Large Sample Theory for Some ARMA Time Series
Model Selection Estimators

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

January 21, 2021

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 asym-
totic distribution is a nonnormal mixture distribution. Hence bootstrap confidence
regions that can handle this complicated distribution were used for hypothesis test-
ing. A bootstrap technique to eliminate selection bias is to fit the model selection
estimator $\hat{\beta}_{MS}^{*}$ to a bootstrap sample to find a submodel, then draw another boot-
strap sample and fit the same submodel to get the bootstrap estimator $\hat{\beta}_{MIX}^{*}$.

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 $\theta$ with a positive sign. Many
references and software will write the model with a negative sign for the moving average
parameters. A moving average $\text{MA}(q)$ 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 $\text{AR}(p)$ 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 $\text{ARMA}(p, q)$ 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 a graduate student 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 $\sigma^2$. 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($\infty$) and MA($\infty$) parameters are square summable. Thus if the ARMA($p,q$) model is weakly stationary, causal, and invertible, then $Y_t$ almost entirely on nearby lags of $Y_t$ and $e_t$, not on the distant past. Also, the time series model $\approx \text{AR}(p_y) \approx \text{MA}(q_y)$ for some positive integers $p_y$ and $q_y$ that do not depend on the sample size $n$.

The mean function $\mu_t = E(Y_t)$ for $t \in \mathbb{Z}$, the set of integers. The autocovariance function $\gamma_{t,s} = \text{Cov}(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s$ for $t, s \in \mathbb{Z}$. A process $\{Y_t\}$ is weakly stationary if a) $E(Y_t) = \mu_t \equiv \mu$ is constant over time, and b) $\gamma_{t,t-k} = \gamma_0$, $k$ for all times $t$ and lags $k$. Hence the covariance function $\gamma_{t,s}$ depends only on the absolute difference $|t - s|$. For a weakly stationary process $\{Y_t\}$, write the autocovariance function as $\gamma_k = \text{Cov}(Y_t, Y_{t-k})$. Then the mean function $E(Y_t) = \mu$ and the variance function $V(Y_t) = \gamma_0$ are constant and do not depend on $t$, while the autocovariance function $\gamma_k$ depends on the lag $k$ but not on the time $t$.

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 ARMA model selection, we may use $p_{\max} = q_{\max} = 5$. For MA model selection, we may use $q_{\max} = 13$. For AR model selection, Granger and Newbold (1977, p. 178) suggest using $p_{\max} = 13$ for nonseasonal time series. 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 Akaiae (1973) AIC, Schwartz (1978) BIC, or Hurvich and Tsai (1989) AICC criterion are used for model selection.

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. For ARMA model selection, let $\beta = (\phi^T, \theta^T)^T = (\phi_1, ..., \phi_{p_{\max}}, \theta_1, ..., \theta_{q_{\max}})^T$ with $b = p_{\max} + q_{\max}$. For AR model selection, let $\beta = (\phi_1, ..., \phi_{p_{\max}})^T$ with $b = p_{\max}$, and for MA model selection, let $\beta = (\theta_1, ..., \theta_{q_{\max}})^T$ with $b = q_{\max}$. Hence $\beta = (\beta_1, ..., \beta_{p_{\max}+q_{\max}+1}, ..., \beta_{p_{\max}+q_{\max}+q_{\max}})^T$. Let $S = \{1, ..., p_{\max} + 1, ..., p_{\max} + q_{\max} + 1, ..., p_{\max} + q_{\max} + q_{\max}\}$ 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 $\hat{\beta}_{I,0}$ be a $b \times 1$ estimator of $\beta$
which is a obtained by padding $\hat{\beta}_I$ with zeroes. If $\beta_I = (\phi_1, ..., \phi_{p_I}, \theta_1, ..., \theta_{q_I})^T$, then $\hat{\beta}_{I,0} = (\hat{\phi}_1, ..., \hat{\phi}_{p_I}, 0, ..., 0, \hat{\theta}_1, ..., \hat{\theta}_{q_I}, 0, ..., 0)^T$. If $q_I = 0$, then $\hat{\beta}_{I,0} = (\hat{\phi}_1, ..., \hat{\phi}_{p_I}, 0, ..., 0)^T$. If $p_I = 0$ then $\hat{\beta}_{I,0} = (0, ..., 0, \hat{\theta}_1, ..., \hat{\theta}_{q_I}, 0, ..., 0)^T$. If $I = \emptyset$ with $p_I = q_I = 0$, then define $\hat{\beta}_{I,0} = 0$, 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 $\hat{\beta}_S = (\hat{\phi}_1, \hat{\theta}_1, \hat{\theta}_2)^T$, $\hat{\beta}_{S,0} = (\hat{\phi}_1, 0, 0, 0, 0, \hat{\theta}_1, \hat{\theta}_2, 0, 0, 0)^T$, and $\hat{\beta}_{I,0} = (\hat{\phi}_1, 0, 0, 0, 0, \hat{\theta}_1, \hat{\theta}_2, 0, 0, 0)^T$.

The model $I_{min}$ corresponds to the model that minimizes the AIC, $AICC$, or BIC criterion. Then the model selection estimator $\hat{\beta}_{MS} = \hat{\beta}_{I_{min},0}$. 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 (2019, 2020) and Rathnayake and Olive (2019).

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

$$F_u(t) = \sum_{j=1}^{J} \pi_j F_{u_j}(t)$$

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

$$E(h(u)) = \sum_{j=1}^{J} \pi_j E[h(u_j)]$$

and

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

If $E(u_j) = \theta$ for $j = 1, ..., J$, then $E(u) = \theta$ and

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

Inference will consider bootstrap hypothesis testing with confidence intervals (CIs) and regions. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ where $\theta_0$ is a known $g \times 1$ vector. A large sample $100(1 - \delta)\%$ confidence region for $\theta$ is a set $\mathcal{A}_n$ such that $P(\theta \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$. Then reject $H_0$ if $\theta_0$ is not in the confidence region. Let the $g \times 1$ vector $T_n$ be an estimator of $\theta$. Let...
Let $T_1, \ldots, T_B$ be the bootstrap sample for $T_0$. Let $A$ be a full rank $g \times b$ constant matrix. For model selection, test $H_0 : A\beta = \theta_0$ versus $H_1 : A\beta \neq \theta_0$ with $\theta = A\beta$. Then let $T_n = A\hat{\beta}_{SEL}$ and let $T_i = A\hat{\beta}_{SEL}$ for $i = 1, \ldots, B$ and $SEL$ is $MS$ or $MIX$. Let $[x]$ 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 $\theta$ is

$$[T_{(s)}^*, T_{(s+c-1)}^*] \text{ with } c = \min(B \lfloor 1 + 1.12\sqrt{\delta/n} \rfloor). \quad (2)$$

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. Let

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

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

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

$$D_i^2 = D_i^2(T, C) = D_i^2(z_i, T, C) = (z_i - T)^T C^{-1}(z_i - T)$$

for each observation $z_i$. Let $T^*$ and $S_T^*$ 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 (2019) hybrid (6) large sample $100(1-\delta)\%$ confidence regions for $\theta$ are $\{w : D_w(T^*, S_T^* \leq D_{(UB)}^2)\} = \{w : (w - T^*)^T [S_T^*]^{-1}(w - T^*) \leq D_{(UB)}^2\}$

where $D_{(UB)}^2$ is computed from $D_i^2 = (T^*_i - T^*)^T [S_T^*]^{-1}(T^*_i - T^*)$ for $i = 1, \ldots, B$ (if $g = 1$, (4) is a closed interval centered at $T^*$ just long enough to cover $UB$ of the $T_i^*$),

$$\{w : D_w(T_n, S_T^*) \leq D_{(UB,T)}^2\} = \{w : (w - T_n)^T [S_T^*]^{-1}(w - T_n) \leq D_{(UB,T)}^2\} \quad (5)$$

where the cutoff $D_{(UB,T)}^2$ is the $100q_B$th sample percentile of the $D_i^2 = (T_i^* - T_n)^T [S_T^*]^{-1}(T_i^* - T_n)$, and $\{w : D_w(T_n, S_T^*) \leq D_{(UB)}^2\} = \{w : (w - T_n)^T [S_T^*]^{-1}(w - T_n) \leq D_{(UB)}^2\}$. \quad (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 (2019) gave simpler proofs and proved that (2) is a large sample CI. Assume $u_n \overset{D}{\to} u$
where \( u_n = \sqrt{n}(T_i^* - T_n), \sqrt{n}(T_i^* - \bar{T}^t) \), \( \sqrt{n}(T_n - \theta) \), or \( \sqrt{n}(\bar{T}^* - \theta) \), and \( nS_T^* \sim P \rightarrow C \) where \( C \) is nonsingular. Let

\[
D_1^2 = D_{T_i}^2(T^t, S_T^*) = \sqrt{n}(T_i^* - T_n)^T(nS_T^*)^{-1}\sqrt{n}(T_i^* - T_n),
\]
\[
D_2^2 = D_{\theta}^2(T_n, S_T^*) = \sqrt{n}(T_n - \theta)^T(nS_T^*)^{-1}\sqrt{n}(T_n - \theta),
\]
\[
D_3^2 = D_{\theta}^2(T^t, S_T^*) = \sqrt{n}(\bar{T}^* - \theta)^T(nS_T^*)^{-1}\sqrt{n}(\bar{T}^* - \theta), \text{ and}
\]
\[
D_4^2 = D_{T_i}^2(T_n, S_T^*) = \sqrt{n}(T_i^* - T_n)^T(nS_T^*)^{-1}\sqrt{n}(T_i^* - T_n).
\]

Then \( D_j^2 \approx u^T(nS_T^*)^{-1}u \approx u^T C^{-1} u \), and the percentiles of \( D_1^2 \) and \( D_2^2 \) 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

\[
\frac{|S_T^*|^{1/2}}{|S_T^*|^{1/2}} \left( \frac{D(U_B)}{D(U_B,T)} \right)^g = \left( \frac{D(U_B)}{D(U_B,T)} \right)^g.
\]

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

Section 2 gives large sample theory for \( \hat{\beta}_{MIX} \) and \( \hat{\beta}_{VS} \). 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,

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

is a consistent estimator of \( \sigma^2 \) where often \( c = 0 \) or \( c = 1 \). See Granger and Newbold (1977, p. 85) and Hannan and Rissanen (1982, p. 89). Let \( \tilde{\sigma}^2 \) be the estimator of \( \sigma^2 \) produced by the time series model. Let

\[
\Gamma_n = \begin{bmatrix}
\gamma_0 & \gamma_1 & \ldots & \gamma_{n-1} \\
\gamma_1 & \gamma_0 & \ldots & \gamma_{n-2} \\
\vdots & \vdots & \ddots & \vdots \\
\gamma_{n-1} & \gamma_{n-2} & \ldots & \gamma_0
\end{bmatrix}
\]

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).

**Theorem 1.** Let the iid zero mean \( e_i \) have variance \( \sigma^2 \), and let the time series have mean \( E(Y_i) = \mu \).
a) Let \( Y_1, \ldots, Y_n \) be a weakly stationary and invertible AR\((p)\) time series, and let \( \beta = (\phi_1, \ldots, \phi_p) \). Let \( \hat{\beta} \) be the Yule-Walker estimator of \( \beta \). Then
\[
\sqrt{n}(\hat{\beta} - \beta) \overset{D}{\rightarrow} N_p(0, V)
\] (9)
where \( V = V(\beta) = \sigma^2\Gamma_p^{-1} \). Equation (9) also holds under mild regularity conditions for the least squares estimator, and the GMLE of \( \beta \).

b) Let \( Y_1, \ldots, Y_n \) be a weakly stationary, causal, and invertible MA\((q)\) time series, and let \( \beta = (\theta_1, \ldots, \theta_q) \). Let \( \hat{\beta} \) be the GMLE. Under regularity conditions,
\[
\sqrt{n}(\hat{\beta} - \beta) \overset{D}{\rightarrow} N_q(0, V).
\] (10)
where \( V = V(\beta) = \sigma^2\Gamma_q^{-1} \).

c) Let \( Y_1, \ldots, Y_n \) be a weakly stationary, causal, and invertible ARMA\((p, q)\) time series, and let \( \beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q) \) with \( g = p + q \). Let \( \hat{\beta} \) be the GMLE. Under regularity conditions,
\[
\sqrt{n}(\hat{\beta} - \beta) \overset{D}{\rightarrow} N_g(0, V).
\] (11)

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\((\infty)\) model which is approximately an AR\((p_q)\) model, and the large sample theory for the AR\((p_q)\) model depends on the zero mean error distribution through \( \sigma^2 \) 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 \( N(0, \sigma^2_0) \), \( V = V(\beta) = I_1^{-1}(\beta) \), the inverse information matrix. Then for the AR\((p)\) model, \( V(\phi) = \sigma^2\Gamma_p^{-1}(\phi) = I_1^{-1}(\phi) \), while for the MA\((q)\) model, \( V(\theta) = \sigma^2\Gamma_q^{-1}(\theta) = I_1^{-1}(\theta) \). See Box and Jenkins (1976, p. 241) and McElroy and Politis (2020, pp. 340-344).

Next we extend the Pelawa Watagoda and Olive (2019, 2020) and Rathnayake and Olive (2020) 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 \( \sqrt{n}(\hat{\beta}_{I_j} - \beta_{I_j}) \overset{D}{\rightarrow} N_{a_j}(0, V_j) \) where \( V_j \) is the covariance matrix of the asymptotic multivariate normal distribution. Then
\[
\sqrt{n}(\hat{\beta}_{I_{j,0}} - \beta_{I_{j,0}}) \overset{D}{\rightarrow} N_b(0, V_{j,0})
\] (12)
where \( V_{j,0} \) adds columns and rows of zeros corresponding to the \( \beta_i \) not indexed by \( I_j \), and \( V_{j,0} \) is singular unless \( I_j \) corresponds to the full model.

The first assumption in Theorem 2 is \( P(S \subseteq I_{\min}) \rightarrow 1 \) as \( n \rightarrow \infty \). Then the model selection estimator corresponding to \( I_{\min} \) underfits with probability going to zero. For AR model selection, the probability of underfitting goes to 0 if the AIC, BIC, or AICC criterion are used, at least if the \( e_t \) are iid \( N(0, \sigma^2) \). Also see Claeskens and Hjort (2008, pp. 39, 40, 45, 46), Hannan and Quinn (1979), and Shibata (1976). Charkhi and Claeskens (2018) show that AIC can be used for a wide variety of error distributions for multiple linear regression variable selection, and it may be possible to extend these results to AR model selection. For MA\((q)\) and ARMA\((p, q)\) model selection, the assumption has perhaps not yet been proved. However, the condition is necessary for the model selection
estimator $\hat{\beta}_{MS}$ to be a consistent estimator of $\beta$. See Rathnayake and Olive (2020). The assumption on $u_{jn}$ in Theorem 2 is reasonable by (12) since $S \subseteq I_j$ for each $\pi_j$, and since $\beta_{MIX}$ uses random selection. The proofs of Theorems 2, 3, and 4 are exactly as in Rathnayake and Olive (2020).

**Theorem 2.** Assume $P(S \subseteq I_{\text{min}}) \to 1$ as $n \to \infty$, and let $\hat{\beta}_{MIX} = \hat{\beta}_{I,0}$ with probabilities $\pi_{kn}$ where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive $\pi_k$ by $\pi_j$. Assume $u_{jn} = \sqrt{n}(\hat{\beta}_{I,j,0} - \beta) \xrightarrow{D} u_j \sim N_b(0, V_{j,0})$. a) Then

$$u_n = \sqrt{n}(\hat{\beta}_{MIX} - \beta) \xrightarrow{D} u \quad (13)$$

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

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

$$v_n = Au_n = \sqrt{n}(A\hat{\beta}_{MIX} - A\beta) \xrightarrow{D} Av = v \quad (14)$$

where $v$ has a mixture distribution of the $v_j = Au_j \sim N_g(0, AV_{j,0}A^T)$ with probabilities $\pi_j$.

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

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

**Proof.** a) Since $u_n$ has a mixture distribution of the $u_{kn}$ with probabilities $\pi_{kn}$, the cdf of $u_n$ is $F_{u_n}(t) = \sum_k \pi_{kn} F_{u_{kn}}(t) \to F_u(t) = \sum_j \pi_j F_{u_j}(t)$ at continuity points of the $F_{u_j}(t)$ as $n \to \infty$.

b) Since $u_n \xrightarrow{D} u$, then $Au_n \xrightarrow{D} Au$.

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). □

Theorem 2 can be used to justify prediction intervals after model selection. 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 variable selection consistency where $\pi_a = \pi_S = 1$ if $P(I_{\text{min}} = S) \to 1$ as $n \to \infty$. See Claeskens and Hjort (2008) for references.

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

The following Pelawa Watagoda and Olive (2019) theorem is useful for bootstrapping model selection estimators. Let $(\overline{T}, S_T)$ be the sample mean and sample covariance matrix computed from $T_1, \ldots, T_B$ which have the same distribution as $T_i$ where $T_i = T_{in}$. Let $D^2_{(U_B)}$ be the cutoff computed from the $D^2_i(\overline{T}, S_T)$ for $i = 1, \ldots, B$. The hyperellipsoids
corresponding to $D^2(T_n, C)$ and $D^2(T, C)$ are centered at $T_n$ and $T$, respectively. Note that $D^2_{T_n}(T_n, C) = D^2_{T_n}(T, C)$. Thus $D^2_{T_n}(T_n, C) \leq D^2_{(U_B)}$ iff $D^2_{T_n}(T, C) \leq D^2_{(U_B)}$. In Theorem 3, since $R_c$ contains $T_c$ with probability $1 - \delta_B$, the region $R_c$ contains $T$ with probability $1 - \delta_B$. Since $T_n$ depends on the sample size $n$, we need $(nS_T)^{-1}$ to be fairly well behaved, e.g. $(nS_T)^{-1} \overset{P}{\rightarrow} \Sigma^{-1}$.

**Theorem 3: Geometric Argument.** Suppose $\sqrt{n}(T_n - \theta) \overset{D}{\rightarrow} u$ with $E(u) = 0$ and $\text{Cov}(u) = \Sigma u \neq 0$. Assume $T_1, \ldots, T_B$ are iid with nonsingular covariance matrix $\Sigma_{T_n}$ where $(nS_T)^{-1} \overset{P}{\rightarrow} \Sigma^{-1}$ Then the large sample $100(1 - \delta)\%$ prediction region $R_p = \{ w : D^2_w(T, S_T) \leq D^2_{(U_B)} \}$ centered at $T$ contains a future value of the statistic $T$ with probability $1 - \delta_B$ which is eventually bounded below by $1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{ w : D^2_w(T_n, S_T) \leq D^2_{(U_B)} \}$ is a large sample $100(1 - \delta)\%$ confidence region for $\theta$ where $T_n$ is a randomly selected $T_i$.

Examining the iid data cloud $T_1, \ldots, T_B$ and the bootstrap sample data cloud $T_1^*, \ldots, T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \theta)$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $u \sim N_p(0, \Sigma_A)$, say, then the bootstrap sample data cloud of $T_1^*, \ldots, T_B^*$ is like the data cloud of iid $T_1, \ldots, T_B$ 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 $\sqrt{n}(T^* - T_n) \overset{P}{\rightarrow} 0$.

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

Pötscher (1991) used the conditional distribution of $|\hat{\beta}_{MS} - \hat{\beta}_{I_k,0}|$ to find the distribution of $w_n = \sqrt{n}(\hat{\beta}_{MS} - \beta)$. Define $P(A|B_k)P(B_k) = 0$ if $P(B_k) = 0$. Let $\hat{\beta}_{I_k,0}^C$ be a random vector from the conditional distribution $\hat{\beta}_{I_k,0} | (\hat{\beta}_{MS} = \hat{\beta}_{I_k,0})$. Let $w_{kn} = \sqrt{n}(\hat{\beta}_{I_k,0} - \beta) | (\hat{\beta}_{MS} = \hat{\beta}_{I_k,0}) \sim \sqrt{n}(\hat{\beta}_{I_k,0}^C - \beta)$. Denote $F_{z}(t) = P(z_1 \leq t_1, \ldots, z_p \leq t_p)$ by $P(z \leq t)$. Then Pötscher (1991) and Pelawa Watagoda and Olive (2020) show

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

Hence $\hat{\beta}_{MS}$ has a mixture distribution of the $\hat{\beta}_{I_k,0}^C$ with probabilities $\pi_{kn}$, and $w_n$ has a mixture distribution of the $w_{kn}$ with probabilities $\pi_{kn}$.

Note that both $\sqrt{n}(\hat{\beta}_{MIX} - \beta)$ and $\sqrt{n}(\hat{\beta}_{MS} - \beta)$ are selecting from the $u_{kn} = \sqrt{n}(\hat{\beta}_{I_k,0} - \beta)$ and asymptotically from the $u_j$. The random selection for $\hat{\beta}_{MIX}$ does not change the distribution of $u_{jn}$, but selection bias does change the distribution of the selected $u_{jn}$ and $u_j$ to that of $w_{jn}$ and $w_j$. The assumption that $w_{jn} \overset{D}{\rightarrow} w_j$ 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 $\hat{\beta}_{MS} = \hat{\beta}_{I_k,0}$ with probabilities $\pi_{kn}$ where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive $\pi_j$ by $\pi_j$. Assume
\[ w_{jn} = \sqrt{n}(\hat{\beta}_{I,0}^C - \beta) \xrightarrow{d} w_j. \] Then
\[ w_n = \sqrt{n}(\hat{\beta}_{MS} - \beta) \xrightarrow{d} w \]
(15)
where the cdf of \( w \) is \( F_w(t) = \sum_j \pi_j F_{w_j}(t) \). Thus \( w \) is a mixture distribution of the \( w_j \) with probabilities \( \pi_j \).

3. Bootstrapping ARMA Time Series Model Selection Estimators

For the bootstrap, we will ignore \( \tau \) 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
\[ Y^*_t = \sum_{k=1}^{p_{max}} \hat{\phi}_k Y^*_{t-k} + e^*_t, \]
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 \( t = 1, \ldots, n \). The ARMA and AR bootstrap use a block of initial values \( (Y^*_{p+1}, \ldots, Y^*_n)^T \) randomly selected from \( Y_1, \ldots, Y_n \). For the parametric bootstrap, assume the full model produces \( m \) residuals \( r_1, \ldots, r_m \). 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 \( j = 1, \ldots, m \). 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 \( V(e^*_t) \approx \hat{\sigma}^2 \). Instead of computing the full model, use model selection and zero padding to compute \( I_k \) and \( \hat{\beta}_{MS,1}^* \). Draw another bootstrap data set and fit model \( I_k \) to get \( \hat{\beta}_{MS,1}^* \). Repeat \( B \) times to get the bootstrap samples \( \hat{\beta}_{MS,1}^*, \ldots, \hat{\beta}_{MS,B}^* \) and \( \hat{\beta}_{MIX,1}^*, \ldots, \hat{\beta}_{MIX,B}^* \). Let the selection probabilities for the bootstrap model selection estimator be \( \rho_{kn} \). Then this bootstrap procedure bootstrap both \( \hat{\beta}_{MS} \) and \( \hat{\beta}_{MIX} \) with \( \pi_{kn} = \rho_{kn} \).

Following McElroy and Politis (2020, pp. 438-439), consider a weakly stationary and invertible time series \( Y_1, \ldots, Y_n \) where the \( e_t \) are iid with mean 0 and variance \( \sigma^2 \). A companion process uses \( \epsilon_t \) that are iid with mean 0 and variance \( \hat{\sigma}^2 \). Both the residual bootstrap and nonparametric 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(\( \infty \)) 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 \( \sqrt{n}(\hat{\beta}^* - \hat{\beta}) \overset{D}{\to} N_b(0, V(\beta)) \) 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 \( n \to \infty \), \( Y_1^*, \ldots, Y_n^* \) with \( \hat{\beta} = \hat{\beta}_n \) satisfies Theorem 1. Hence if \( n \) is fixed and the time series \( Y_1^*, \ldots, Y_n^* \) is generated with \( \beta_n \), then on the set \( A \) the estimator \( \hat{\beta}^* \) satisfies \( \sqrt{m}(\hat{\beta}^* - \hat{\beta}_n) \overset{D}{\to} N_b(0, V(\hat{\beta}_n)) \) as \( m \to \infty \). Since \( V(\hat{\beta}) \overset{P}{=} V(\beta) \) if \( \hat{\beta}_n \overset{P}{=} \beta \) as \( n \to \infty \), it follows that \( \sqrt{n}(\hat{\beta}^* - \hat{\beta}_n) \overset{D}{\to} N_b(0, V(\beta)) \) as \( n \to \infty \). \( \square \)

The key idea is that for the parametric bootstrap, \( Y_1^*, \ldots, Y_n^* \) satisfies the Gaussian time series model with \( \hat{\beta}_n \) as the parameter vector and \( \hat{\beta}_n \) is a \( \sqrt{n} \) consistent estimator of \( \beta \). Hence the Gaussian time series \( Y_1^*, \ldots, Y_n^* \) with \( \hat{\beta}_n \) will be weakly stationary, causal, and invertible on a set \( A \) going to one in probability. Since \( \hat{\beta}_n \) 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 \( \sqrt{n}(\hat{\beta}_{I_j,0} - \beta) \overset{D}{\to} N_b(0, V_{j,0}) \), we would like to show that \( \sqrt{n}(\hat{\beta}_{I_j,0} - \hat{\beta}_{I_j,0}) \overset{D}{\to} N_b(0, V_{j,0}) \) 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

\[
Y_t^* = \sum_{k=1}^{p_{I_j}} \hat{\phi}_k Y_{t-k}^* + e_t^* + e_t^*(I_j),
\]

or

\[
Y_t^* = \sum_{k=1}^{p_{I_j}} \hat{\phi}_k Y_{t-k}^* + e_t^* + e_t^*(I_j),
\]

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

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 \( T_1^*, \ldots, T_B^* \) can be
written as
\[ T_{1,1}^*, ..., T_{B_{in},1}^*, ..., T_{1,J}^*, ..., T_{B_{in},J}^*. \]

Denote \( T_{ij}^* \) as the \( j \)th bootstrap component of the bootstrap sample with sample mean \( \bar{T}_j^* \) and sample covariance matrix \( S_{T,j}^* \). Similarly, we can define the \( j \)th component of the iid sample \( T_1, ..., T_B \) to have sample mean \( \bar{T}_j \) and sample covariance matrix \( S_{T,j} \).

Let \( T_n = \hat{\beta}_{MIX} \) and \( T_{ij} = \hat{\beta}_{I,j} \). If \( S \subseteq I_j \), assume \( \sqrt{n}(\hat{\beta}_{I,j} - \beta_{I,j}) \overset{D}{\rightarrow} N_0(0, V_j) \) and \( \sqrt{n}(\hat{\beta}_{I,j}^* - \hat{\beta}_{I,j}) \overset{D}{\rightarrow} N_0(0, V_j) \). Then by Equation (12),
\[ \sqrt{n}(\hat{\beta}_{I,j} - \beta) \overset{D}{\rightarrow} N_p(0, V_{j,0}) \quad \text{and} \quad \sqrt{n}(\hat{\beta}_{I,j,0}^* - \hat{\beta}_{I,j,0}) \overset{D}{\rightarrow} N_0(0, V_{j,0}). \] (16)

This result means that the component clouds have the same variability asymptotically. The iid data component clouds are all centered at \( \beta \). If the bootstrap data component clouds were all centered at the same value \( \tilde{\beta} \), then the bootstrap cloud would be like an iid data cloud shifted to be centered at \( \beta \), and (5) would be a confidence region for \( \theta = \beta \). Instead, the bootstrap data component clouds are shifted slightly from a common center, and are each centered at a \( \beta_{I,j,0} \). 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 \( \bar{T}^* \) tends to estimate \( \theta \) at least as well as \( T_n \) for a mixture distribution. A similar argument holds if \( T_n = A\hat{\beta}_{MIX}, T_{ij} = A\hat{\beta}_{I,j,0} \) and \( \theta = A\beta \).

In the simulations of Section 4 for \( H_0 : A\beta = B\beta_S = \theta_0 \) with \( n \geq 20b \), the coverage tended to get close to \( 1 - \delta \) for \( B \geq \max(200, 50b) \) so that \( S_T^* \) is a good estimator of \( \text{Cov}(T^*) \). In the simulations where \( S \) is not the full model, inference with backward elimination with \( I_{\min} \) using \( AIC \) was often more precise than inference with the full model if \( n \geq 20b \) and \( B \geq 50b \).

The matrix \( S_T^* \) can be singular due to one or more columns of zeros in the bootstrap sample for \( \beta_1, ..., \beta_b \). The \( \beta_j \) 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 \( \hat{\beta}^* = \hat{\beta}_{FULL} \) to the bootstrap sample. For example, take \( k = \lceil cB \rceil \) with \( c = 0.01 \). A confidence interval \( [L_n, U_n] \) can be computed without \( S_T^* \) 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. Simulations

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. Often $p_{\text{max}}$ and $q_{\text{max}}$ were rather small to make the simulation time shorter.

We simulated AR model selection with the Yule Walker estimators and AIC. Let $k = p_{\text{max}}$. 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 $\phi = (0.5, 0.33)$. Error types were $N(0,1)$, $t_5$, uniform(-1,1), and $e \sim W - 1$ where $W \sim \text{exponential}(1)$. The parametric bootstrap and residual bootstrap were used. Nominal 95% confidence regions and intervals were used with 1% augmentation from the bootstrapped full model. The simulations bootstrapped the full model $\beta = \hat{\phi}$, the model selection estimator $\hat{\beta}_{MS}$, and $\hat{\beta}_{MIX}$.

The tables give two rows for each of the three estimators giving the observed CI coverage and average CI length. The term “full” is for the AR($p_{\text{max}}$) 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 $H_0 : \beta_E = \mathbf{0}$ where $\beta_E = (\beta_{p+1}, ..., \beta_k)^T$. The 1 indicates the test $H_0 : \beta_S = (\phi_1, ..., \phi_S)^T$. Note that $H_0$ is true for both tests.

Table 1: AR(p) Model Selection, parametric bootstrap, $n=100, \phi = 0.5, B=100$, $p_{\text{max}}=5$

<table>
<thead>
<tr>
<th>e</th>
<th>$\phi_1$</th>
<th>$\phi_2$</th>
<th>$\phi_{p_{\text{max}}-1}$</th>
<th>$\phi_{p_{\text{max}}}$</th>
<th>pr0</th>
<th>hyb0</th>
<th>br0</th>
<th>pr1</th>
<th>hyb1</th>
<th>br1</th>
</tr>
</thead>
<tbody>
<tr>
<td>N,full</td>
<td>0.9486</td>
<td>0.9536</td>
<td>0.9662</td>
<td>0.9724</td>
<td>0.9588</td>
<td>0.9610</td>
<td>0.9724</td>
<td>0.9344</td>
<td>0.9480</td>
<td>0.9578</td>
</tr>
<tr>
<td>len</td>
<td>0.4274</td>
<td>0.4611</td>
<td>0.4443</td>
<td>0.3965</td>
<td>3.1913</td>
<td>3.1913</td>
<td>3.8924</td>
<td>1.9595</td>
<td>1.9595</td>
<td>2.0673</td>
</tr>
<tr>
<td>N,MS</td>
<td>0.9434</td>
<td>0.9982</td>
<td>0.9998</td>
<td>1.0000</td>
<td>0.9976</td>
<td>0.9960</td>
<td>0.9980</td>
<td>0.9590</td>
<td>0.9660</td>
<td>0.9728</td>
</tr>
<tr>
<td>len</td>
<td>0.4327</td>
<td>0.4410</td>
<td>0.3661</td>
<td>0.2718</td>
<td>4.1515</td>
<td>4.1515</td>
<td>4.4179</td>
<td>1.9633</td>
<td>1.9633</td>
<td>2.0642</td>
</tr>
<tr>
<td>N,MIX</td>
<td>0.9426</td>
<td>0.9984</td>
<td>0.9998</td>
<td>1.0000</td>
<td>0.9992</td>
<td>0.9980</td>
<td>0.9988</td>
<td>0.9576</td>
<td>0.9660</td>
<td>0.9760</td>
</tr>
<tr>
<td>len</td>
<td>0.4251</td>
<td>0.3885</td>
<td>0.2853</td>
<td>0.1842</td>
<td>4.6250</td>
<td>4.6250</td>
<td>4.8664</td>
<td>1.9650</td>
<td>1.9650</td>
<td>2.0758</td>
</tr>
</tbody>
</table>

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 $P(S \subseteq I_{\text{min}}) \rightarrow 1$ as $n \rightarrow \infty$ 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).


Simulations were done in $R$. See R Core Team (2016). 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 msarsim simulates AR model selection using

REFERENCES


