ARIMA Model Selection and Prediction Intervals

W.A. Dhanushka M. Welagedara, Mulubrhan G. Haile, and David J. Olive*

Department of Mathematics
University of Minnesota Duluth
Duluth, Minnesota 55812
dwelaged@d.umn.edu

Department of Mathematics and Physics
Westminster College
Fulton, Missouri 65251
mule.haile@westminster-mo.edu

School of Mathematical & Statistical Sciences
Southern Illinois University
Carbondale, Illinois 62901-4408
dolive@siu.edu

* Corresponding Author

Keywords AIC, BIC, Data splitting, GMLE.

Mathematics Subject Classification Primary 62M10; Secondary 62M20.

Abstract

Inference after model selection is a very important problem. Model selection algorithms for ARIMA time series, with criterion such as AIC and BIC, tend to select an inconsistent model with positive probability, making data splitting inference unreliable. One technique was fairly reliable for sample sizes greater than 600, and a modification also worked. When consistent estimators are used, the forecast residuals are consistent estimators of the forecast errors. Find a prediction interval for a future forecast error, then shift the interval to be centered at the point estimator of the h-step ahead forecast. A few prediction intervals perform fairly well even after model selection.

1. Introduction

The abstract gives the main results of this paper. This section reviews some time series models, and model selection for ARIMA time series models. We will use the R software

notation and write a moving average parameter θ with a positive sign. Many references and software will write the model with a negative sign for the moving average parameters. A *moving average* MA(q) times series is

$$Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t$$

where $\theta_q \neq 0$. An *autoregressive* AR(p) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

where $\phi_p \neq 0$. An *autoregressive moving average* ARMA(p, q) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t \qquad (1)$$

where $\theta_q \neq 0$ and $\phi_p \neq 0$. A time series X_t follows an ARIMA(p, d, q) model with known d if the differenced time series model Y_t follows an ARMA(p, q) model. See Box and Jenkins (1976) for more on these models. We will assume that the e_t are independent and identically distributed (iid) with zero mean and variance σ^2. The observed time series is $\{Y_t\} = Y_1, ..., Y_n$.

We usually want the ARMA(p, q) model to be weakly stationary, causal, and invertible. Let $Z_t = Y_t - \mu$ where $\mu = E(Y_t)$ if $\{Y_t\}$ is weakly stationary. Then the causal property implies that $Z_t = \sum_{j=1}^{\infty} \psi_j e_{t-j} + e_t$, which is an MA($\infty$) representation, where the $\psi_j \to 0$ rapidly as $j \to \infty$. Invertibility implies that $Z_t = \sum_{j=1}^{\infty} \chi_j Z_{t-j} + e_t$, which is an AR($\infty$) representation, where the $\chi_j \to 0$ rapidly as $j \to \infty$. We will make the usual assumption that the AR(∞) and MA(∞) parameters are square summable. Thus if the ARMA(p, q) model is weakly stationary, causal, and invertible, then Y_t depends almost entirely on nearby lags of Y_t and e_t, not on the distant past.

This paper considers model selection where it is assumed that it is known that the model is ARMA, AR, or MA, but the order needs to be determined. For ARMA model selection, let the full model be an ARMA(p_{max}, q_{max}) model. For AR model selection $q_{max} = 0$, while for MA model selection $p_{max} = 0$. Granger and Newbold (1977, p. 178) suggested using $p_{max} = 13$ for AR model selection, and we may use $p_{max} = q_{max} = 5$ for ARMA model

selection, and $q_{max} = 13$ for MA model selection. For ARMA model selection, there are $J = (p_{max}+1)(q_{max}+1)$ ARMA(p, q) submodels where p ranges from 0 to p_{max} and q ranges from 0 to q_{max}. For AR and MA model selection there are $J = p_{max} + 1$ and $J = q_{max} + 1$ submodels, respectively. Assume the true (optimal) model is an ARMA(p_S, q_S) model with $p_S \leq p_{max}$ and $q_S \leq q_{max}$. Let the selected model I be an ARMA(p_I, q_I) model. Then the model underfits unless $p_I \geq p_S$ and $q_I \geq q_S$.

More notation is needed for model selection. Let the full model be the AR(p_{max}), MA(q_{max}), or ARMA(p_{max}, q_{max}) model. Let $\boldsymbol{\beta}$ be a $b \times 1$ vector. For ARMA model selection, let $\boldsymbol{\beta} = (\boldsymbol{\phi}^T, \boldsymbol{\theta}^T)^T = (\phi_1, ..., \phi_{p_{max}}, \theta_1, ..., \theta_{q_{max}})^T$ with $b = p_{max} + q_{max}$. For AR model selection, let $\boldsymbol{\beta} = (\phi_1, ..., \phi_{p_{max}})^T$ with $b = p_{max}$, and for MA model selection, let $\boldsymbol{\beta} = (\theta_1, ..., \theta_{q_{max}})^T$ with $b = q_{max}$. Hence $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p_{max}}, \beta_{p_{max}+1}, ..., \beta_{p_{max}+q_{max}})^T$. Let $S = \{1, ..., p_S, p_{max} + 1, ..., p_{max} + q_S\}$ index the true ARMA(p_S, q_S) model. If $S = \emptyset$ is the empty set, then the time series random variables $Y_1, ..., Y_n$ are iid. Let $I = \{1, ..., p_I, p_{max} + 1, ..., p_{max} + q_I\}$ index the ARMA(p_I, q_I) model. Let $\hat{\boldsymbol{\beta}}_{I,0}$ be a $b \times 1$ estimator of $\boldsymbol{\beta}$ which is a obtained by padding $\hat{\boldsymbol{\beta}}_I$ with zeroes. If $\boldsymbol{\beta}_I = (\phi_1, ..., \phi_{p_I}, \theta_1, ..., \theta_{q_I})^T$, then $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, ..., \hat{\phi}_{p_I}, 0, .., 0, \hat{\theta}_1, ..., \hat{\theta}_{q_I}, 0, ..., 0)^T$. If $q_I = 0$, then $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, ..., \hat{\phi}_{p_I}, 0, .., 0)^T$. If $p_I = 0$ then $\hat{\boldsymbol{\beta}}_{I,0} = (0, ..., .., 0, \hat{\theta}_1, ..., \hat{\theta}_{q_I}, 0, ..., 0)^T$. If $I = \emptyset$ with $p_I = q_I = 0$, then define $\hat{\boldsymbol{\beta}}_{I,0} = \mathbf{0}$, the $b \times 1$ vector of zeroes. The submodel I underfits unless $S \subseteq I$.

For example, if $p_{max} = q_{max} = 5$, then $S = \{1, 6, 7\}$ corresponds to the ARMA(1,2) model, and $I = \{1, 6, 7, 8\}$ corresponds to the ARMA(1,3) model. Then $\hat{\boldsymbol{\beta}}_S = (\hat{\phi}_1, \hat{\theta}_1, \hat{\theta}_2)^T$, $\hat{\boldsymbol{\beta}}_{S,0} = (\hat{\phi}_1, 0, 0, 0, 0, \hat{\theta}_1, \hat{\theta}_2, 0, 0, 0)^T$, and $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, 0, 0, 0, 0, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, 0, 0)^T$.

The model I_{min} corresponds to the model that minimizes the AIC, AIC_C, or BIC criterion. Then the model selection estimator $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$. Haile and Olive (2023) gave the large sample theory for $\hat{\boldsymbol{\beta}}_{MS}$.

For AR model selection, the probability of underfitting goes to 0 if the Akaike (1973) AIC, Schwartz (1978) BIC, or Hurvich and Tsai (1989) AIC_C criterion are used. See Hannan and Quinn (1979) and Shibata (1976). Although Hannan (1980), Hannan and Kavalieris (1984), and Huang et al. (2022) gave similar results for ARMA models, in simulations, BIC did not

appear to select a consistent model with probability going to one. AIC and AIC_C appear to fail due to the following Theorem 1 for the Gaussian maximum likelihood estimator (GMLE).

Let the r_i be the m (one step ahead) residuals where often $m = n$ or $m = n - p$. Under regularity conditions,

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{m} r_i^2}{m - p - q - c} \tag{2}$$

is a consistent estimator of σ^2 where often $c = 0$ or $c = 1$. See Davis (1977), Granger and Newbold (1977, p. 85), and Huang et al. (2022). Let $\hat{\sigma}^2$ be the estimator of σ^2 produced by the time series model, and let $\gamma_k = Cov(Y_t, Y_{t-k})$. Let

$$\mathbf{\Gamma}_n = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{bmatrix}.$$

The following large sample theorem for the AR(p) model is due to Mann and Wald (1943). Also see McElroy and Politis (2020, p. 333) and Anderson (1971, pp. 210-217). For large sample theory for MA and ARMA models, see Hannan (1973), Kreiss (1985), and Yao and Brockwell (2006).

There is a strong regularity condition for the GMLE for the ARMA model. Assume the ARMA(p_S, q_S) model is the true model. If both $p > p_S$ and $q > q_S$, then the GMLE is not a consistent estimator. See Chan, Ling, and Yau (2020) and Hannan (1980).

Theorem 1 *Let the iid zero mean e_i have variance σ^2, and let the time series have mean $E(Y_t) = \mu$.*

a) Let $Y_1, ..., Y_n$ be a weakly stationary and invertible AR(p) time series, and let $\boldsymbol{\beta} = (\phi_1, ..., \phi_p)$. Let $\hat{\boldsymbol{\beta}}$ be the Yule Walker estimator of $\boldsymbol{\beta}$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}) \tag{3}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \mathbf{\Gamma}_p^{-1}$. Equation (3) also holds under mild regularity conditions for the least squares estimator, and the GMLE of $\boldsymbol{\beta}$.

b) Let $Y_1, ..., Y_n$ be a weakly stationary, causal, and invertible MA(q) time series, and let $\boldsymbol{\beta} = (\theta_1, ..., \theta_q)$. Let $\hat{\boldsymbol{\beta}}$ be the GMLE. Under regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_q(\mathbf{0}, \boldsymbol{V}) \tag{4}$$

where $\boldsymbol{V}$ is given, for example, by McElroy and Politis (2022, pp. 340-341).

c) Let $Y_1, ..., Y_n$ be a weakly stationary, causal, and invertible ARMA(p, q) time series, and let $\boldsymbol{\beta} = (\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$ with $g = p + q$. Let $\hat{\boldsymbol{\beta}}$ be the GMLE. Under regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{V}) \tag{5}$$

where $\boldsymbol{V}$ depends on the autocorrelation function and σ^2.

The main point of Theorem 1 is that the theory can hold even if the e_t are not iid $N(0, \sigma^2)$. The basic idea for the GMLE is that $\{Y_t\}$ satisfies an AR(∞) model which is approximately an AR(p_y) model, and the large sample theory for the AR(p_y) model depends on the zero mean error distribution through σ^2 by Theorem 1a). See Anderson (1971: ch. 5, 1977), Durbin (1959), Hamilton (1994, pp. 117, 429), and Hannan and Rissanen (1982, p. 85). When the e_t are iid $N(0, \sigma^2)$, $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\beta})$, the inverse information matrix. Then for the AR(p) model, $\boldsymbol{V}(\boldsymbol{\phi}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\phi})$. See Box and Jenkins (1976, p. 241) and McElroy and Politis (2020, pp. 340-344).

Pötscher (1990) showed how to estimate $r_S = \max(p_S, q_S)$ consistently. Section 2 reviews this method and gives a modification that can lead to a more parsimonious model. Section 3 illustrates h-step ahead prediction intervals with ARIMA models. Section 4 gives some examples and simulations.

2. Model Selection Algorithms

In the literature and software, the AIC and BIC criteria can take many forms since the criterion can be multiplied by a positive constant, such as $1/n$, and a constant d_n can be added to the criterion without changing the model that minimizes the criterion. Parameters that are in every model, such as σ^2 and possibly a constant, can be absorbed in a constant d_n. For ARMA(p, q) models, let $\log(\hat{L})$ be the log likelihood for the GMLE. Then the

AIC and BIC criteria have the form $-2\log(\hat{L}) + (p+q)c(n)$ where $c(n) = 2$ for AIC and $c(n) = \log(n)$ for BIC. From McElroy and Politis (2020, p. 360) and Huang et al. (2022), $-2\log(\hat{L}) \approx n\log(\hat{\sigma}_I^2) + a_n$ where $\hat{\sigma}_I^2$ is the GMLE of the error variance of model I and a_n is a constant that depends on n. Hence if I is an ARMA(p, q) model, take

$$AIC(I) = n\log(\hat{\sigma}_I^2) + 2(p+q) \text{ and } \mathrm{BIC(I)} = \mathrm{n}\log(\hat{\sigma}_\mathrm{I}^2) + (\mathrm{p}+\mathrm{q})\log(\mathrm{n}). \quad (6)$$

For AIC given by (6), let $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, and models with $4 \leq \Delta(I) \leq 7$ are borderline. See Duong (1984). Claeskens and Hjort (2008, pp. 39, 111) use slightly different formulas for AR(p) models. Pötscher and Srinivasan (1994) multiply the Equation (6) formulas by $1/n$.

In simulations for ARMA model selection, the model selection methods often failed to select a consistent model with high probability in that an inconsistent model was selected with probability > 0.1. A model I is inconsistent due to underfitting if $p_I < p_S$ or $q_I < q_S$. A model I is inconsistent due to overfitting if $p_I > p_S$ and $q_I > q_S$. A model I is consistent if $p_I = p_S$ and $q_I \geq q_S$ or if $q_I = q_S$ and $p_I \geq p_S$. If the model selection procedure was restricted to AR models or MA models, then model I is only inconsistent due to underfitting. The BIC criterion appeared to work for large n if the only models considered were the ARMA(k, k) models for $k = 0, ..., k_{max}$. Then the only consistent model is the ARMA(r_S, r_S) model. For this set of restricted models, AIC and AIC_C tend to overfit with positive probability, and hence do not select a consistent model with probability going to one as $n \to \infty$.

In simulations, the Pötscher (1990) method to estimate $r_S = \max(p_S, q_S)$ often worked rather well. Also see Pötscher and Srinivasan (1994). Chan, Ling, and Yau (2020) suggested that this method is reliable for $n \geq 1000$. In our simulations, the method was fairly reliable for $n \geq 600$, but some models needed much larger n, and there were some models where the method did not simulate well. For the Pötscher (1990) method, let k_{max} be a positive integer such as $p_{max} = q_{max} = k_{max} = 5$. Fit the ARMA$(k, k)$ model for $k = 0, 1, ..., k_{max}$. For each of these $k_{max}+1$ models, compute the BIC–type criterion $z(k) = \log(\hat{\sigma}_k^2) + 2k\log(n)/n$ where $\hat{\sigma}_k^2$ is the GMLE estimator of the error (or innovation) variance σ^2. This criterion is Equation (6) divided by n, and thus (6) could be used instead. The estimator $\hat{r}$ of r_S is the first local

minimum of the series $z(0), z(1), ..., z(k_{max})$. Hence $\hat{r} = 0$ if $z(0) \leq z(1)$; $\hat{r} = 1$ if $z(0) > z(1)$ and $z(1) \leq z(2)$; $\hat{r} = 2$ if $z(0) > z(1)$, $z(1) > z(2)$, and $z(2) \leq z(3)$; $\hat{r} = k$ if $z(r) > z(r+1)$ for $0 \leq r < k$ and $z(k) \leq z(k+1)$ for $k = 0, ..., k_{max} - 1$; and $\hat{r} = k_{max}$ if $z(k)$ is not a local minimum for any $k = 0, 1, ..., k-1$. Note that $r_S \leq k_{max}$ is necessary for $\hat{r}$ to be a consistent estimator of r_S.

The following method is new, and can have fewer parameters than the ARMA$(\hat{r}, \hat{r})$ model. Use the AIC(I) criterion of Equation (6) after finding $\hat{r}$ as above. Then a decrease of AIC > 2 when one parameter is omitted suggests that the parameter was not needed. Let *pen* be a penalty such as $pen = 2$ (used in the simulations) or $pen = 0$. The algorithm computes the $crit = AIC(I) - pen$ for the ARMA$(\hat{r}, \hat{r})$ model, and fits the ARMA$(\hat{r} - i, \hat{r})$ and ARMA$(\hat{r}, \hat{r} - i)$ models for $i = 0, ..., \hat{r} - 1$. If one of the models has AIC(I) $< crit$, then the set crit = AIC(I) $-pen$. This process is repeated at each step. The value of *crit* is updated only if a decrease of more than *pen* from the current value of *crit* is observed. The final model I is the model selected by this algorithm. This additional penalty decreased the amount of underfitting. Note that $2\hat{r}$ models are fitted after finding $\hat{r}$, which fits $k_{max} + 1$ models. This method is faster than computing the AIC for $(k_{max} + 1)^2$ models. Take the ARMA$(p, \hat{r})$ or ARMA$(\hat{r}, q)$ model that has the smallest value of *crit*. Then at least one of p and q will equal $\hat{r}$. Huang et al. (2022) use a similar method with BIC.

3. Prediction Intervals

For forecasting, predict the test data $Y_{n+1}, ..., Y_{n+L}$ given the past training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ prediction interval (PI) for Y_{n+h} is $[L_n, U_n]$ where the coverage $P(L_n \leq Y_{n+h} \leq U_n) = 1 - \delta_n$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. We often want $1 - \delta_n \to 1 - \delta$ as $n \to \infty$. By construction, some of the prediction intervals will have training data coverage $\approx 1 - \alpha_n$ where $1 - \alpha_n \geq 1 - \delta$, and $1 - \alpha_n \to 1 - \delta$ as $n \to \infty$. A large sample $100(1-\delta)\%$ PI is asymptotically optimal if it has the shortest asymptotic length: the length of $[L_n, U_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the population shorth: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

The shorth estimator of the population shorth will be defined below and used to create

large sample PIs that do not require knowing the distribution of the errors e_t. If the data are $Z_1, ..., Z_n$, let $Z_{(1)} \leq \cdots \leq Z_{(n)}$ be the order statistics. Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (e.g., $\lceil 7.7 \rceil = 8$). Consider intervals that contain c cases $[Z_{(1)}, Z_{(c)}], [Z_{(2)}, Z_{(c+1)}], ..., [Z_{(n-c+1)}, Z_{(n)}]$. Compute $Z_{(c)} - Z_{(1)}, Z_{(c+1)} - Z_{(2)}, ..., Z_{(n)} - Z_{(n-c+1)}$. Then the estimator shorth$(c) = [Z_{(s)}, Z_{(s+c-1)}]$ is the interval with the shortest length.

Suppose the data $Z_1, ..., Z_n$ are iid and a large sample $100(1-\delta)\%$ PI is desired for a future value Z_f such that $P(Z_f \in [L_n, U_n]) \to 1 - \delta$ as $n \to \infty$. The shorth(c) interval is a large sample $100(1-\delta)\%$ PI if $c/n \to 1 - \delta$ as $n \to \infty$, that often has the asymptotically shortest length. Frey (2013) showed that for large $n\delta$ and iid data, the shorth$(k_n = \lceil n(1-\delta) \rceil)$ prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the large sample $100(1-\delta)\%$ PI shorth$(c) =$

$$[Z_{(s)}, Z_{(s+c-1)}] \text{ with } \mathrm{c} = \min(\mathrm{n}, \lceil \mathrm{n}[1 - \delta + 1.12\sqrt{\delta/\mathrm{n}}\ ]\ \rceil). \tag{7}$$

Some more notation is needed before deriving PIs for time series. Suppose the training data set is $Y_1, ..., Y_t$. The h-step ahead forecast for a future value Y_{t+h} is $\hat{Y}_t(h)$ and the h-step ahead forecast residual is $\hat{e}_t(h) = Y_{t+h} - \hat{Y}_t(h)$. For example, a common choice for model $Y_t = \tau + \sum_i \psi_i Y_{t-ik_i} + \sum_j \nu_j e_{t-jk_j} + e_t$ is $\hat{Y}_t(h) = \hat{\tau} + \sum_i \hat{\psi}_i Y^*_{t+h-ik_i} + \sum_j \hat{\nu}_j \hat{e}^*_{t+h-jk_j}$ where $\hat{e}_t$ is the tth residual, $Y^*_{t+h-ik_i} = Y_{t+h-ik_i}$ if $h - ik_i \leq 0$, $Y^*_{t+h-ik_i} = \hat{Y}_t(h - ik_i)$ if $h - ik_i > 0$, $\hat{e}^*_{t+h-jk_j} = \hat{e}_{t+h-jk_j}$ if $h - jk_j \leq 0$, and $\hat{e}^*_{t+h-jk_j} = 0$ if $h - jk_j > 0$, and the forecasts $\hat{Y}_t(1), \hat{Y}_t(2), ..., \hat{Y}_t(L)$ are found recursively if there is data $Y_1, ..., Y_t$. Typically the residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ are the 1-step ahead forecast residuals and the fitted or predicted values $\hat{Y}_t = \hat{Y}_{t-1}(1)$ are the 1-step ahead forecasts.

Example 1 is useful to illustrate the forecasts. The R software produces $\hat{e}_t$ and $\hat{Y}_t = Y_t - \hat{e}_t$ for $t = m+1, ..., m+n_1$ where there are n_1 1-step ahead forecast residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ available, often with $m = 0$ and $n_1 = n$. In the examples, we get the formulas $\hat{Y}_n(h)$, and then replace n by t so that the test data formula is applied to the training data. Then the general formula for an ARMA(p, q) model is $\hat{Y}_t(h) = \hat{\tau} + \hat{\phi}_1 \hat{Y}_t(h-1) + \hat{\phi}_2 \hat{Y}_t(h-2) + \cdots + \hat{\phi}_{h-1} \hat{Y}_t(1) + \hat{\phi}_h Y_t + \cdots + \hat{\phi}_p Y_{t+h-p} + \hat{\theta}_h \hat{e}_t + \cdots + \hat{\theta}_q \hat{e}_{t+h-q}$ for $1 < h \leq min(p, q)$. Assume there

are n_h forecast residuals $\hat{e}_t(h)$ available from the training data.

EXAMPLE 1. Suppose the training data is $Y_1, ..., Y_n$. a) Consider an MA(2) model: $Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t$. The R software produces $\hat{e}_t$ and $\hat{Y}_t = Y_t - \hat{e}_t$ for $t = 1, ..., n$ where $\hat{Y}_t = \hat{Y}_{t-1}(1) = \hat{\tau} + \hat{\theta}_1 \hat{e}_{t-1} + \hat{\theta}_2 \hat{e}_{t-2}$ and $\hat{e}_t(1) = Y_{t+1} - \hat{Y}_t(1)$ for $t = 3, ..., n$. Also, $\hat{Y}_n(1) = \hat{\tau} + \hat{\theta}_1 \hat{e}_n + \hat{\theta}_2 \hat{e}_{n-1}$. Hence there are $n_1 = n$ 1-step ahead forecast residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ available. Similarly, $\hat{Y}_t(2) = \hat{\tau} + \hat{\theta}_2 \hat{e}_t$ for $t = 1, ..., n$. Hence the 2-step ahead forecast residuals are available for $t = 3, ..., n-2$. Now $\hat{Y}_t(h) = \hat{\tau} \approx \overline{Y}$ for $h > 2$. Hence there are n h-step ahead forecast residuals $Y_t - \overline{Y}$ for $h > 2$ and $t = 1, ..., n$.

b) Consider an ARMA(1,1) model: $Y_t = \tau + \phi_1 Y_{t-1} + \theta_1 e_{t-1} + e_t$. For $h = 1$, $\hat{Y}_t(1) = \hat{\tau} + \hat{\phi}_1 Y_t + \hat{\theta}_1 \hat{e}_t$. For $h > 1$, $\hat{Y}_t(h) = \hat{\tau} + \hat{\phi}_1 \hat{Y}_t(h-1)$.

c) Consider an AR(1) model: $Y_t = \tau + \phi_1 Y_{t-1} + e_t$. For $h = 1$, $\hat{Y}_t(1) = \hat{\tau} + \hat{\phi}_1 Y_t$. If $\hat{Y}_t(0) = Y_t$, then $\hat{Y}_t(h) = \hat{\tau} + \hat{\phi}_1 \hat{Y}_t(h-1) = \hat{\tau}(1 + \hat{\phi}_1 + \cdots + \hat{\phi}_1^{h-1}) + \hat{\phi}_1^h Y_t = \dfrac{1 - \hat{\phi}_1^h}{1 - \hat{\phi}_1} \hat{\tau} + \hat{\phi}_1^h Y_t$. For a weakly stationary AR(1) time series, a good estimation method will have $|\hat{\phi}_1| < 1$.

d) Consider an ARIMA(1,1,1) model with $\tau = 0$: $Y_t = (1 + \phi_1) Y_{t-1} - \phi_1 Y_{t-2} + \theta_1 e_{t-1} + e_t$. Then $\hat{Y}_t(1) = (1 + \hat{\phi}_1) Y_t - \hat{\phi}_1 Y_{t-1} + \hat{\theta}_1 \hat{e}_t$, $\hat{Y}_t(2) = (1 + \hat{\phi}_1) \hat{Y}_t(1) - \hat{\phi}_1 Y_t$, and $\hat{Y}_t(h) = (1 + \hat{\phi}_1) \hat{Y}_t(h-1) - \hat{\phi}_1 \hat{Y}_t(h-2)$ for $h > 2$.

e) Consider an ARIMA(0,1,1) model with $\tau = 0$: $Y_t = Y_{t-1} + \theta_1 e_{t-1} + e_t$. Then $\hat{Y}_t(1) = Y_t + \hat{\theta}_1 \hat{e}_t$, and $\hat{Y}_t(h) = \hat{Y}_t(h-1) = \hat{Y}_t(1)$ for $h \geq 2$.

f) Consider an ARIMA(0,2,2) model with $\tau = 0$: $Y_t = 2Y_{t-1} - Y_{t-2} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t$. Then $\hat{Y}_t(1) = 2Y_t - Y_{t-1} + \hat{\theta}_1 \hat{e}_t + \hat{\theta}_2 \hat{e}_{t-1}$, $\hat{Y}_t(2) = 2\hat{Y}_t(1) - Y_t + \hat{\theta}_2 \hat{e}_t$, and $\hat{Y}_t(h) = 2\hat{Y}_t(h-1) - \hat{Y}_t(h-2)$ for $h \geq 3$.

The basic idea for getting prediction intervals for the test data is now given. Find the forecast formulas for the test data $Y_{n+1}, ..., Y_{n+L}$, and apply the formulas to the training data $Y_1, ..., Y_n$ to get forecast residuals. Assume consistent estimators are used so that the forecast residuals are consistent estimators of the forecast errors. Apply the shorth to the n_h forecast residuals $\hat{e}_t(h)$ to get $[L_n(h), U_n(h)]$, a PI for a future forecast error. Then the PI for Y_{n+h} is $[\hat{Y}_n(h) + L_n(h), \hat{Y}_n(h) + U_n(h)]$. Since the forecast residuals tend to underestimate the forecast errors, small correction factors are needed for small n. This idea is

illustrated for ARIMA models, but also works for many other time series methods, including seasonal ARIMA models. Similar PIs and prediction regions were derived for multiple linear regression, nonlinear models of the form $Y = m(\boldsymbol{x}) + e$, and multivariate linear regression by Olive (2007, 2013, 2017ab, 2018).

Often time series PIs assume normality, and do not work well unless the errors e_t are iid $N(0, \sigma_e^2)$. For many time series models, a large sample normal $100(1-\delta)\%$ PI for Y_{t+h} is

$$[L_n, U_n] = \hat{Y}_t(h) \pm t_{1-\delta/2, n-p-q} SE(\hat{Y}_t(h)). \tag{8}$$

Suppose that as $n \to \infty$, $\hat{Y}_t(h) \xrightarrow{P} E(Y_{t+h}) = \mu_{t+h}$ and $SE(\hat{Y}_t(h)) \xrightarrow{P} SD(Y_{t+h}) = \sigma_{t+h}$. Thus $\hat{Y}_t(h)$ and $SE(\hat{Y}_t(h))$ are consistent estimators of μ_{t+h} and σ_{t+h}, respectively. These quantities are conditional on the past, but the conditioning is suppressed. Then
$P(Y_{t+h} \in [L_n, U_n]) \approx P(Y_{t+h} \in [\mu_{t+h} - z_{1-\delta/2}\sigma_{t+h}, \mu_{t+h} + z_{1-\delta/2}\sigma_{t+h}]) =$
$P[|Y_{t+h} - \mu_{t+h}| < z_{1-\delta/2}\sigma_{t+h}]$ " $\geq$ " $1 - \frac{1}{z_{1-\delta/2}^2}$ assuming Chebyshev's inequality holds to a good approximation. Hence a 95% PI could have coverage as low as 74% and a 99.7% PI could have coverage as low as 89%. If n is large, a nominal 95% PI uses $t_{1-\delta/2, n-p-q} \approx 1.96$ while using $z_{1-\delta/2} = 5$ has coverage that is eventually bounded below by 96% as $n \to \infty$. The t cutoff 1.96 tends to be too low while the Chebyshev cutoff 5 tends to be too high in that the PI length will be too long and the coverage too high.

The following new PI ignores the time series structure of the data. Let $\overline{e}_t = Y_t - \overline{Y}$, and let shorth$(c_1 = \lceil n(1-\delta) \rceil) = [L_n(h), U_n(h)]$ be computed from the $\overline{e}_t$. Then the large sample $100(1-\delta)\%$ shorth(c_1) PI for Y_{t+h} is

$$[L_n, U_n] = [\overline{Y} + b_n L_n(h), \overline{Y} + b_n U_n(h)] \tag{9}$$

where $b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+1}{n-1}}$. Note that this PI is the same for all h. For weakly stationary, causal, and invertible ARMA(p, q) models, this PI is too long for h near 1, but should have short length for large h and if $h > q$ for an MA(q) model. This PI is the Olive (2013) PI suggested for Y_f when $Y_1, ..., Y_n$ and Y_f are iid.

PI (9) works for two reasons. First, a weakly stationary, causal ARMA(p, q) time series follows an MA(∞) model which is approximately an MA(q_y) time series where q_y depends

on the time series but not on n. Such time series tend to be ergodic: see White (1984, p. 46). For ergodic data from a unimodal distribution, Chen and Shao (1999) proved the sample shorth converges to the unique population shorth. Second, we show that PI (9) works for MA(q) models. Thus PI (9) will also work for MA(∞) time series models. For the MA(q) model, $e_t(h) = \theta_1 e_{t+h-1} + \theta_2 e_{t+h-2} + \cdots + \theta_{h-1} e_{t+1} + e_{t+h}$ for $h \leq q$, $e_t(h) = Y_{t+h} - \mu$ for $h > q$, the $e_t(h)$ are identically distributed for fixed h, and the random variables $e_j(h), e_{j+h}(h), e_{j+2(h)}(h), ...$ are iid for fixed $h \leq q$. For $h \leq q$, there are h iid sequences starting at $j = 1, 2, ..., h$, respectively. For $h > q$ there are $q + 1$ iid sequences starting at $j = 1, ..., (q+1)$. Since the sample percentiles of the iid sequences converge in probability to the population percentiles for fixed h, so do the sample percentiles of all of the data. Hence $P(e_t(h) \in [L_n(h), U_n(h)]) \approx 1 - \delta$ as $n \to \infty$ for the MA(q) model if consistent estimators are used.

The following PI is new and takes into account the time series structure of the data. A similar idea in Masters (1995, p. 305) is to find the n_h h-step ahead forecast residuals and use percentiles to make PIs for Y_{t+h} for $h = 1, ..., L$. Let the full model be the ARMA(k_{max}, k_{max}) model. Let I_m be the ARMA(p_m, q_m) model that was selected by the model selection algorithm. Often $I_m = I_{min}$. Find $\hat{Y}_n(h)$ and the forecast residuals $\hat{e}_t(h)$ for the selected model I_m. For $h = 1$ we will use the residuals $\hat{e}_t$. Let $k = p_m + q_m$, and

$$\tilde{e}_t(h) = \left(1 + \frac{15}{n_h}\right) \sqrt{\frac{n_h}{n_h - k}} \hat{e}_t(h).$$

Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + k/n_h)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta k/n_h), \quad \text{otherwise.}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Then compute the shorth(c_{mod}) PI $[\hat{L}_n(h), \hat{U}_n(h)]$ from the n_h scaled forecast residuals $\tilde{e}_t(h)$ with

$$c_{mod} = \min(n_h, \lceil n_h[q_n + 1.12\sqrt{\delta/n_h}\ ]\ \rceil). \tag{10}$$

Then the new large sample $100(1 - \delta)\%$ PI for Y_{n+h} is

$$[L_n, U_n] = [\hat{Y}_t(h) + \hat{L}_n(h), \hat{Y}_t(h) + \hat{U}_n(h)]. \tag{11}$$

Similar correction factors were used by Olive, Rathnayake, and Haile (2022) for prediction intervals for regression models, such as generalized linear models, after variable selection. Note that for $h = 1$, an estimator for $\sigma^2 = V(e)$ is

$$\hat{\sigma}^2 = \frac{1}{n_1 - k} \sum_{i=1}^{n_1} \hat{e}_i^2 \approx \frac{1}{n_1} \sum_{i=1}^{n_1} e_i^2,$$

suggesting that

$$\sqrt{\frac{n_1}{n_1 - k}} \ \ \hat{e}_i \approx e_i.$$

Why might PIs (11) have good coverage? For both the test data and the training data, $Y_{t+h} = \hat{Y}_t(h) + \hat{e}_t(h) = \mu_{t+h} + e_t(h)$. First, consider the training data where n_h forecast residuals $\hat{e}_t(h)$ exist. Then the proportion of $Y_{t+h} \in [\hat{Y}_t(h) + L_n(h), \hat{Y}_t(h) + U_n(h)] =$ the proportion of the n_h forecast residuals $\hat{e}_t(h) \in [L_n(h), U_n(h)] \approx 1-\delta_n \geq 1-\delta$ by construction. Hence the training data coverage is good. If the selected fitted model is good, and the test data behaves like the training data, then we expect the test data coverage to be good. Hence we need consistent estimators and large n.

Second, assume the time series follow a weakly stationary ARMA model, and suppose $\hat{Y}_t(h)$ is a consistent estimator of μ_{t+h} and $\hat{e}_t(h)$ estimates $e_t(h)$ in that $\hat{e}_t(h) - e_t(h) \xrightarrow{D} 0$ as $n \to \infty$. Also assume that the percentiles of $\hat{e}_t(h)$ estimate the percentiles of $e_t(h)$ such that $P(e_t(h) \in [L_n(h), U_n(h)]) \to 1 - \delta$ as $n \to \infty$. Then $P(Y_{n+h} \in [\hat{Y}_n(h) + L_n(h), \hat{Y}_n(h) + U_n(h)]) \approx P(e_t(h) \in [L_n(h), U_n(h)]) \approx 1-\delta$. These assumptions are roughly the assumptions made when normality is assumed, which makes the time series strictly stationary. For $h = 1$, the $\{\hat{e}_{t+1}\} = \{\hat{e}_t(1)\}$ estimate the iid $\{e_t\}$, and these assumptions may be reasonable if consistent estimators are used and n is large. If the model selection estimator selects a consistent estimator with probability that goes to 1 as $n \to \infty$, then the model selection estimator tends to be consistent by Haile and Olive (2023). For weakly stationary ARMA models, $\mu_{t+h} \to \mu$, $\hat{Y}_t(h) \to \mu$, and $\hat{e}_t(h)$ estimates $Y_{t+h} - \mu$ as $h \to \infty$. Lee and Scholtes (2014) discuss when the percentiles of forecast errors are consistent for ARMA models.

4. Example and Simulations

Model selection can be done using the R function `auto.arima` from the Hyndman and

Khandakar (2008) *forecast package*. Also see Hyndman and Athanasopoulos (2018). The AIC and BIC criteria used by this function differ from those given by Equation (6).

EXAMPLE 2. The monthly Brent crude oil spot price Y_t (dollars per barrel) with 396 observations was collected over the period of 01/1990 - 12/2022. This data set is available from (https://github.com/rishabh89007/Time_Series_Datasets). The differenced time series did not have constant variance. Hence the differenced time series X_t of log(price) was used. The plot of time series in Figure 1 shows several outliers, cases 7, 8, 362, 363, 364, and 365, which create white space in the plot. The outliers near 2020 may be due to covid. These six outliers which were replaced by missing values. Hence if $X_t = log(Y_t) - log(Y_{t-1})$ is the original time series, then W_t is the new time series with $W_t = X_t$ if X_t is not one of the outliers, and $W_t = NA$ if X_t was an outlier, where NA is R notation for missing. See Figure 2 for the plot of W_t. The `auto.arima` function was used for model selection and picked an AR(1) model, which appeared to be reasonable from ACF and PACF plots. The new model selection procedure and the Pötscher (1990) method both selected an ARMA(1,1) model, which is consistent if the AR(1) model is consistent.

For the model selection and PI simulations, there were four error types for the iid e_t: 1) N(0,1), 2) t_5, 3) $U(-1, 1)$, or 4) (EXP(1) - 1), a shifted exponential distribution. All these distributions have mean 0, but the fourth distribution is not symmetric. The 6 time series types were tstype=1 for an AR(1) model with $\phi = 0.5$, tstype=2 for an AR(2) model with $\boldsymbol{\phi} = (0.5, 0.33)^T$, tstype=3 for an MA(1) model with $\theta = -0.5$, tstype=4 for an MA(2) model with $\boldsymbol{\theta} = (-0.5, 0.5)^T$, tstype=5 for an ARMA(3,1) model with $\boldsymbol{\phi} = (0.7, 0.1, -0.4)^T$ and $\theta = 0.1$. Finally, tstype=6 allows the user to specify $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ for an ARMA(p, q) model with $p \geq 1$, $q \geq 1$, and $p, q \leq kmax$ where kmax is the largest value of r for the fitted ARMA(r, r) models, $r = 0, 1, ...,$kmax.

Model Selection Simulations

We used the `auto.arima` function with "AIC", the Pötscher (1990) method that selects an ARMA$(\hat{r}, \hat{r})$ model, and the new ARMA model selection method given in Section 2. In Tables 1–2, these methods are denoted by R AIC, $\hat{r}$, and I, respectively. AIC was used with

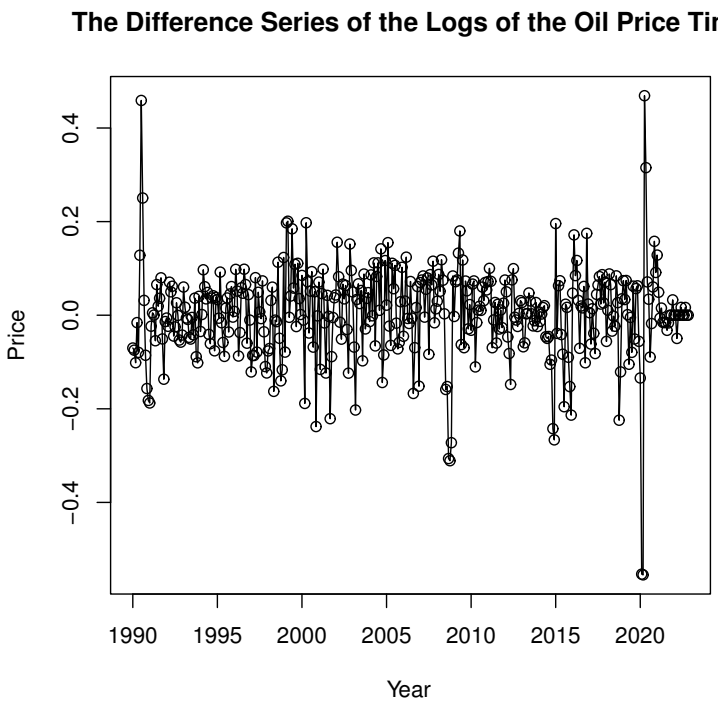

Figure 1: Difference Series of Logs of Oil Price

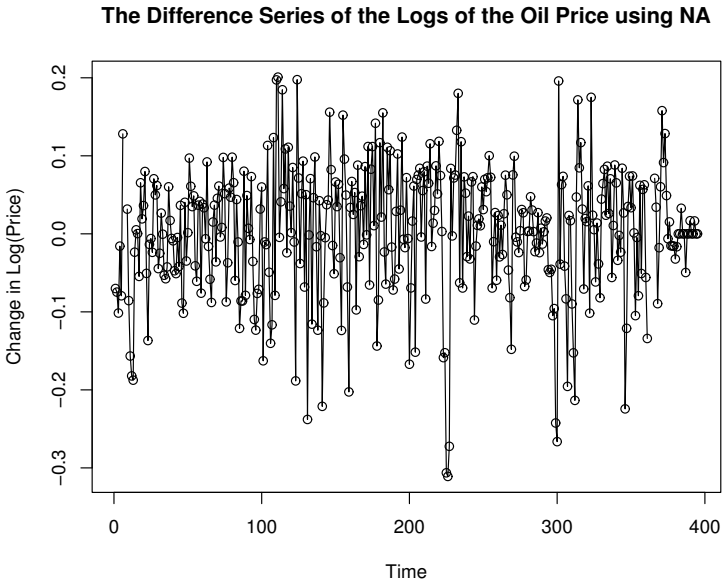

Figure 2: Difference Series of Logs of Oil Price using NA

Table 1: ARMA, Proportion Consistent Model is Selected, nruns=1000,$\phi = 0.4$,$\theta = -0.7$

n	dist	R AIC	$\hat{r}$	I
50	N	0.151	0.561	0.561
200	N	0.239	0.979	0.979
500	N	0.152	1.00	1.00
2000	N	0.249	1.00	1.00
50	t	0.151	0.609	0.609
200	t	0.239	0.944	0.944
500	t	0.109	1.00	1.00
2000	t	0.233	1.00	1.00
50	U	0.153	0.559	0.559
200	U	0.241	0.953	0.953
500	U	0.229	0.982	0.982
2000	U	0.309	1.00	1.00
50	sEXP	0.149	0.569	0.569
200	sEXP	0.241	0.959	0.959
500	sEXP	0.179	0.995	0.995
2000	sEXP	0.263	1.00	1.00

Table 2: Proportion Consistent Model is Selected, nruns=1000, tstype=5

n	dist	R AIC	$\hat{r}$	I
500	N	0.251	0.381	0.364
800	N	0.279	0.692	0.609
1000	N	0.361	0.648	0.619
1500	N	0.357	0.943	0.832
2000	N	0.514	1.00	0.954
500	t	0.219	0.354	0.324
800	t	0.283	0.654	0.584
1000	t	0.354	0.793	0.739
1500	t	0.359	0.939	0.793
2000	t	0.389	0.969	0.904
500	U	0.219	0.409	0.339
800	U	0.229	0.713	0.68
1000	U	0.324	0.839	0.761
1500	U	0.434	0.964	0.879
2000	U	0.524	0.983	0.963
500	sEXP	0.279	0.362	0.301
800	sEXP	0.323	0.619	0.519
1000	sEXP	0.419	0.774	0.663
1500	sEXP	0.369	0.939	0.889
2000	sEXP	0.394	0.994	0.909

`auto.arima` since in the simulations for $n \leq 500$, underfitting was much more of a problem than overfitting. The simulations give the proportion of times a consistent model I was selected. Thus $p_I = p_S$ and $q_I \geq q_S$ or $q_I = q_S$ and $p_I \geq p_S$.

For the 6 time series types, searching all 36 models with `auto.arima` would select a consistent model about 75% to 96% of the time if the true model was AR(p) or MA(q) and n was large (tstype 1 to 4), but did not perform well for ARMA(1,1) models.

The ARMA models were sensitive the values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. For ARMA(1,1) models with $(\phi, \theta)^T = (0.5, 0.2), (0.2, 0.1), (0.4, -0.1), (0.6, -0.3), (-0.2, 0.6), (-0.4, 0.8), (-0.6, 1.0),$ $(0.2, -0.5), (0.4, -0.7), (-0.2, -0.1), (-0.4, 0.1), (-0.6, 0.4), (-0.8, 0.6)$, the Pötscher (1990) method worked well with $n = 1000$, but $(\phi, \theta)^T = (0.5, -0.5)$ did not. In the simulations, the Pötscher (1990) model selection method could work fairly well for n as low as 80, and often worked fairly well for $n = 600$, but often much larger sample sizes were needed. The tstype = 5 model needed $n \geq 1500$. Chan, Ling, and Yau (2020) suggested that the Pötscher (1990) method is reliable for $n \geq 1000$. More simulations are in Welagedara (2023).

For the simulated ARMA(1,1) time series in Table 1, the Pötscher (1990) and new methods were reliable for picking a consistent model for $n \geq 200$, while `auto.arima` with AIC picked an inconsistent model in at least 69% of the 1000 runs. Table 2 used the tstype = 5 model ARMA(1,3) model. The new methods were reliable for $n \geq 1500$, while `auto.arima` with AIC picked an inconsistent model in at least 47% of the 1000 runs.

Prediction Intervals after Model Selection

For ease of programming, we used one step ahead prediction intervals after model selection using the `auto.arima` function, the GMLE, and AIC_C. Haile (2022) gave additional prediction intervals and simulations. With 5000 runs, coverages between 0.94 and 0.96 suggest that there is no reason to believe that the nominal coverage is not 0.95. The iid error distributions for e_t were N(0,1), t_5, $U(-1, 1)$, or (EXP(1) - 1), a shifted exponential distribution. For $h = 1$, the asymptotic optimal lengths of the 95% PIs are 3.92, 5.141, 1.9, and 2.996, while the asymptotic lengths of the normal (Chebyshev) nominal 95% PIs are 3.92σ = 3.92, 5.061, 2.263, and 3.92 for the N(0,1), t_5, $U(-1, 1)$, and (EXP(1) - 1) distributions,

Table 3: One Step Ahead PIs after Model Selection, Coverages and Lengths

n	dist	cov/len	PI (11)	PI (F)	PI (8)
100	N	cov	0.9592	0.9442	0.9476
100		len	4.3214	3.8857	3.9341
100	t5	cov	0.9550	0.9412	0.9434
100		len	5.6747	5.0015	5.0637
100	U	cov	0.9776	0.9842	0.9860
100		len	2.1992	2.2538	2.2819
100	sEXP	cov	0.9540	0.9406	0.9424
100		len	3.7989	3.8504	3.8983
400	N	cov	0.9500	0.9470	0.9476
400		len	3.9990	3.9119	3.9239
400	t5	cov	0.9444	0.9404	0.9412
400		len	5.2364	5.0455	5.0609
400	U	cov	0.9576	0.9988	0.9992
400		len	1.9644	2.2593	2.2662
400	sEXP	cov	0.9578	0.9508	0.9518
400		len	3.2935	3.9047	3.9166
800	N	cov	0.9526	0.9514	0.9520
800		len	3.9445	3.9147	3.9206
800	t5	cov	0.9480	0.9452	0.9456
800		len	5.1604	5.0491	5.0568
800	U	cov	0.9524	0.9994	0.9994
800		len	1.9255	2.2605	2.2640
800	sEXP	cov	0.9438	0.9410	0.9410
800		len	3.1842	3.9147	3.9207

respectively. For iid data, and likely MA(∞) errors, the asymptotic coverages of the nominal 95% Chebyshev intervals for the four error distributions are 0.95, 0.948, 1.00, and 0.948.

Table 3 gives some results for nominal 95% PIs. The full model was the ARMA(5,5) model. The true model was an MA(2) model. PIs (8) and (11) were used, as well as the normal (Chebyshev) nominal 95% PI given `auto.arima`, denoted by (F). Two lines per distribution-sample size combination were given. The first line gives the simulated coverage, which tended to be higher than 0.94. The second line gives the average PI length. PIs (8) and (F) were very similar. For $n = 800$, the PI lengths and coverages were close to the asymptotic values.

Table 4: One Step Ahead PIs after Model Selection, $\phi = 0.4, \theta = -0.7$

n	model selection method	cov/len	PI (9)	PI (11)	PI (F)
100	A	cov	0.9508	0.9510	0.9380
100	A	len	5.6964	5.6948	5.0013
100	P	cov		0.9574	0.9334
100	P	len		5.9887	4.9387
200	A	cov	0.9476	0.9508	0.9426
200	A	len	5.4792	5.3964	5.0300
200	P	cov		0.9554	0.9432
200	P	len		5.5153	5.0007
400	A	cov	0.9478	0.9514	0.9482
400	A	len	5.3971	5.2408	5.0444
400	P	cov		0.9516	0.9490
400	P	len		5.2918	5.0323

In limited simulations, one step ahead prediction intervals (11) and the Chebyshev "95%" h-step ahead prediction intervals (8) simulated fairly well after model selection using `auto.arima` or the Pötscher method. Prediction interval (9) does not depend on model

selection. For example, in Table 4, the true model was an ARMA$(0.4, -0.7)$ model, the e_t were iid t_5, A stands for `auto.arima`, and P stands for the Pötscher method. PI (9) does not depend on the model selection method, and hence the coverage and length were given for A but not for P. For PI (11) and PI (F), the coverages and lengths were similar for A and P. Note that PI (11) was slightly longer than PI (9) for P and $n = 100, 200$. Also, the PI (F) average length is close to the asymptotic length 5.061 while the PI (11) length is not as close to the asymptotically optimal length 5.141.

5. Conclusions

ARMA and ARIMA model selection that searches all $(p_{max} + 1)(q_{max} + 1)$ models was unreliable in simulations. Using BIC from Equation (6), $k_{max} = p_{max} = q_{max}$, and fitting the ARMA$(k, k)$ models for $k = 0, 1, ..., k_{max}$ with the Pötscher method had much better performance, but still needed $n \geq 600$ to be fairly reliable. For data splitting, suppose the first n_H values of the time series are used to select the ARMA(k_I, k_I) model I, and the remaining $n_V = n - n_H$ values for inference. Then $n_H \geq 600$ and $n_V \geq 20k_I$ should be used. Hence the time series length needs to be fairly long, $n \geq 600 + 20k_{max}$, in order to use data splitting inference. Much larger values of n_H and n_V are sometimes needed.

There is a large literature on ARIMA time series PIs, especially for AR(p) models, and the bootstrap is often used. Most of the literature assumes that the model and the order are known, ignoring model selection. Theory needs the model selection estimator to select a consistent estimator with probability going to one. Hence the Pötscher method is better for theory than the standard model selection method that searches all $(p_{max} + 1)(q_{max} + 1)$ models. See Haile (2022), Hyndman and Athanasopoulos (2018), Lu and Wang (2020), and Pan and Politis (2016) for references. See Hong, Kuffner, and Martin (2018) for why classical PIs after AIC variable selection do not work. Hyndman and Athanasopoulos (2018, last paragraph of §8.8) note that ARIMA-based prediction intervals tend to be too narrow, so actual coverage is less than the nominal coverage. See Bhansali (1981) for the effects of estimating the order of the time series model. Data sets where the future data does not behave like the past data are common, and then the prediction intervals tend to perform

poorly.

Plots and simulations were done in *R*. See R Core Team (2020). Programs are in the collection of functions *tspack.txt*. See (http://parker.ad.siu.edu/Olive/tspack.txt). The function `armamsel1` performs Pötscher (1990) ARMA model selection method, and the function `armamsel2` also performs the new ARMA model selection method described in Section 2. The function `armasim3` did the simulation for Tables 1-2.

If $X_t = Y_t - Y_{t-1}$, then $Y_{t+1} = Y_t + X_{t+1}$ which is a random walk where X_{t+1} follows an MA(∞) model. If $X_t = Y_t - 2Y_{t-1} + Y_{t-2}$, then $Y_{t+1} = 2Y_t - Y_{t-1} + X_{t+1}$. Apply PI (9) to the X_t to get $[L_n, U_n]$. Then a nonparametric one step ahead PI for Y_{n+1} is $[Y_n + L_n, Y_n + U_n]$ or $[2Y_n - Y_{n-1} + L_n, 2Y_n - Y_{n-1} + U_n]$. The function `nonpisim` simulates these PIs.

One step ahead prediction intervals (8), (9), and (11) and the Chebyshev "95%" h-step ahead prediction intervals simulated fairly well after model selection using `auto.arima` or the Pötscher method. Huang et al. (2022) show that the variance estimator and the estimator $\hat{\boldsymbol{\beta}}$ are still useful even when the model overfits. PI (9) does not depend on model selection, and h-step ahead PIs should become similar to PI (9) as $h \to \infty$ for MA(∞) time series. The Chebyshev "95%" prediction intervals such as (8) are useful even after model selection, provided that consistent estimators of μ_{t+h} and σ_{t+h} are used, but the asymptotic coverage could be between 0.74 and 1.0, depending on the error distribution.

The function `locpi` gets PI (9). The function `locpi2` needs the forecast residuals, and finds $[L_n, U_n]$ used in PI (11). One step ahead PIs similar to (11) are easy to compute if the one step ahead residuals are given by the model selection output. The function `onesteppi` gets the one step ahead PI for seasonal ARIMA$(p, d, q) \times (P, D, Q)_s$ models with period s where the 6 estimated parameters need to be given. The function can handle missing values entered as NA. For Table 3, the function `pitsvssim` simulates PIs (8) and (11) after model selection using the GMLE with AIC_C using the *R* function `auto.arima`.

For Table 4, the functions `armapisim` and `arimapisim` compare `auto.arima` and the Pötscher method for four 1-step ahead methods. The function `tspisim` compares `auto.arima` and the Pötscher method for the normal Chebyshev h-step ahead PIs. These three function

would often fail if 5000 runs were used.

Acknowledgments

The authors thank the referees and editors.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings, 2nd international symposium on information theory*, ed. B. N. Petrov and F. Csakim, 267-281. Budapest: Akademiai Kiado.

Anderson, T. W. 1971. *The statistical analysis of time series*. Hoboken, NJ: Wiley.

Anderson, T. W. 1977. Estimation for autoregressive moving average models in the time and frequency domains. *The Annals of Statistics* 5 (5):842-865. doi:10.1214/aos/1176343942.

Bhansali, R. J. 1981. Effects of not knowing the order of an autoregressive process on the mean squared error of prediction-I. *Journal of the American Statistical Association* 76 (375):588-597. doi:10.1080/01621459.1981.10477690.

Box, G., and G. M. Jenkins. 1976. *Time series analysis: forecasting and control*. revised ed., Oakland, CA: Holden-Day.

Chan, N. H., S. Ling, and C. Y. Yau. 2020. Lasso-based variable selection of ARMA models. *Statistica Sinica* 30 (4):1925-1948. doi:10.5705/ss.202017.0500.

Chen, M.-H., and Q.-M. Shao. 1999. Monte carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* 8 (1):69-92. doi:10.1080/10618600.1999.10474802.

Claeskens, G., and N. L. Hjort. 2008. *Model selection and model averaging*. New York, NY: Cambridge University Press.

Davis, W. W. 1977. Robust interval estimation of the innovation variance of an Arma model. *The Annals of Statistics* 5 (4):700-708. doi:10.1214/aos/1176343893.

Duong, Q. P. 1984. On the choice of the order of autoregressive models: A ranking and selection approach. *Journal of Time Series Analysis* 5 (3):145-157. doi:10.1111/j.1467-9892.1984.tb00383.x.

Durbin, J. 1959. Efficient estimation of parameters in moving-average models. *Biometrika* 46 (3/4):306-316. doi:10.2307/2333528.

Frey, J. 2013. Data-driven nonparametric prediction intervals. *Journal of Statistical Planning and Inference* 143 (6):1039-1048. doi:10.1016/j.jspi.2013.01.004.

Granger, C. W. J., and P. Newbold. 1977. *Forecasting economic time series.* New York, NY: Academic Press.

Haile, M. G. 2022. Inference for Time Series after Variable Selection. (Ph.D. Thesis), Southern Illinois University, USA, at (http://parker.ad.siu.edu/Olive/shaile.pdf).

Haile, M. G., and D. J. Olive. 2023. Bootstrapping ARMA time series models after model selection. *Communications and Statistics - Theory and Methods* to appear. doi:10.1080/03610926.2023.2280546.

Hamilton, J. D. 1994. *Time series analysis.* Princeton NJ: Princeton University Press.

Hannan, E. J. 1973. The asymptotic theory of linear time-series models. *Journal of Applied Probability* 10 (1):130-145. doi:10.2307/3212501.

Hannan, E. J. 1980. The estimation of the order of an ARMA process. *The Annals of Statistics* 8 (5):1071-1081. doi:10.1214/aos/1176345144.

Hannan, E. J., and L. Kavalieris. 1984. A method for autoregressive-moving average estimation. *Biometrika* 71 (2):273-280. doi:10.1093/biomet/71.2.273.

Hannan, E. J., and B. G. Quinn. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, B* 41 (2):190-195. doi:10.1111/j.2517-6161.1979.tb01072.x.

Hannan, E. J., and J. Rissanen. 1982. Recursive estimation of mixed autoregressive-moving average order. *Biometrika* 69 (1): 81-94. doi:10.1093/biomet/69.1.81.

Hong, L., T. A. Kuffner, and R. Martin. 2018. On overfitting and post-selection uncertainty assessments. *Biometrika* 105 (1):221-224. doi:10.1093/biomet/asx083.

Huang, H. H., N. H. Chan, K. Chen, and C. K. Ing. (2022). Consistent order selection for ARFIMA processes. *The Annals of Statistics* 50 (3):1297-1319. doi:10.1214/21-AOS2149.

Hurvich, C., and C. L. Tsai. 1989. Regression and time series model selection in small

samples. *Biometrika* 76 (2):297-307. doi:10.1093/biomet/76.2.297.

Hyndman, R. J., and G. Athanasopoulos. 2018. *Forecasting: Principles and practice*. 2nd ed., Melbourne, Aus.: OTexts. (https://OTexts.org/fpp2/).

Hyndman, R. J., and Y. Khandakar. 2008. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software* 27 (3):1-22. doi:10.18637/jss.v027.i03.

Kreiss, J. P. 1985. A note on M-estimation in stationary ARMA processes. *Statistics & Risk Modeling* 3 (3-4):317-336. doi:10.1524/strm.1985.3.34.317.

Lee, Y. S., and S. Scholtes, S. 2014. Empirical prediction intervals revisited. *International Journal of Forecasting* 30 (2):217-234. doi:10.1016/j.ijforecast.2013.07.018.

Lu, X., and L. Wang. 2020. Bootstrap prediction interval for ARMA models with unknown orders. *Revstat-Statistical Journal* 18 (3):375-396.

Mann, H. B., and A. Wald 1943. On the statistical treatment of linear stochastic difference equations. *Econometrica* 11 (3/4):173-220. doi:10.2307/1905674.

Masters, T. 1995. *Neural, novel, & hybrid algorithms for time series prediction*. New York, NY: Wiley.

McElroy, T. S., and D. N. Politis. 2020. *Time series: A first course with bootstrap starter*. Boca Raton, FL: CRC Press Taylor & Francis.

Olive, D. J. 2007. Prediction intervals for regression models. *Computational Statistics & Data Analysis* 51 (6):115-3122. doi:0.1016/j.csda.2006.02.006.

Olive, D. J. 2013. Asymptotically optimal regression prediction intervals and prediction regions for multivariate data. *International Journal of Statistics and Probability*, 2 (1):90-100. doi:10.5539/ijsp.v2n1p90.

Olive, D. J. 2017a. *Robust multivariate analysis*. New York, NY: Springer.

Olive, D. J. 2017b. *Linear regression*. New York, NY: Springer.

Olive, D. J. 2018. Applications of hyperellipsoidal prediction regions. *Statistical Papers* 59 (3):913-931. doi:10.1007/s00362-016-0796-1.

Olive, D. J., R. C. Rathnayake, and M. G. Haile. 2022. Prediction intervals for GLMs, GAMs, and some survival regression models. *Communication in Statistics: Theory and*

Methods 51 (22): 8012-8026. doi:10.1080/03610926.2021.1887238.

Pan, L., and D. N. Politis. 2016. Bootstrap prediction intervals for linear, nonlinear, and nonparametric autoregressions. *Journal of Statistical Planning and Inference* 177 1-27. doi:10.1016/j.jspi.2014.10.003.

Pötscher, B. M. 1990. Estimation of autoregressive moving-average order given an infinite number of models and approximation of spectral densities. *Journal of Time Series Analysis* 11 (2):165-179. doi:10.1111/j.1467-9892.1990.tb00049.x.

Pötscher, B. M., and S. Srinivasan. 1994. A comparison of order estimation procedures for ARMA Models. *Statistica Sinica* 4 (1):29-50.

R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. www.R-project.org.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6 (2):461-464. doi:10.1214/aos/1176344136.

Shibata, R. 1976. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* 63 (1):117-126. doi:10.1093/biomet/63.1.117.

Welagedara, W. A. D. M. (2023), Model Selection, Data Splitting for ARMA Time Series, and Visualizing Some Bootstrap Confidence Regions. (Ph.D. Thesis), Southern Illinois University, USA, at (http://parker.ad.siu.edu/Olive/swelagedara.pdf).

White, H. 1984. *Asymptotic theory for econometricians*. San Diego, CA: Academic Press.

Yao, Q., and P. J. Brockwell. 2006. Gaussian maximum likelihood estimation for ARMA models I: Time series. *Journal of Time Series Analysis* 27 (6): 857-875. doi:10.1111/j.1467-9892.2006.00492.x.