Bootstrapping ARMA Time Series Models After Model Selection

Mulubrhan G. Haile and David J. Olive *
Southern Illinois University

August 30, 2022

Abstract

Inference after model selection is a very important problem. This paper derives the asymptotic distribution of some model selection estimators for autoregressive moving average (ARMA) time series models. Under strong regularity conditions, the model selection estimators are asymptotically normal, but generally the asymptotic distribution is a nonnormal mixture distribution. Hence bootstrap confidence regions that can handle this complicated distribution were used for hypothesis testing. A bootstrap technique to eliminate selection bias is to fit the model selection estimator $\hat{\boldsymbol{\beta}}^*_{MS}$ to a bootstrap sample to find a submodel, then draw another bootstrap sample and fit the same submodel to get the bootstrap estimator $\hat{\boldsymbol{\beta}}^*_{MIX}$.

KEY WORDS: ARIMA, confidence region, variable selection.

1. Introduction

This section reviews autoregressive moving average (ARMA) time series models, model selection, and some results on bootstrap confidence regions. 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

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

*Mulubrhan G. Haile is Visiting Assistant Professor, Westminster College, Fulton, MO, and David J. Olive is Professor, School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901-4408 (E-mail: *dolive@siu.edu*).

where $\theta_q \neq 0$ and $\phi_p \neq 0$. The results in this paper also apply to a time series X_t that 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. Also, the time series model $\approx$ AR$(p_y) \approx$ MA(q_y) for some positive integers p_y and q_y that do not depend on the sample size n.

This paper considers ARMA, AR, and MA model selection. 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 $p_{max} = 0$. For nonseasonal time series, 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 $q_I \geq q_S$. 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 (1980) for similar results for ARMA models. Also see Claeskens and Hjort (2008, pp. 39, 40, 45, 46), Hannan and Quinn (1979), and Shibata (1976).

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 $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$. With this notation, the ARMA time series model selection theory developed in this paper is very similar to the variable selection theory for regression models, such as multiple linear regression and generalized linear models, developed by Pelawa Watagoda and Olive (2021ab) and Rathnayake and Olive (2021).

Assume $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$. Let $\hat{\boldsymbol{\beta}}_{MIX}$ be a random vector with a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities equal to π_{kn}. Hence $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with the same probabilities π_{kn} of the model selection estimator $\hat{\boldsymbol{\beta}}_{MS}$, but the I_k are randomly selected. A random vector $\boldsymbol{u}$ has a mixture distribution of random vectors $\boldsymbol{u}_j$ with probabilities π_j if $\boldsymbol{u}$ equals the randomly selected random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$. Let $\boldsymbol{u}$ and $\boldsymbol{u}_j$ be $p \times 1$ random vectors. Then the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of $\boldsymbol{u}_j$. Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)] \quad \text{and}$$

$$\text{Cov}(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j \text{Cov}(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T.$$

If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and

$$\text{Cov}(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j \text{Cov}(\boldsymbol{u}_j).$$

Inference will consider bootstrap hypothesis testing with confidence intervals (CIs) and regions. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. A large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region. Let the $g \times 1$ vector T_n be an estimator of $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $\boldsymbol{A}$ be a full rank $g \times b$ constant matrix. For model selection, test $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{SEL}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL}$ for $i = 1, ..., B$ and SEL is MS or MIX. Let $\lceil x \rceil$ be the smallest integer $\geq x$. For $g = 1$, let the shortest closed interval containing at least c of the T_i^* be the shorth(c) estimator. See Frey (2013). Then the large sample $100(1-\delta)\%$ shorth(c) CI for θ is

$$[T^*_{(s)}, T^*_{(s+c-1)}] \quad \text{with} \quad \text{c} = \min(\text{B}, \lceil \text{B}[1 - \delta + 1.12\sqrt{\delta/\text{n}}\ ]\ \rceil). \tag{2}$$

The shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples.

The confidence regions use Mahalanobis distances D_i and a correction factor to get better coverage when $B \geq 50g$. This result is useful because the bootstrap confidence regions can be slow to simulate and tend to have undercoverage. Let

$$q_B = \min(1-\delta+0.05, 1-\delta+g/B) \text{ for } \delta > 0.1 \text{ and}$$

$$q_B = \min(1-\delta/2, 1-\delta+10\delta g/B), \quad \text{otherwise.} \tag{3}$$

If $1-\delta < 0.999$ and $q_B < 1-\delta+0.001$, set $q_B = 1-\delta$. Let $D_{(U_B)}$ be the $100q_B$th sample percentile of the D_i. Let T be $g \times 1$ and let $\boldsymbol{C}$ be a $g \times g$ symmetric positive definite matrix. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = D^2_{\boldsymbol{z}_i}(T, \boldsymbol{C}) = (\boldsymbol{z}_i - T)^T \boldsymbol{C}^{-1}(\boldsymbol{z}_i - T)$$

for each observation $\boldsymbol{z}_i$. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample.

The Olive (2017ab, 2018) prediction region method (4), modified Bickel and Ren (2001) (5), and Pelawa Watagoda and Olive (2021a) hybrid (6) large sample $100(1-\delta)\%$ confidence regions for $\boldsymbol{\theta}$ are $\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} \tag{4}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$ (if $g = 1$, (4) is a closed interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*), $\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\} =$

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} \tag{5}$$

where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample percentile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$, and $\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\}. \tag{6}$$

Under regularity conditions, Olive (2017b, 2018) proved that (4) is a large sample confidence region. See Bickel and Ren (2001) for (5), while Pelawa Watagoda and Olive (2021a) gave simpler proofs and proved that (2) is a large sample CI. Assume $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$ where $\boldsymbol{u}_n = \sqrt{n}(T_i^* - T_n), \sqrt{n}(T_i^* - \overline{T}^*)$, $\sqrt{n}(T_n - \boldsymbol{\theta})$, or $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta})$, and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{C}$ where $\boldsymbol{C}$ is nonsingular. Let

$$D_1^2 = D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D^2_{\boldsymbol{\theta}}(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_n - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D^2_{T_i^*}(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - T_n)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - T_n).$$

Then $D_j^2 \approx \boldsymbol{u}^T (n\boldsymbol{S}_T^*)^{-1} \boldsymbol{u} \approx \boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$, and the percentiles of D_1^2 and D_4^2 can be used as cutoffs. Confidence regions (4) and (6) have the same volume.

The ratio of the volumes of regions (4) and (5) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}} \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g . \tag{7}$$

The volume of confidence region (5) tends to be greater than that of (4) since the T_i^* are closer to $\overline{T}^*$ than T_n on average.

Section 2 gives large sample theory for $\hat{\boldsymbol{\beta}}_{MIX}$ and $\hat{\boldsymbol{\beta}}_{MS}$. Section 3 shows how to bootstrap these two estimators, and Section 4 gives a simulation.

2. Large Sample Theory for Some Model Selection Estimators

Some notation and preliminary results are needed for the large sample theory. The Gaussian maximum likelihood estimator (GMLE) will be used. The Yule Walker and least squares estimators will also be used for AR(p) models. Let the r_i be the m (one step ahead) residuals where often $m = n$ or $m = n - p$. Under regularity conditions,

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^m r_i^2}{m - p - q - c} \tag{8}$$

is a consistent estimator of σ^2 where often $c = 0$ or $c = 1$. See Granger and Newbold (1977, p. 85) and Pankratz (1983, p. 206). 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

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

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). Pötscher (1990) showed how to estimate $\max(p_S, q_S)$ consistently.

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}) \tag{9}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}$. Equation (9) 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,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_q(\boldsymbol{0}, \boldsymbol{V}). \tag{10}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{\Gamma}_q^{-1}$.

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,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{V}) \tag{11}$$

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), Hannan and Rissanen (1982, p. 85), and Whittle (1953). 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, $\boldsymbol{V}(\boldsymbol{\phi}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\phi})$, while for the MA($q$) model, $\boldsymbol{V}(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{\Gamma}_q^{-1}(\boldsymbol{\theta}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\theta})$. See Box and Jenkins (1976, p. 241) and McElroy and Politis (2020, pp. 340-344).

Next we extend the Pelawa Watagoda and Olive (2021ab) and Rathnayake and Olive (2021) theory for variable selection estimators to time series model selection estimators. Suppose the full model is as in Section 1 and that if $S \subseteq I_j$ where the dimension of I_j is a_j, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j$ is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_b(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \tag{12}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the β_i not indexed by I_j, and $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model.

The first assumption in Theorem 2 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the model selection estimator corresponding to I_{min} underfits with probability going to zero. This assumption follows from Hannan (1980). The assumption also requires $p_S \leq p_{max}$ and $q_S \leq q_{max}$. The assumption on $\boldsymbol{u}_{jn}$ in Theorem 2 is reasonable by (12) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection. The proofs of Theorems 2, 3, and 4 are exactly as in Rathnayake and Olive (2021).

Theorem 2. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_b(\boldsymbol{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{13}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$. Thus $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times b$ full rank matrix with $1 \leq g \leq b$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{14}$$

where $\boldsymbol{v}$ has a mixture distribution of the $\boldsymbol{v}_j = \boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

c) The estimator $\hat{\boldsymbol{\beta}}_{MS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_a = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_b(\boldsymbol{0}, \boldsymbol{V}_{a,0})$ where SEL is MS or MIX.

Proof. a) Since $\boldsymbol{u}_n$ has a mixture distribution of the $\boldsymbol{u}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{u}_n$ is $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ as $n \to \infty$.
b) Since $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$, then $\boldsymbol{A}\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{u}$.
c) The result follows since selecting from a finite number K of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959).
d) If $\pi_a = 1$, there is no selection bias, asymptotically. The result also follows by Pötscher (1991, Lemma 1). $\square$

Theorem 2 can be used to justify prediction intervals after model selection. See Haile and Olive (2022). Typically the mixture distribution is not asymptotically normal unless a $\pi_a = 1$ (e.g. if S is the full model). Theorem 2d) is useful for *model selection consistency* where $\pi_a = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to \infty$. See Hannan (1980) and Claeskens and Hjort (2008) for references.

The following subscript notation is useful. Subscripts before the MIX are used for subsets of $\hat{\boldsymbol{\beta}}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_b)^T$. Let $\hat{\boldsymbol{\beta}}_{i,MIX} = \hat{\beta}_i$. Similarly, if $I = \{i_1, ..., i_a\}$, then $\hat{\boldsymbol{\beta}}_{I,MIX} = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_a})^T$. Subscripts after MIX denote the ith vector from a sample $\hat{\boldsymbol{\beta}}_{MIX,1}, ..., \hat{\boldsymbol{\beta}}_{MIX,B}$. Similar notation is used for other estimators such as $\hat{\boldsymbol{\beta}}_{MS}$. The subscript 0 is still used for zero padding. We may use $FULL$ to denote the full model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$.

The following Pelawa Watagoda and Olive (2021a) theorem is useful for bootstrapping model selection estimators. Let $(\overline{T}, \boldsymbol{S}_T)$ be the sample mean and sample covariance matrix computed from $T_1, ..., T_B$ which have the same distribution as T_n where $T_i = T_{in}$. Let $D^2_{(U_B)}$ be the cutoff computed from the $D^2_i(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. The hyperellipsoids corresponding to $D^2(T_n, \boldsymbol{C})$ and $D^2(\overline{T}, \boldsymbol{C})$ are centered at T_n and $\overline{T}$, respectively. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \leq D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \leq D^2_{(U_B)}$. In Theorem 3, since R_p contains T_f with probability $1 - \delta_B$, the region R_c contains $\overline{T}$ with probability $1 - \delta_B$. Since T_n depends on the sample size n, we need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved, e.g. $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$.

Theorem 3: Geometric Argument. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} \neq \boldsymbol{0}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$ where $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p =$

$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B$ which is eventually bounded below by $1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ where T_n is a randomly selected T_i.

Examining the iid data cloud $T_1, ..., T_B$ and the bootstrap sample data cloud $T_1^*, ..., T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, say, then the bootstrap sample data cloud of $T_1^*, ..., T_B^*$ is like the data cloud of iid $T_1, ..., T_B$ shifted to be centered at T_n. Then the hybrid region (6) is a confidence region by the geometric argument (as is region (5) which tends to use a larger cutoff), and (4) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$.

For $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, we have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (14) where $E(\boldsymbol{v}) = \boldsymbol{0}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. By Theorem 3, if we had iid data $T_1, ..., T_B$, then R_c would be a large sample confidence region for $\boldsymbol{\theta}$. If $\sqrt{n}(T_n^* - T_n) \xrightarrow{D} \boldsymbol{v}$, then we could use the bootstrap sample and confidence regions (4) to (6). This condition holds only under strong regularity conditions such as $\pi_a = 1$. Section 3 will explain why the bootstrap confidence regions are still useful.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{MS}|(\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta})$. Define $P(A|B_k)P(B_k) = 0$ if $P(B_k) = 0$. Let $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_k,0})$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_k,0}) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \leq t_1, ..., z_p \leq t_p)$ by $P(\boldsymbol{z} \leq \boldsymbol{t})$. Then Pötscher (1991) and Pelawa Watagoda and Olive (2021b) show

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta}) \leq \boldsymbol{t}] = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

Hence $\hat{\boldsymbol{\beta}}_{MS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ with probabilities π_{kn}, and $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}.

Note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta})$ are selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$. The random selection for $\hat{\boldsymbol{\beta}}_{MIX}$ does not change the distribution of $\boldsymbol{u}_{jn}$, but selection bias does change the distribution of the selected $\boldsymbol{u}_{jn}$ and $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_{jn}$ and $\boldsymbol{w}_j$. The assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild. The proof for Equation (15) is the same as that for (13).

Theorem 4. *Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then*

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{15}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$. Thus $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

3. Bootstrapping ARMA Time Series Model Selection Estimators

For the bootstrap, we will ignore τ and build the bootstrap time series data set $\{Y_t^*\}$ sequentially. Fit the full model to get the $\hat{\phi}_k$ and $\hat{\theta}_j$. Let

$$Y_t^* = \sum_{k=1}^{p_{max}} \hat{\phi}_k Y_{t-k}^* + e_t^*,$$

$$Y_t^* = \sum_{k=1}^{q_{max}} \hat{\theta}_k e_{t-k}^* + e_t^*,$$

or

$$Y_t^* = \sum_{k=1}^{p_{max}} \hat{\phi}_k Y_{t-k}^* + \sum_{k=1}^{q_{max}} \hat{\theta}_k e_{t-k}^* + e_t^*$$

for $t = 1, ..., n$. The ARMA and AR bootstrap may use a block of initial values $(Y_{-p+1}^*, ..., Y_0^*)^T = (Y_{j+1}, Y_{j+2}, ..., Y_{j+p})^T$ randomly selected from $Y_1, ..., Y_n$. For the *parametric bootstrap*, the e_t^* are iid $N(0, \hat{\sigma}^2)$ where $\hat{\sigma}^2$ is the estimate from fitting the full model with (p_{max}, q_{max}). For the *residual bootstrap*, assume the full model produces m residuals $r_1, ..., r_m$. Often $m = n$ or $m = n - p_{max}$. Refer to Equation (8) with (p, q) replaced by (p_{max}, q_{max}) and $b = p_{max} + q_{max}$. Let

$$\hat{e}_j = \sqrt{\frac{m}{m - b - c}}\,(r_j - \overline{r})$$

for $j = 1, ..., m$. Let the e_t^* be obtained by sampling with replacement from the $\hat{e}_j$. With respect to this bootstrap distribution, the e_t^* are iid with $E(e_t^*) = 0$ and $V(e_t^*) \approx \tilde{\sigma}^2$. Instead of computing the full model, use model selection and zero padding to compute I_k and $\hat{\boldsymbol{\beta}}^*_{MS,1}$. Draw another bootstrap data set and fit model I_k to get $\hat{\boldsymbol{\beta}}^*_{MIX,1}$. Repeat B times to get the bootstrap samples $\hat{\boldsymbol{\beta}}^*_{MS,1}, ..., \hat{\boldsymbol{\beta}}^*_{MS,B}$ and $\hat{\boldsymbol{\beta}}^*_{MIX,1}, ..., \hat{\boldsymbol{\beta}}^*_{MIX,B}$. Let the selection probabilities for the bootstrap model selection estimator be ρ_{kn}. Then this bootstrap procedure bootstraps both $\hat{\boldsymbol{\beta}}_{MS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ with $\pi_{kn} = \rho_{kn}$.

Following McElroy and Politis (2020, pp. 438-439), consider a weakly stationary and invertible time series $Y_1, ..., Y_n$ where the e_t are iid with mean 0 and variance σ^2. A companion process uses ϵ_t that are iid with mean 0 and variance $\hat{\sigma}^2$. Both the residual bootstrap and parametric bootstrap produce companion processes $\{Y_t^*\}$. The residual bootstrap for an AR(p_{max}) model is closely related to the sieve bootstrap for AR(p) and AR(∞) models. See McElroy and Politis (2020, pp. 430, 434).

It is important to note that for the parametric bootstrap, **we are not assuming** that the e_t are iid $N(0, \sigma^2)$. The following theorem is for bootstrapping the full model.

Theorem 5. Assume the time series is such that Theorem 1 holds. Then $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$ if the GMLE is used with the parametric bootstrap. This result also holds for the AR(p) model if the Yule Walker or least squares estimator is used with the parametric bootstrap or the residual bootstrap.

Proof. On a set A of probability going to one as $n \to \infty$, $Y_1^*, ..., Y_n^*$ with $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ satisfies Theorem 1. Hence if n is fixed and the time series $Y_1^*, ..., Y_m^*$ is generated with $\hat{\boldsymbol{\beta}}_n$, then on the set A the estimator $\hat{\boldsymbol{\beta}}^*$ satisfies $\sqrt{m}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_n) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}}_n))$ as $m \to \infty$.

Since $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ if $\hat{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}$ as $n \to \infty$, it follows that $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_n) \xrightarrow{D} N_b(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$ as $n \to \infty$. $\square$

The basic idea is that for the parametric bootstrap, $Y_1^*, ..., Y_n^*$ satisfies the Gaussian time series model with $\hat{\boldsymbol{\beta}}_n$ as the parameter vector and $\hat{\boldsymbol{\beta}}_n$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$. Hence the Gaussian time series $Y_1^*, ..., Y_n^*$ with $\hat{\boldsymbol{\beta}}_n$ will be weakly stationary, causal, and invertible on a set A going to one in probability. Since $\hat{\boldsymbol{\beta}}_n$ depends on n, convergence along a triangular array needs to be used. Bootstrap results such as Theorem 5 are rather rare in the time series literature. Bühlmann (1994) has such a result for the AR(p) model.

If Equation (12) holds so $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_b(\boldsymbol{0}, \boldsymbol{V}_{j,0})$, we would like to show that $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j,0} - \hat{\boldsymbol{\beta}}_{I_j,0}) \xrightarrow{D} N_b(\boldsymbol{0}, \boldsymbol{V}_{j,0})$ if I_j was selected with random selection. This result holds for the full model by Theorem 5. Suppose $S \subseteq I_j$. Then the bootstrap data set $\{Y_t^*\}$ satisfies

$$Y_t^* = \sum_{k=1}^{p_{I_j}} \hat{\phi}_k Y^*_{t-k} + e_t^* + e_t^*(I_j),$$

$$Y_t^* = \sum_{k=1}^{q_{I_j}} \hat{\theta}_k e^*_{t-k} + e_t^* + e_t^*(I_j),$$

or

$$Y_t^* = \sum_{k=1}^{p_{I_j}} \hat{\phi}_k Y^*_{t-k} + \sum_{k=1}^{q_{I_j}} \hat{\theta}_k e^*_{t-k} + e_t^* + e_t^*(I_j)$$

where $e_t^*(I_j) = \sum_{k=p_{I_j}+1}^{p_{max}} \hat{\phi}_k Y^*_{t-k}$ for the AR(p_{max}) model, $e_t^*(I_j) = \sum_{k=q_{I_j}+1}^{q_{max}} \hat{\theta}_k e^*_{t-k}$ for the MA(q_{max}) model, and $e_t^*(I_j) = \sum_{k=p_{I_j}+1}^{p_{max}} \hat{\phi}_k Y^*_{t-k} + \sum_{k=q_{I_j}+1}^{q_{max}} \hat{\theta}_k e^*_{t-k}$ for the ARMA(p_{max}, q_{max}) model. When $S \subseteq I_j$, the $e_t^*(I_j) \xrightarrow{P} 0$ rapidly as $n \to \infty$. For the MA model with the parametric bootstrap, $e_t^*(I_j) \sim N(0, \hat{\sigma}^2 \sum_{k=q_{I_j}+1}^{q_{max}} \hat{\theta}_k^2)$ which has a variance proportional to $1/n$ if $S \subseteq I_j$. We could also modify $\hat{\boldsymbol{\beta}}^*_{MIX}$ to omit the $e_t^*(I_j)$ resulting in a new bootstrap estimator $\hat{\boldsymbol{\beta}}^*_{MX}$.

The key idea is to show that the bootstrap data cloud is slightly more variable than the iid data cloud, so confidence region (5) applied to the bootstrap data cloud has coverage bounded below by $(1-\delta)$ for large enough n and B. Let B_{jn} count the number of times $T_i^* = T_{ij}^*$ in the bootstrap sample. Then the bootstrap sample $T_1^*, ..., T_B^*$ can be written as

$$T^*_{1,1}, ..., T^*_{B_{1n},1}, ..., T^*_{1,J}, ..., T^*_{B_{Jn},J}.$$

Denote $T^*_{1j}, ..., T^*_{B_{jn},j}$ as the jth bootstrap component of the bootstrap sample with sample mean $\overline{T}^*_j$ and sample covariance matrix $\boldsymbol{S}^*_{T,j}$. Similarly, we can define the jth component of the iid sample $T_1, ..., T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$.

Let $T_n = \hat{\boldsymbol{\beta}}_{MIX}$ and $T_{ij} = \hat{\boldsymbol{\beta}}_{I_j,0}$. If $S \subseteq I_j$, assume $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$. Then by Equation (12),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \text{ and } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\beta}}^*_{\mathrm{I_j,0}} - \hat{\boldsymbol{\beta}}_{\mathrm{I_j,0}}) \xrightarrow{\mathrm{D}} \mathrm{N_b}(\boldsymbol{0}, \boldsymbol{V}_{\mathrm{j,0}}). \qquad (16)$$

This result means that the component clouds have the same variability asymptotically. The iid data component clouds are all centered at $\boldsymbol{\beta}$. If the bootstrap data component clouds were all centered at the same value $\tilde{\boldsymbol{\beta}}$, then the bootstrap cloud would be like an iid data cloud shifted to be centered at $\tilde{\boldsymbol{\beta}}$, and (5) would be a confidence region for $\boldsymbol{\theta} = \boldsymbol{\beta}$. Instead, the bootstrap data component clouds are shifted slightly from a common center, and are each centered at a $\hat{\boldsymbol{\beta}}_{I_j,0}$. Geometrically, the shifting of the bootstrap component data clouds makes the bootstrap data cloud similar but more variable than the iid data cloud asymptotically (we want $n \geq 20b$), and centering the bootstrap data cloud at T_n results in the confidence region (5) having slightly higher asymptotic coverage than applying (5) to the iid data cloud. Also, (5) tends to have higher coverage than (6) since the cutoff for (5) tends to be larger than the cutoff for (6). Region (4) has the same volume as region (6), but tends to have higher coverage since empirically, the bagging estimator $\overline{T}^*$ tends to estimate $\boldsymbol{\theta}$ at least as well as T_n for a mixture distribution. A similar argument holds if $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$, $T_{ij} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$, and $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$.

In the simulations of Section 4 for $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{\beta}_S = \boldsymbol{\theta}_0$ with $n \geq 20b$, the coverage tended to get close to $1 - \delta$ for $B \geq \max(200, 50b)$ so that $\boldsymbol{S}^*_T$ is a good estimator of $\text{Cov}(T^*)$.

The matrix $\boldsymbol{S}^*_T$ can be singular due to one or more columns of zeros in the bootstrap sample for $\beta_1, ..., \beta_b$. The β_j corresponding to these columns are likely not needed in the model given that the other predictors are in the model. A simple remedy is to add k bootstrap samples of the full model estimator $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}^*_{FULL}$ to the bootstrap sample. For example, take $k = \lceil cB \rceil$ with $c = 0.01$. Getting a good full model for the ARMA model can be difficult. A confidence interval $[L_n, U_n]$ can be computed without $\boldsymbol{S}^*_T$ for (4), (5), and (6). Using the confidence interval $[\max(L_n, T^*_{(1)}), \min(U_n, T^*_{(B)})]$ can give a shorter covering region.

Undercoverage can occur if bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-b)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

4. Example and Simulations

Example 1. The WWWusage data set, available from the *R* software, is from Durbin and Koopman (2001). This data set is a time series of the numbers of users connected to the Internet through a server every minute, and has $n = 100$. First differences were taken. The ARIMA(3,1,0) and ARIMA(1,1,1) models were interesting. The second model minimized the BIC criterion for ARIMA(p,1,q) models with $0 \leq p, q, \leq 5$, and BIC(model 1) - BIC(model 2) = 0.29. Let Y be the differenced time series of length $n = 99$, and consider model selection with AR(p_{max}=10) since $n = 99$ is small. The parameter ϕ_3 was not statistically significant at the 0.05 level using the full AR(10) model output, but was significant at the 0.1 level. The bootstrap often selected the AR(1), AR(2), and AR(3) models, but rarely selected the AR(8), AR(9), or AR(10) models. The parametric and residual bootstrap were both used with model selection and MIX to get 90% confidence intervals for ϕ_j for j = 1,...,10. The confidence intervals for the full AR(10) model did not use the bootstrap. Table 1 shows that the bootstrap confidence intervals were often

much shorter than those of the full model. We also used the bootstrap confidence regions to test $\boldsymbol{\phi}_I = (\phi_k, ..., \phi_{10})^T = \mathbf{0}$ for $k = 1, ..., 10$.
$D_{\mathbf{0}}$: 12.100,2.284,1.575,...,0.227,0.224,0.084
D_{U_B}: 5.223,5.094,4.952,...,0.227,0.224,0.084.
Shown above are the test statistics $D_{\mathbf{0}}$ and the cutoffs D_{U_B} where $H_0 : \boldsymbol{\phi}_I = \mathbf{0}$ is rejected if $D_{\mathbf{0}} > D_{U_B}$. For $k = 8, 9, 10$, $D_{\mathbf{0}} = D_{U_B}$, which means that the 90% prediction region method confidence region only contained $\boldsymbol{\phi}_I^* = \mathbf{0}$ where $\boldsymbol{\beta}_I = \boldsymbol{\phi}_I$. The full AR(10) model was fit for 10 bootstrap samples with $c = 0.01$, and $B = 1000$ bootstrap samples used model selection with the parametric bootstrap. See the second to last paragraph in Section 3.

Table 1: 90% Bootstrap Confidence Intervals for ϕ_j,B=1000,c=0.01,pmax=10

j	par boot	par mix	res boot	res mix	full
1	[0.767,1.163]	[0.752,1.167]	[0.764,1.174]	[0.741,1.164]	[0.987,1.322]
2	[−0.772,−0.140]	[−0.646,0]	[−0.640,0]	[−0.637,0]	[−0.907,−0.397]
3	[−0.004,0.411]	[−0.0003,0.354]	[−0.004,0.408]	[0,0.362]	[0.069,0.627]
4	[−0.175,0.220]	[−0.083,0.160]	[−0.153,0.217]	[−0.081,0.154]	[−0.247,0.328]
5	[−0.152,0.165]	[−0.117,0.066]	[−0.165,0.173]	[−0.118,0.066]	[−0.404,0.180]
6	[−0.243,0.033]	[−0.103,0.034]	[−0.278,0]	[−0.116,0.037]	[−0.140,0.452]
7	[−0.203,0.003]	[−0.107,0.005]	[−0.202,0]	[−0.065,0.003]	[−0.505, 0.089]
8	[0,0]	[0,0]	[0,0]	[0,0]	[−0.314,0.273]
9	[0,0]	[0,0]	[0,0]	[0,0]	[−0.138,0.402]
10	[0,0]	[0,0]	[0,0]	[0,0]	[−0.202,0.145]

We simulated AR model selection with the Yule Walker estimators and AIC. For MA and ARMA model selection, the GMLE with AIC_C was used. Let $b = p_{max} + q_{max}$. We recommend $n \geq 10b$ and $B \geq 20b$. We used 5000 runs. Then coverage within [0.94,0.96] suggests that the true coverage is near the nominal coverage 0.95. Often B, p_{max}, and q_{max} were rather small to make the simulation time shorter. More simulations are in Haile (2022).

For the AR simulations, the true model was an AR(1) model with $p_S = 1$ and $\phi_1 = 0.5$, or an AR(2) model with $p_S = 2$ and $\boldsymbol{\phi} = (0.5, 0.33)^T$ corresponding to tstype = 1 or 2. For the MA simulations, the true model was an MA(1) model with $p_S = 1$ and $\theta_1 = -0.5$, or an MA(2) model with $p_S = 2$ and $\boldsymbol{\theta} = (-0.5, 0.5)^T$ corresponding to tstype = 1 or 2. Error types were N(0,1), t_5, uniform(−1,1), and $e \sim W - 1$ where $W \sim$ exponential(1) corresponding to etype = 1, 2, 3, or 4. The parametric bootstrap and residual bootstrap were used corresponding to btype =1 or 2. Nominal 95% confidence regions and intervals were used with 1% augmentation from the bootstrapped full model. The simulations bootstrapped the full model estimator, the model selection estimator $\hat{\boldsymbol{\beta}}_{MS}$, and $\hat{\boldsymbol{\beta}}_{MIX}$.

The tables give two rows for each of the three estimators giving the observed CI coverage and average CI length. For the tests, the length gives the average cutoff $D_{(U_B)}$.

Table 2: AR(p) MS, residual bootstrap,n=400,$\phi = (0.5, 0.33)$,B=200,pmax=5

e	ϕ_1	ϕ_2	$\phi_{p_{max}-1}$	$\phi_{p_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.959	0.950	0.964	0.962	0.955	0.955	0.966	0.940	0.953	0.965
len	0.215	0.234	0.230	0.204	2.851	2.851	3.465	2.475	2.475	2.648
N,MS	0.959	0.955	1−	1−	0.998	0.994	0.999	0.954	0.962	0.968
len	0.215	0.233	0.201	0.152	3.545	3.545	3.827	2.493	2.493	2.648
N,MIX	0.962	0.947	1.000	1−	0.999	0.997	0.999	0.954	0.959	0.970
len	0.213	0.224	0.161	0.111	3.970	3.970	4.212	2.509	2.509	2.655
t,full	0.950	0.948	0.959	0.968	0.957	0.953	0.966	0.944	0.958	0.968
len	0.215	0.234	0.229	0.204	2.857	2.857	3.463	2.479	2.479	2.657
t,MS	0.955	0.952	0.999	1.000	0.997	0.992	0.998	0.955	0.963	0.973
len	0.215	0.233	0.201	0.152	3.560	3.560	3.837	2.499	2.499	2.656
t,MIX	0.955	0.946	0.999	1.000	0.998	0.996	0.998	0.954	0.961	0.969
len	0.213	0.223	0.160	0.108	3.992	3.992	4.231	2.511	2.511	2.653
U,full	0.953	0.949	0.962	0.963	0.950	0.952	0.963	0.937	0.954	0.963
len	0.215	0.235	0.230	0.204	2.854	2.854	3.460	2.477	2.477	2.641
U,MS	0.956	0.951	1−	1.000	0.997	0.991	0.997	0.952	0.959	0.966
len	0.215	0.234	0.202	0.152	3.539	3.539	3.820	2.495	2.495	2.611
U,MIX	0.959	0.943	1−	1.000	0.999	0.995	0.998	0.952	0.958	0.967
len	0.213	0.224	0.162	0.111	3.968	3.968	4.208	2.508	2.507	2.650
E,full	0.956	0.954	0.958	0.962	0.953	0.953	0.965	0.944	0.954	0.966
len	0.215	0.236	0.229	0.203	2.865	2.865	3.466	2.485	2.485	2.661
E,MS	0.961	0.957	1−	1−	0.998	0.993	0.998	0.954	0.961	0.973
len	0.215	0.235	0.201	0.150	3.563	3.563	3.841	2.501	2.501	2.653
E,MIX	0.956	0.950	1−	1−	0.999	0.996	0.999	0.951	0.959	0.968
len	0.213	0.225	0.160	0.107	3.984	3.984	4.220	2.519	2.519	2.663

The term "full" is for the full model, the term "MS" is for model selection, and the term "MIX" for random selection. The terms pr, hyb and br are for the prediction region method, hybrid region, and Bickel and Ren region. The 0 indicates that the test was $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ where $\boldsymbol{\beta}_E = (\beta_{p_S+1}, ..., \beta_k)^T$. The 1 indicates the test $H_0 : \boldsymbol{\beta}_S = (\phi_1, ..., \phi_S)^T$. Note that H_0 is true for both tests.

Coverage was often low for $n = 100$. Coverage tended to be lower for the residual bootstrap than for the parametric bootstrap and lower for the two parameter true model than for the 1 parameter true model. For $n = 400$, coverage tended to be near or higher than the nominal 0.95. The MIX coverage was often better than the model selection coverage, which was not the case for regression variable selection simulations in Rathnayake and Olive (2022). Coverage could be near 1 if many zeroes were produced, but then the model selection CI length tended to be shorter than the full model CI length. See Table 2 for the AR simulation and Table 3 for the MA simulation. An entry of 1− means the coverage was more than 0.9995 but less than 1.0.

Table 3: MA(q) Model Selection,residual bootstrap,n=400,tstype=2,B=100,qmax=5,btype=2

e	θ_1	θ_2	$\theta_{q_{max}-1}$	$\theta_{q_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.949	0.953	0.952	0.952	0.931	0.982	0.990	0.932	0.960	0.973
len	0.215	0.242	0.246	0.223	2.876	2.876	3.678	2.478	2.478	2.709
N,MS	0.955	0.961	1.000	1.000	0.999	0.998	0.999	0.950	0.957	0.966
len	0.215	0.231	0.155	0.103	4.321	4.321	4.540	2.499	2.499	2.606
N,MIX	0.952	0.961	1.000	1.000	0.999	0.999	1−	0.948	0.950	0.959
len	0.204	0.217	0.129	0.088	4.598	4.598	4.791	2.496	2.496	2.572
t,full	0.952	0.956	0.958	0.956	0.939	0.987	0.992	0.940	0.962	0.973
len	0.214	0.242	0.246	0.222	2.881	2.881	3.651	2.482	2.482	2.704
t,MS	0.960	0.966	1.000	1.000	0.999	0.999	1−	0.956	0.964	0.970
len	0.214	0.230	0.154	0.102	4.328	4.328	4.546	2.502	2.502	2.607
t,MIX	0.957	0.961	1.000	1.000	1−	0.999	1.000	0.954	0.954	0.962
len	0.203	0.216	0.127	0.087	4.601	4.601	4.791	2.495	2.495	2.571
U,full	0.951	0.957	0.953	0.942	0.938	0.987	0.992	0.943	0.967	0.978
len	0.215	0.243	0.246	0.223	2.874	2.874	3.681	2.479	2.479	2.711
U,MS	0.956	0.971	1.000	1.000	1−	0.999	1−	0.961	0.967	0.972
len	0.215	0.232	0.155	0.104	4.327	4.327	4.549	2.504	2.504	2.611
U,MIX	0.958	0.965	1.000	1.000	1−	1−	1−	0.957	0.955	0.962
len	0.204	0.217	0.129	0.089	4.598	4.598	4.789	2.495	2.495	2.571
E,full	0.959	0.952	0.957	0.954	0.935	0.985	0.993	0.940	0.964	0.975
len	0.213	0.241	0.245	0.222	2.878	2.878	3.673	2.479	2.479	2.701
E,MS	0.967	0.964	1.000	1.000	1−	1−	1−	0.957	0.962	0.969
len	0.213	0.229	0.153	0.102	4.329	4.329	4.550	2.497	2.497	2.604
E,MIX	0.962	0.960	1.000	1.000	1.000	1−	1.000	0.949	0.955	0.962
len	0.204	0.215	0.126	0.087	4.598	4.598	4.791	2.496	2.496	2.569

ARMA models are much harder to bootstrap since it is much harder to get a consistent full model. In the simulation, we used a consistent full model. Estimating the order of the full model with Pötscher (1990) would likely cause the coverages to be worse. There were also convergence problems with the program, which would run with 1000 runs but not for 5000. Hence we ran the program 5 times with 1000 runs, and averaged the results. The true model was the ARMA(3,1) model with $\boldsymbol{\phi} = (0.7, 0.1, -0.4)^T$ and $\theta = 0.1$. The CI coverages for ϕ_1, ϕ_2 and θ_1 were too high. With $n = 100$, the coverages for tests were sometimes low. See Table 4.

Table 4: ARMA(p,q), parametric bootstrap, n=100, runs=5000, B=100, pmax=3, qmax=3

e	ϕ_1	ϕ_2	θ_1	θ_2	pr0	hyb0	br0	pr1	hyb1	br1
N,full	1.000	0.996	0.999	0.971	0.945	0.938	0.969	0.975	0.893	0.985
len	1.782	1.797	1.894	1.565	2.643	2.643	3.020	3.629	3.629	4.163
N,MS	0.998	0.988	0.999	1−	0.994	0.937	0.997	0.970	0.936	0.974
len	1.704	1.518	1.871	1.256	2.996	2.996	3.314	3.530	3.530	3.913
N,MIX	1.000	0.995	1.000	1.000	1−	0.940	0.998	0.964	0.923	0.968
len	1.670	1.387	1.837	1.215	3.095	3.095	3.425	3.582	3.582	4.007
t,full	0.999	0.997	0.998	0.972	0.947	0.937	0.969	0.978	0.903	0.988
len	1.741	1.768	1.846	1.461	2.643	2.643	2.997	3.607	3.607	4.111
t,MS	1.000	0.991	0.999	0.999	0.992	0.950	0.980	0.975	0.944	0.979
len	1.647	1.505	1.793	1.202	3.012	3.012	3.321	3.537	3.537	3.893
t,MIX	1.000	0.996	1.000	0.999	0.994	0.953	0.995	0.971	0.934	0.971
len	1.617	1.371	1.766	1.154	3.101	3.101	3.408	3.595	3.595	3.990
U,full	1.000	0.997	0.999	0.972	0.939	0.930	0.969	0.972	0.883	0.982
len	1.753	1.784	1.867	1.528	2.647	2.647	3.043	3.634	3.634	4.150
U,MS	0.999	0.989	0.998	0.999	0.992	0.937	0.992	0.964	0.930	0.966
len	1.695	1.537	1.859	1.246	2.996	2.996	3.318	3.547	3.547	3.923
U,MIX	1.000	0.996	0.999	1.000	0.995	0.941	0.996	0.963	0.920	0.964
len	1.652	1.398	1.828	1.208	3.091	3.091	3.421	3.606	3.606	4.025
E,full	0.999	0.998	0.998	0.977	0.953	0.941	0.972	0.976	0.900	0.984
len	1.779	1.801	1.895	1.562	2.646	2.646	3.016	3.641	3.640	4.160
E,MS	0.999	0.991	0.998	1.000	0.996	0.945	0.996	0.976	0.941	0.975
len	1.701	1.525	1.866	1.260	3.008	3.008	3.323	3.531	3.531	3.913
E,MIX	0.999	0.998	0.999	0.999	0.998	0.951	0.998	0.970	0.933	0.970
len	1.676	1.394	1.847	1.220	3.113	3.113	3.425	3.594	3.594	4.007

5. Discussion

Although there is a massive literature for variable selection and model selection, this paper may give the first large sample theory for ARMA time series model selection estimators. More theory is needed for the assumption $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$

and for the regularity conditions for the asymptotic normality of the GMLE for MA and ARMA time series. More bootstrap theory for Equation (16) is also needed.

A competitor for model selection is data splitting. Perform model selection on $Y_1, ..., Y_{n_h}$ to obtain model I. Then fit model I on the remaining cases $Y_{n_h+1}, ..., Y_n$ and perform inference. Inference is correct provided $S \subseteq I$. See Hurvich and Tsai (1989).

Bhansali (1981) discusses the effects of estimating the time series order, and there is a large literature for bootstrapping time series. See, for example, Bühlmann (1994, 1997, 2002), Härdle, Horowitz, and Kreiss (2003), Kreiss and Lahiri (2012), Kreiss, Paparoditis, and Politis (2011), and Lahiri (2003).

Simulations were done in *R*. See R Core Team (2020). The collection of *R* functions *tspack*, available from (http://parker.ad.siu.edu/Olive/tspack.txt), has some useful functions for the inference. The *tspack* function `arboottest` was used to get the confidence intervals for Example 1. The *tspack* function `msarsim` simulates AR model selection using the Yule Walker equations with AIC and the *R* function `ar.yw` for Table 2. The *tspack* function `msmasim` simulates MA model selection using the GMLE with AIC_C for Table 3. The *tspack* function `msarmamasim4` was used for Table 4. The last two functions used the *R* function `auto.arima` from the Hyndman and Khandakar (2008) *forecast package*. Also see Hyndman and Athanasopoulos (2018).

REFERENCES

Akaike, H. (1973), "Information Theory as an Extension of the Maximum Likelihood Principle," in *Proceedings, 2nd International Symposium on Information Theory*, eds. Petrov, B.N., and Csakim, F., Akademiai Kiado, Budapest, 267-281.

Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, Wiley, Hoboken, NJ.

Anderson, T.W. (1977), "Estimation for Autoregressive Moving Average Models in the Time and Frequency Domains," *The Annals of Statistics*, 5, 842-865.

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, 588-597.

Bickel, P.J., and Ren, J.–J. (2001), "The Bootstrap in Hypothesis Testing," in *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet*, eds. de Gunst, M., Klaassen, C., and van der Vaart, A., The Institute of Mathematical Statistics, Hayward, CA, 91-112.

Box, G.E.P, and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control*, revised ed., Holden-Day, Oakland, CA.

Bühlmann, P. (1994), "Blockwise Bootstrapped Empirical Process for Stationary Sequences," *The Annals of Statistics*, 22, 995-1012.

Bühlmann, P. (1997), "Sieve Bootstrap for Time Series," *Bernoulli*, 3, 5123-148.

Bühlmann, P. (2002), "Bootstraps for Time Series," *Statistical Science*, 17, 52-72.

Chan, N.H., Ling, S., and Yau, C.Y. (2020), "Lasso-Based Variable Selection of ARMA Models," *Statistica Sinica*, 30, 1925-1948.

Claeskens, G., and Hjort, N.L. (2008), *Model Selection and Model Averaging*, Cambridge University Press, New York, NY.

Durbin, J. (1959), "Efficient Estimation of Parameters in Moving-Average Models," *Biometrika*, 46, 306-316.

Durbin, J., and Koopman, S.J. (2001), *Time Series Analysis by State Space Methods*, Oxford University Press, Oxford, UK.

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference*, 143, 1039-1048.

Granger, C.W.J., and Newbold, P. (1977), *Forecasting Economic Time Series*, Academic Press, New York, NY.

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

Haile, M.G. and Olive, D.J. (2022), "Prediction Intervals and Regions for Some Time Series, Random Walks, and Renewal Processes," unpublished manuscript. See (http://parker.ad.siu.edu/Olive/pptspi.pdf).

Hall, P. (1988), "Theoretical Comparisons of Bootstrap Confidence Intervals," (with discussion), *The Annals of Statistics*, 16, 927-985.

Hamilton, J.D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Hannan, E.J. (1973), "The Asymptotic Theory of Linear Time-Series Models," *Journal of Applied Probability*, 10, 130-145.

Hannan, E.J. (1980), "The Estimation of the Order of an ARMA Process," *The Annals of Statistics*, 8, 1071-1081.

Hannan, E.J., and Quinn, B.G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, B*, 41, 190-195.

Hannan, E.J., and Rissanen, J. (1982), "Recursive Estimation of Mixed Autoregressive-Moving Average Order," *Biometrika*, 69, 81-94.

Härdle, W., Horowitz, J., and Kreiss, J.-P. (2003), "Bootstrap Methods for Time Series," *International Statistical Review*, 71, 435-459.

Hurvich, C., and Tsai, C.L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297-307.

Hyndman, R.J., and Athanasopoulos, G. (2018), *Forecasting: Principles and Practice*, 2nd edition, OTexts: Melbourne, Australia. https://OTexts.org/fpp2/

Hyndman, R.J., and Khandakar, Y. (2008), "Automatic Time Series Forecasting: the Forecast Package for R." *Journal of Statistical Software*, 26, 1-22.

Kreiss, J.P., (1985), "A Note on M-Estimation in Stationary ARMA Processes," *Statistics & Risk Modeling*, 3, 317-336.

Kreiss, J.-P., and Lahiri, S.N. (2012), "Bootstrap Methods for Time Series," in *Handbook of Statistics 30, Time Series Analysis Methods and Applications*, eds. Rao, T.S., Rao, S.S., and Rao, C.R., Elsevier, Oxford, UK, 3-26.

Kreiss, J.-P., Paparoditis, E., and Politis, D.N. (2011), "On the Range of Validity of the Autoregressive Sieve Bootstrap," *The Annals of Statistics,* 39, 2103-2130.

Lahiri, S.N. (2003), *Resampling Methods for Dependent Data*, Springer, New York, NY.

Mann, H.B., and Wald, A. (1943), "On the Statistical Treatment of Linear Stochastic Difference Equations," *Econometrica*, 11, 173-220.

McElroy, T.S., and Politis, D.N. (2020), *Time Series: a First Course With Bootstrap Starter*, CRC Press Taylor & Francis, Boca Raton, FL.

Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2018), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, 59, 913-931.

Pankratz, A. (1983), *Forecasting with Univariate Box-Jenkins Models*, Wiley, New York, NY.

Pelawa Watagoda, L. C. R., and Olive, D.J. (2021a), "Bootstrapping Multiple Linear Regression after Variable Selection," *Statistical Papers,* 62, 681-700.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021b), "Comparing Six Shrinkage Estimators With Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers,* 62, 2407-2431.

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, 165-179.

Pötscher, B. (1991), "Effects of Model Selection on Inference," *Econometric Theory*, 7, 163-185.

Pratt, J.W. (1959), "On a General Concept of "in Probability"," *The Annals of Mathematical Statistics,* 30, 549-558.

Rathnayake, R.C., and Olive, D.J. (2021), "Bootstrapping Some GLM and Survival Regression Variable Selection Estimators," *Communications in Statistics: Theory and Methods*, to appear. (http://parker.ad.siu.edu/Olive/ppbootglm.pdf).

R Core Team (2020), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

Shibata, R. (1976), "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," *Biometrika*, 63, 117-126.

Whittle, P. (1953), "Estimation and Information in Stationary Time Series," *Arkiv för Matematik*, 2, 423-34.

Yao, Q. and Brockwell, P.J. (2006), "Gaussian Maximum Likelihood Estimation for ARMA Models I: Time Series," *Journal of Time Series Analysis*, 27, 857-875.