

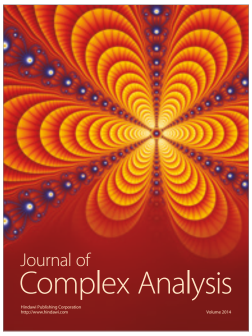

Journal of Probability and Statistics

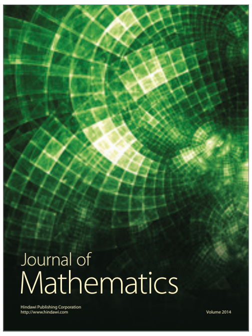
The Scientific World Journal

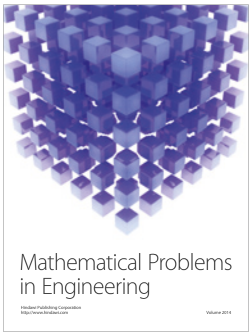

International Journal of Differential Equations

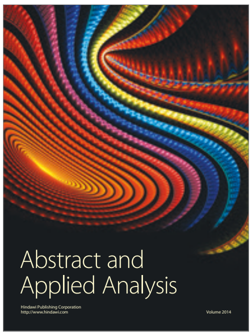

International Journal of Combinatorics

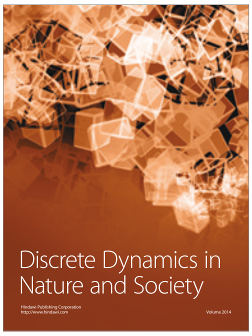

Advances in Mathematical Physics

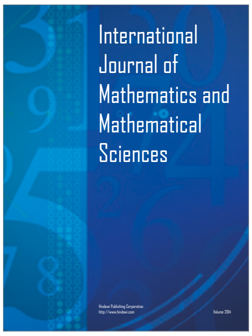

Journal of Complex Analysis


Journal of Mathematics


Mathematical Problems in Engineering

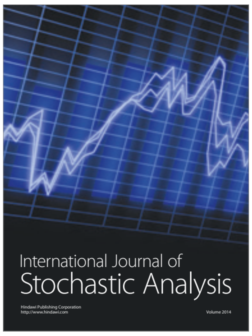

Abstract and Applied Analysis


Discrete Dynamics in Nature and Society


International Journal of Mathematics and Mathematical Sciences

Journal of Discrete Mathematics

Journal of Function Spaces


International Journal of Stochastic Analysis

Journal of Optimization