Time Series Data Splitting and Outlier Detection

Welagedara, W.A.D.M. and David J. Olive *
Southern Illinois University

July 26, 2022

Abstract

This paper discusses data splitting for time series inference and outlier detection for some time series models.

KEY WORDS: ARIMA, prediction interval, variable selection.

1 Introduction

This section reviews autoregressive moving average (ARMA) time series models. We follow Haile (2022) and Haile and Olive (2022a) closely for sections 1 and 2. A *time series* $Y_1, ..., Y_n$ consists of observations Y_t collected sequentially at times $1, ..., n$. 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. For the time series models described below, we will assume that the errors e_t are independent and identically distributed (iid) with zero mean and variance σ^2. The backshift operator or lag operator B satisfies $BW_t = W_{t-1}$ and $B^j W_t = W_{t-j}$.

A *moving average* MA(q) times series is

$$Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t = \tau + (1 + \theta_1 B + \cdots + \theta_q B^q) e_t = \tau + \theta(B) e_t$$

where $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ and $\theta_q \neq 0$. Note that $E(Y_t) = \mu = \tau = \theta_0$ for $t \geq 1$. Since the e_t are iid, the Y_t are identically distributed, and $Y_j, Y_{j+q+1}, Y_{j+2(q+1)}$, ... are iid.

An *autoregressive* AR(p) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \text{ or } \ (1 - \phi_1 \mathrm{B} - \cdots - \phi_\mathrm{p} \mathrm{B}^\mathrm{p}) \mathrm{Y_t} = \tau + \mathrm{e_t},$$

or $\phi(B) Y_t = \tau + e_t$ where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ and $\phi_p \neq 0$. If $E(Y_t) = \mu$ for $t \geq 1$, write $Y_t - \mu = \sum_{j=1}^{p} \phi_j (Y_{t-j} - \mu) + e_t$ to get $\tau = \phi_0 = \mu(1 - \sum_{j=1}^{p} \phi_j)$.

An *autoregressive moving average* ARMA(p, q) times series is

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

*

or $\phi(B)Y_t = \tau + \theta(B)e_t$ where $\theta_q \neq 0$ and $\phi_p \neq 0$. The ARMA(0,q) model is the MA(q) model, and the ARMA(p,0) model is the AR(p) model. Again $\tau = \mu(1 - \sum_{j=1}^{p} \phi_j)$ if $p \geq 1$, and $\tau = \mu$ if $p = 0$. The ARMA(0,0) model is $Y_t = \mu + e_t$, often called the location model.

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. To describe ARIMA models, let the difference operator $\bigtriangledown = (1 - B)$. Let $Y_t = \bigtriangledown^d X_t = (1-B)^d X_t$ be the differenced time series. The first difference is $Y_t = \bigtriangledown X_t = (1-B)X_t = X_t - X_{t-1}$. The second difference is $Y_t = \bigtriangledown^2 X_t = \bigtriangledown(\bigtriangledown X_t) = X_t - 2X_{t-1} + X_{t-2}$. If X_t follows an ARIMA(p, d, q) model, want Y_t to follow a weakly stationary, causal, and invertible ARMA(p, q) = ARIMA$(p, 0, q)$ model. Typically $d = 0$ or 1, but occasionally $d = 2$. Usually $\tau = 0$ if $d > 1$. The ARIMA$(p, d = 1, q)$ model is $X_t = \tau + (1 + \phi_1)X_{t-1} + (\phi_2 - \phi_1)X_{t-2} + \cdots + (\phi_p - \phi_{p-1})X_{t-p} - \phi_p X_{t-p-1} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$. The ARIMA$(p, d, q)$ model can be written compactly as $\phi(B) \bigtriangledown^d X_t = \tau + \theta(B)e_t$. See Box and Jenkins (1976) for more on these models.

A *stochastic process* $\{Y_t, t \in \mathbb{T}\}$ is a collection of random variables where often $\mathbb{T} = \mathbb{Z}$, the set of integers. The observed time series is $\{Y_t\} = Y_1, ..., Y_n$. The *mean function* $\mu_t = E(Y_t)$ for $t \in \mathbb{Z}$. The *autocovariance function* $\gamma_{t,s} = Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s$ for $t, s \in \mathbb{Z}$. The *autocorrelation function* $\rho_{t,s} = Corr(Y_t, Y_s) = \dfrac{Cov(Y_t, Y_s)}{\sqrt{Var(T_t)Var(Y_s)}} = \dfrac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}$ 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 $\gamma_k = Cov(Y_t, Y_{t-k})$ and the *autocorrelation function* as $\rho_k = corr(Y_t, Y_{t-k}) = \gamma_k/\gamma_0$. Note that the mean function $E(Y_t) = \mu$ and the variance function $V(Y_t) = Var(Y_t) = \gamma_0$ are constant and do not depend on t. The autocovariance and autocorrelation functions γ_k and ρ_k depend on the lag k but not on the time t.

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 and μ is some origin otherwise. 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.

Consider $\theta(B)$ and $\phi(B)$ as polynomials in B. An ARMA(p, q) model is invertible if all of the roots of the polynomial $\theta(B) = 0$ have modulus > 1, and weakly stationary if all of the roots of the polynomial $\phi(B) = 0$ have modulus > 1. (Let the complex number $W = W_1 + W_2 \ \ i$ have modulus $|W| = W_1^2 + W_2^2$.) Hence the roots of both polynomials lie outside the unit circle. An AR(p) model is always invertible and an MA(q) model is always causal. For the AR(1) model, need $|\phi_1| < 1$. For the MA(1) model, need $|\theta_1| < 1$.

For the ARMA(1,1) model, need $|\phi_1| < 1$ and $|\theta_1| < 1$.

Let τ_i stand for θ_i or ϕ_i. Let k stand for q or p, and let $\psi(B) = 1 - \tau_1 B - \tau_2 B^2 - \cdots - \tau_k B^k$ stand for $\phi(B)$ or $\theta(B)$. A necessary but not sufficient condition for the roots of $\psi(B) = 0$ to all be greater than 1 in modulus is $\tau_1 + \cdots + \tau_k < 1$ and $|\tau_k| < 1$.

1.1 Large Sample Theory

Some notation is 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,

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

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}) \tag{2}$$

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

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

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

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,

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

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 $N(0, \sigma_e^2)$, $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\beta})$, 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).

Section 2 reviews model selection.

2 Model Selection

Let I be a time series model. The $AIC(I)$ statistic is used to pick a model from several ARIMA models. The model I_{min} with the smallest AIC is always of interest but often overfits: has too many unnecessary parameters. Imagine fitting an ARIMA(p, d, q) model where $d = 0, 1$ or 2 is fixed and p and q run from 0 to j for small j. The number of parameters in the model for fixed d is $p+q+2$ where $\sigma = \sqrt{V(X_t)}$, τ, $\phi_1, ..., \phi_p$, $\theta_1, ..., \theta_q$ are the parameters. $AIC(I)$ tends to be large when the model does not have enough terms, to drop as needed terms are added, and then to rise as unnecessary terms are added. If $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline. See Brockwell and Davis (1987, p. 269), Duong (1984), and Burnham and Anderson (2004).

Haile and Olive (2022a) extend regression variable selection notation to ARMA time series model selection as in the next few paragraphs. Consider regression models where the response variable Y is independent of the $p \times 1$ vector of predictors $\boldsymbol{x}$ given $\boldsymbol{x}^T\boldsymbol{\beta}$, written $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$. Many important regression models satisfy this condition, including multiple linear regression and generalized linear models (GLMs).

Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{5}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (5) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{I/S} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ if $S \subseteq I$. The model using $\boldsymbol{x}^T\boldsymbol{\beta}$ is the full model.

To clarify notation, suppose $p = 4$, a constant $x_1 = 1$ corresponding to β_1 is always in the model, and $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, 0)^T$. Then the $J = 2^{p-1} = 8$ possible subsets of $\{1, 2, ..., p\}$ that always contain 1 are $I_1 = \{1\}$, $S = I_2 = \{1, 2\}$, $I_3 = \{1, 3\}$, $I_4 = \{1, 4\}$, $I_5 = \{1, 2, 3\}$, $I_6 = \{1, 2, 4\}$, $I_7 = \{1, 3, 4\}$, and $I_8 = \{1, 2, 3, 4\}$. There are $2^{p-a_S} = 4$ subsets I_2, I_5, I_6, and I_8 such that $S \subseteq I_j$. Also, $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4)^T$ is obtained by regressing Y on $\boldsymbol{x}_{I_7} = (x_1, x_3, x_4)^T$.

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or backward elimination. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. Also use zero padding for the model I_{min}. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets. For example, if each subset contains at least one variable, then there are $J = 2^p - 1$ subsets.

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 $p_{max} = 0$. If model selection is restricted to AR models, Granger and Newbold (1977, p. 178) suggest using $p_{max} = 13$ for nonseasonal time series, quarterly seasonal time series, and short monthly seasonal time series. They recommend $p_{max} = 25$ for longer monthly seasonal time series. 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. See Example 1 where there are 36 submodels.

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, 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).

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 $S \subseteq I$. Note that the full model, e.g. the ARMA(p_{max}, q_{max}) model, is

a submodel.

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 $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$. Assume $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$. Haile and Olive (2022a) gave the large sample theory for $\hat{\boldsymbol{\beta}}_{MS}$, and used bootstrap confidence regions for hypothesis testing.

Example 1. Shown below is the aicmatrix of $\Delta(I) = AIC(I) - AIC(I_{\min})$ for the R `WWW usage` time series, which gives the number of users connected to the Internet through a server every minute where $n = 100$. First differences were used so $d = 1$. From this output, I_{min} is the ARIMA(5,1,4) model. Some interesting models are the ARIMA(3,1,0) model and the ARIMA(1,1,1) model.

```
aicmat(WWWusage,dd=1,pmax=5)
$aics            q
  p      0     1    2     3    4    5
  0 119.86 38.67 8.74  9.13 8.24 7.72
  1  18.10  3.16 5.11  3.44 3.96 5.14
  2  11.04  5.15 6.22  4.63 2.10 6.95
  3   0.85  2.80 4.48  3.27 3.62 5.29
  4   2.79  1.74 5.04  7.94 4.26 6.99
  5   4.72  6.50 2.40 10.50 0.00 1.63
```

3 Data Splitting

Data splitting is used to get valid inference. If the model was selected without using the time series, then the model has an asymptotic normal distribution that can be used for inference. If the entire time series is used to build or select the model, then the resulting model tends not to have an asymptotic normal distribution due to selection bias. If the first half of the time series is used to build or select the model, and that model is fit on the second half of the time series, then inference is valid (the model for the second half of the time series was selected without using the second half). Time series models are often built or selected using the entire data set with transformations such as the log transformation, model selection with AIC, BIC, or AIC_C. Plots such as the ACF and PACF are also used to select the model.

A problem with using a half set is that the efficiency using $n/2$ is less than that of using n. Sequential data splitting can be used as a remedy. Data splitting divides the training data set of n cases into two sets H and the validation set V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. A common method of data splitting randomly divides the training data into the two sets H and V. Often $n_H \approx \lceil n/2 \rceil$ where $\lceil x \rceil$ is the ceiling function = least integer function, e.g $\lceil 7.7 \rceil = 8$. See, for example, Hurvich and Tsai (1989).

An application of data splitting is to use a model selection method on H to get a model I. On the validation set V, fit time series model I. Then use the standard time series inference. For AR model selection and MA model selection, data splitting works if the selected model does not underfit. For the GMLE and an ARMA model, assume the ARMA(p_S, q_S) model is the true model. Then the selected model I is an ARMA(p_I, q_I) models. The model I should not underfit ($p_I \geq p_S$ and $q_I \geq q_S$), and needs $p_I = p_S$ or $q_I = q_S$ for valid inference.

Sequential data splitting may be useful. Let $\lfloor x \rfloor$ be the integer part of x, e.g., $\lfloor 7.7 \rfloor = 7$. Let the ceiling function $\lceil x \rceil$, e.g. $\lceil 7.7 \rceil = 8$. Initially, divide the data set into two sets H_1 with the first $n_1 \leq n/2$ cases Y_t and V_1 with the last $n - n_1$ cases where $n_1 = 30$ or $n_1 = 2(p_{max} + q_{max})$. Apply model selection on H_1 to get model I_1, an AR(p_{I_1}), MA(q_{I_1}), or ARMA(p_{I_1}, q_{I_1}) model if AR, MA or ARMA model selection is used. Let $a_1 = p_{I_1} + q_{I_1}$. If $n_1 \geq 10a_1$, set $H = H_1$ and $V = V_1$. Otherwise, let $n_2 = 2n_1$, use $Y_1, ..., Y_{n_2}$ in H_2 and the last $n - n_2$ cases in V_2. Apply model selection on H_2 to get model I_2. Let $a_2 = p_{I_2} + q_{I_2}$. If $n_2 \geq 10a_2$, set $H = H_2$ and $V = V_2$. Continue in this manner, forming sets $(H_1, V_1), (H_2, V_2), ..., (H_d, V_d)$ where H_i has $n_i = in_1$. Stop when $n_d \geq 10a_d$ or $n_{d+1} > n/2$. For the second case, use $n_d = \lfloor n/2 \rfloor$. Then $H = H_d$ and $V = V_d$. Use the model I_d for inference with the data in $V = V_d$.

Use simulation to examine whether the model underfits (and for ARMA data splitting, picks $p_I = p_S$ or $q_I = q_S$ where $I = I_d$).

Sequential data splitting was used for regression models in Zhang and Olive (2022).

4 Outlier Detection

Outliers are cases that lie far away from the pattern set by the bulk of the data, and can be often be detected from the plot of t versus Y_t and from the response plot of $\hat{Y}_t$ versus Y_t with the identity line that has zero intercept and unit slope added as a visual aid. In both plots Y_t is on the vertical axis, and the vertical deviations of Y_t from the identity line are the residuals $\hat{e}_t = Y_t - \hat{Y}_t$. The residual plot of $\hat{Y}_t$ versus $\hat{e}_t$ is also useful.

The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,} \tag{6}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation MED(n) = MED$(Y_1, ..., Y_n)$ will also be used. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{7}$$

Assume the time series Y_t is weakly stationary with an MA(∞) representation. Let $up = \text{MED}(n) + k\text{MAD}(n)$ and $low = \text{MED}(n) - k\text{MAD}(n)$ where $k = 6$ is the default. Make a new time series W_t where if $low \leq Y_t \leq up$, then $W_t = Y_t$. Otherwise, make W_t a missing value: let $W_t = NA$ if $Y_t < low$ or $Y_t > up$. This method is useful since software methods for handling missing values are widely available. See, for example, Jones (1980). The method may also be useful for handling heavy tailed time series, where the first or second moment of the Y_t does not exist. Since the W_t do not depend

on the ARMA model, plug in $W_1, ..., W_n$ into the time series software in place of the $Y_1, ..., Y_n$ to to get robust estimators of other quantities, such as the ACF and PACF. Compare Agnieszka and Magdalena (2018), Allende and Heiler (1992), Bhatia, et al. (2016), Bustos and Yohai (1986), Chakhchoukh (2010), Chang, Tiao, and Chen (1988), Chen and Liu (1993), Choy (2001), de Luna and Genton (2001), Denby and Martin (1979), Deutsch, Richards, and Swain (1990), Fox (1972), Justel, Peña, and Tsay (2001), Lawrence (2014), Ledolter (1989), Liu, Kumar, and Palomar (2019), Lucas, Franses, and Van Dijk (2009), Ma and Genton (2000), Muler, Peña, and Yohai (2009), Stockinger and Dutter (1987), Tsay (1986, 1988).

One alternative is to get a robustly Winzorize the time series W_t: if $Y_t > up$, then $W_t = max(Y_k \leq up)$. If $Y_t < low$, then $W_t = min(Y_k \geq low)$. If $low \leq Y_t \leq up$, then $W_t = Y_t$. Then fit the time series to W_t. A second alternative would set $W_t = \text{MED}(n)$ instead of NA. Variants would use the fitted time series to predict the W_t that were changed, fit the time series again, and perhaps repeat this step. These methods impute the potential outliers, but the existing methods for handling missing values likely impute better.

For an MA(q) model, the $Y_j, Y_{j+q+1}, Y_{j+2(q+1)}$, ... are iid. Hence there are $q+1$ iid sequences starting at $j = 1, ..., (q+1)$. Since the sample percentiles of the iid sequences converge in probability to the population percentiles for fixed h, so do the sample percentiles of all of the data. Hence the sample median and sample median absolute deviation converge to the corresponding population quantities, and similar results hold for MA(∞) models. Haile and Olive (2022b) used similar results to justify a time series prediction interval.

To see why $k = 6$ is recommended, examine the approximate proportion of cases not changed to NA for several distributions when no outliers are present. See Table 1. Let MED(X) and MAD(X) be the population median and median absolute deviation. Notation for the random variables is as in Olive (2008, ch. 10; 2014, ch. 10).

Table 1: Probability X $\in$ [MED(X) $-$ 6MAD(X), MED(X) + 6MAD(X)]

distribution of X	MED(X)	MAD(X)	prob
Cauchy(μ, σ)	μ	σ	0.8949
double exponential(θ, λ)	θ	$\log(2)\lambda$	0.9844
exponential(θ, λ)	$\theta + \log(2)\lambda$	$\lambda/2.0781$	0.9721
logistic(μ, σ)	μ	$\log(3)\sigma$	0.9973
N(μ, σ^2)	μ	σ	0.9999
uniform(θ_1, θ_2)	$(\theta_1 + \theta_2)/2$	$(\theta_2 - \theta_1)/4$	1

The AR(p) model is useful for illustrating problems outliers cause. Write the AR(p)

equations $Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t$ in matrix form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ or

$$\begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Y_p & Y_{p-1} & \dots & Y_1 \\ 1 & Y_{p+1} & Y_p & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{n-1} & Y_{n-2} & \dots & Y_{n-p} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} e_{p+1} \\ e_{p+2} \\ \vdots \\ e_n \end{bmatrix}$$

where $\boldsymbol{X}$ is of full rank with more rows than columns $p+1$ and $\boldsymbol{\beta} = (\phi_0, \boldsymbol{\phi}^T)^T = (\phi_0, \phi_1, ..., \phi_p)^T$. Note that if Y_{p+1} is an outlier, then Y_{p+1} is an outlier in the kth row and kth column of $\boldsymbol{X}$ for $k = 2, ..., p+1$. Differencing can cause even more outliers in the data.

"Robust" multiple linear regression estimators can be applied to ARIMA$(p, d, 0)$ data or data from the dynamic linear model to create a "robust" estimator. These estimators tend to work poorly for several reasons. First, the "robust" multiple linear regression estimators that are practical to compute tend to be inconsistent with poor outlier resistance. See Hawkins and Olive (2002), Huber and Ronchetti (2009), and Olive (2017b, 2022b).

The Olive (2017b) `rmreg2` estimator will be used as the robust multiple linear regression estimator. The (ordinary) least squares estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, $\hat{\phi}_{0,OLS} = \overline{Y} - \hat{\boldsymbol{\phi}}_{OLS}^T \overline{\boldsymbol{x}}$, and $\hat{\boldsymbol{\phi}}_{OLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$. Here $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ are the usual estimated covariance matrices used when $\boldsymbol{w}_i = (\boldsymbol{x}_i, Y_i)^T$ are iid from some population. The `rmreg2` estimator plugs in robust covariance estimators in place of the classical estimators. More details follow. Let

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{x} \\ Y \end{pmatrix}, \quad E(\boldsymbol{w}) = \boldsymbol{\mu}_{\boldsymbol{w}} = \begin{pmatrix} E(\boldsymbol{x}) \\ E(Y) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{x}} \\ \mu_Y \end{pmatrix}, \quad \text{and} \quad \text{Cov}(\boldsymbol{w}) = \boldsymbol{\Sigma}_{\boldsymbol{w}} =$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{x},Y} \\ \boldsymbol{\Sigma}_{Y,\boldsymbol{x}} & \boldsymbol{\Sigma}_{Y,Y} \end{pmatrix}.$$

Let $(T, \boldsymbol{C}) = (\tilde{\boldsymbol{\mu}}_{\boldsymbol{w}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}})$ be a robust estimator of multivariate location and dispersion. Then the robust plug in estimator $\tilde{\phi}_0 = \tilde{\mu}_Y - \tilde{\boldsymbol{\phi}}^T \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\tilde{\boldsymbol{\phi}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$. The robust estimator $(T, \boldsymbol{C})$ used will be the RMVN estimator of Olive (2017b), Olive and Hawkins (2010), and Zhang, Olive, and Ye (2012) that has been used to make robust estimators of multiple linear regression and multivariate linear regression. See Olive (2017b). The robust estimator has not yet been shown to be consistent for AR(p) data, but the robust estimator can be used as an outlier diagnostic.

Example 2. Here we examine outliers for the AR(p) model and use the Cryer and Chan (2008) *R* package `TSA` data set `deere1` which gives 82 consecutive values for the amount of deviation from a specified target value in an industrial machining process at Deere & Co. If there is an outlier at Y_k where k is not too close to 1 or n, then fitted values will use the outlier for $t = k+1, ..., k+p$. So the outlier appears $p+1$ times in the equations for the AR(p) model.

An AR(2) model will be used for the Deere time series, and the plot of the time series in Figure 1 shows that there is one large outlier, corresponding to case 27. Figure 2 shows

the response and residual plots for the AR(2) model. Only one outlier, instead of two, appears in the fitted values since $\hat{\phi}_1 = 0.027$ is quite small. The plots for the robust fit are similar and are not shown.

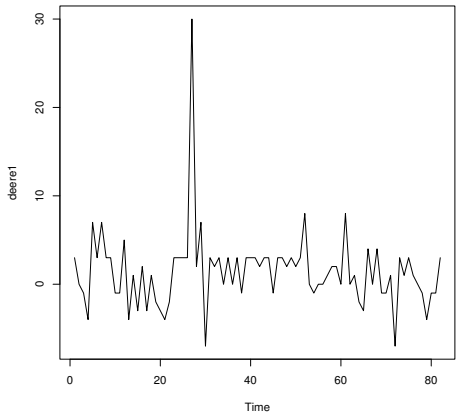

Figure 1: The Deere Time Series Has One Outlier

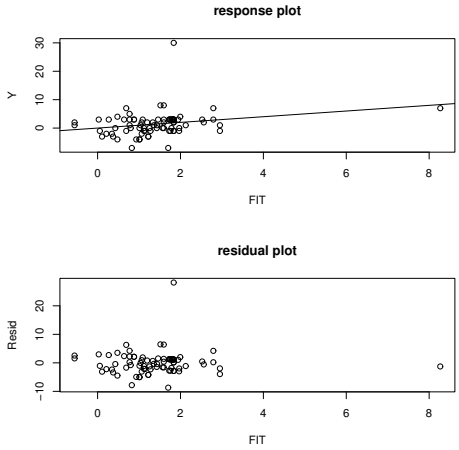

Figure 2: Response and Residual Plots for the AR(2) Model

The outlier Y_{27} is changed from 30 to a more reasonable value 8 to create "cleaned data." The robust AR(2) model was refit using the cleaned data resulting in "cleaned fitted values." In the original data, cases Y_7 and Y_{76} were changed to 25 and 26. The fitted values from the robust AR(2) models versus the cleaned fitted values showed some tilt. Next cases Y_7 and Y_{76} were changed to 250 and 260. Figure 3 shows fitted values from the robust AR(2) models versus the cleaned fitted values with the identity line added as a visual aid. The two sets of fitted values for the bulk of the data are similar since big outliers are easier to detect.

The *R* code below corresponds to the following.

a) Gives the plot of the time series. See Figure 1.

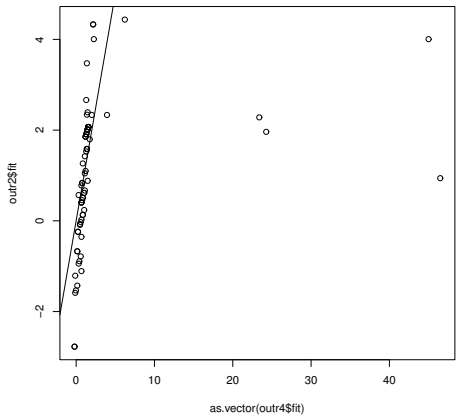

Figure 3: Fitted Values from the Cleaned Data Versus Robust Fitted Values from the Data with 3 Outliers

b) Gives the output table for the AR(2) model as well as the response and residual plots. See Figure 2.

c) The commands for this part fit a robust AR(2) model and gives the coefficient values and the response and residual plots.

d) The command for this part change the outlier from 30 to a more reasonable value 8 and refits the AR(2) model producing the output table for the AR(2) model, and the response and residual plots.

e) The commands for this part fit a robust AR(2) model on the cleaned data, giving the coefficient values for the AR(2) model, and the response and residual plots.

f) The commands for this part plot the fitted values from the robust AR(2) model fit to the data with the outlier versus the fitted values from the classical AR(2) model fit to the clean data. The fitted values are similar except for the outlier in Y_t and one of the outliers in $\hat{Y}_t$.

g) The commands for this part change the values of two cases (7 and 76) to 25 and 26, and fits the robust estimator. Then the commands plot the fitted values versus the fitted values of the robust estimator to the cleaned data. The fitted values are tilted some.

h) Now the values of the two cases (7 and 76) are changed to 250 and 260. Then the commands plot the fitted values versus the fitted values of the robust estimator to the cleaned data. The fitted values for the bulk of the data are similar since big outliers are easier to detect. See Figure 3.

i) These commands change the potential outliers to NA. For the deere1 data set, case 27 is changed to NA. The output table and response and residual plots are given.

j) These commands take the data set from g) and change the potential outliers to NA. The three outliers got NA. The output table and response and residual plots are given.

k) These commands take the data set from h) and change the potential outliers to NA. The three outliers got NA. The output table and response and residual plots are

given.

```
source("http://parker.ad.siu.edu/Olive/tspack.txt")

#library("TSA")
#data(deere1)
#plot(deere1)

deere1 <- c(3,0,-1,-4,7,3,7,3,3,-1,-1,5,-4,1,-3,2,-3,1,-2,-3,
-4,-2,3,3,3,3,30,2,7,-7,3,2,3,0,3,0,3,-1,3,3,3,2,3,3,-1,3,3,
2,3,2,3,8,0,-1,0,0,1,2,2,0,8,0,1,-2,-3,4,0,4,-1,-1,1,-7,3,1,
3,1,0,-1,-4,-1,-1,3)

#a)
plot(deere1,type="l")

#b)
out2 <- arima(deere1,c(2,0,0))
resplots(deere1,out2)

#c)
outr1 <- robar(deere1,2)
outr1$phihat
#right click Stop on the plot twice
#For each Y outlier in an AR(p) model,
#there will be p outliers in the X matrix
#which could cause up to p outliers in the fitted values.
#So p+1 outliers in the XY matrix used to compute the robust estimator.

#d)
deerem2 <- deere1
deerem2[27] <- 8
out2m <- arima(deerem2,c(2,0,0))
resplots(deerem2,out2m)

#e)
outr2 <- robar(deerem2,2)
outr2$phihat

#f)
cleanfit <- as.vector(deere1) - as.vector(out2m$resid)
plot(as.vector(outr1$fit,),cleanfit)
abline(0,1)

#g)
```

```
deerem3 <- deere1
deerem3[c(7,76)] <- c(25,26)
outr3 <- robar(deerem3,2)
plot(as.vector(outr3$fit),outr2$fit) #right click Stop 2 times, hit Enter
abline(0,1)
#identify(as.vector(outr3$fit),outr2$fit)
#NAs mean the identified points are off by 2
#74, 25, 5 instead of 76,27,7

#h)
deerem4 <- deere1
deerem4[c(7,76)] <- c(250,260)
outr4 <- robar(deerem4,2)
plot(as.vector(outr4$fit),outr2$fit) #right click Stop 2 times, hit Enter
abline(0,1)
#identify(as.vector(outr4$fit),outr2$fit
#works better with massive outliers
#outliers are easy to spot with response plot
#since there Y values are outlying

#i)
YNA <- tsNA(deere1)$W
out5 <- arima(YNA,c(2,0,0))
resplots(YNA,out5)

#j)
W2 <- tsNA(deerem3)$W
out6 <- arima(W2,c(2,0,0))
resplots(W2,out6)

#k)
W3 <- tsNA(deerem4)$W
out7 <- arima(W3,c(2,0,0))
resplots(W3,out7)
```

5 Dynamic Linear Models

Le $Y_i = \beta_1 + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + e_i$ for $i = 1, ..., n$ where the $Y_i, x_{i,j}$ and e_i each follow a time series for $j = 2, ..., p$. Then $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is a dynamic linear model in matrix form. If the e_i are iid, multiple linear regression methods with least squares (OLS) can be used for inference. Unfortunately, the iid error assumption rarely holds for dynamic linear model.

6 Discussion

Plots and simulations were done in *R*. See R Core Team (2018). Programs are in the collection of functions *tspack.txt*. See (http://parker.ad.siu.edu/Olive/tspack.txt). The function `robar` fits a robust AR(p) model. The function `tsNA` changes potential outliers to NA.

REFERENCES

Agnieszka, D., and Magdalena, L. (2018), "Detection of Outliers in the Financial Time Series Using ARIMA Models," *Applications of Electromagnetics in Modern Techniques and Medicine (PTZE)*, 2018, 49-52.

Allende, H., and Heiler, S. (1992), "Recursive Generalized M Estimates For Autoregressive Moving-Average Models," *Journal of Time Series Analysis,* 13, 1-18.

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.

Bhansali, R.J. (1981), "Effects of Not Knowing the Order of an Autoregressve Process on the Mean Squared Error of Prediction-I," *Journal of the American Statistical Association*, 76, 588-597.

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2016), "Efficient and Consistent Robust Time Series Analysis," arXiv preprint arXiv:1607.00146, arxiv.org.

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

Brockwell, P.J., and Davis, R.A. (1987), *Time Series: Theory and Methods*, Springer, New York, NY.

Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

Bustos, O.H., and Yohai, V.J. (1986), "Robust Estimates for ARMA Models," *Journal of the American Statistician*, 81, 155-168.

Chakhchoukh, Y. (2010), "A New Robust Estimation Method for ARMA Models," *IEEE Transactions on Signal Processing,* 58, 3512-3522.

Chan, N.H., Ling, S., and Yau, C.Y. (2020), "Lasso-based variable selection of ARMA models," *Statistica Sinica*, 30, 1925-1948.

Chang, I., Tiao, G.C., and Chen, C. (1988), "Estimation of Time Series Parameters in the Presence of Outliers," *Technometrics*, 30, 193-204.

Charkhi, A., and Claeskens, G. (2018), "Asymptotic Post-Selection Inference for the Akaike Information Criterion," *Biometrika*, 105, 645-664.

Chen, C. and Liu, L. (1993), "Joint Estimation of Model Parameters and Outlier Effects in Time Series," *Journal of the American Statistical Association*, 88, 284-297.

Choy, K. (2001), "Outlier Detection for Stationary Time Series," *Journal of Statistical Planning and Inference,* 99, 111-127.

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

Cryer, J.D., and Chan, K.-S. (2008), *Time Series Analysis: with Applications in R*, 2nd ed., Springer, New York, NY.

de Luna, X., and Genton, M.G. (2001), "Robust Simulation-Based Estimation of ARMA Models," *Journal of Computational and Graphical Statistics*, 10, 370-387.

Denby, L., and Martin, R.D. (1979), "Robust Estimation of the First-Order Autoregressive Parameter," *Journal of the American Statistical Association*, 74, 365, 140-146.

Deutsch, S.J., Richards, J.E., and Swain, J.J. (1990), "Effects of a Single Outlier on ARMA identification," *Communications in Statistics*, 19, 2207-2227.

Duong, Q.P. (1984), "On the Choice of the Order of Autoregressive Models: a Ranking and Selection Approach," *Journal of Time Series Analysis,* 5, 145-157.

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

Fox, A.J. (1972), "Outliers in Time Series," *Journal of the Royal Statistical Society: B*, 34, 350-363.

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. (2022a), "Bootstrapping ARMA Time Series Models after Model Selection," preprint at (http://parker.ad.siu.edu/Olive/pptsboot.pdf).

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

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.

Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," *Journal of the American Statistical Association,* (with discussion), 97, 136-148.

Huber, P.J., and Ronchetti, E.M. (2009), *Robust Statistics,* 2nd ed., Wiley, Hoboken, NJ.

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/

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Jones, R.H. (1980), "Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations," *Technometrics*, 22, 389-395.

Justel, A., Peña, D., and Tsay, R.S. (2001), "Detection of Outlier Patches in Autoregressive Time Series," *Statistica Sinica,* 11, 651-673.

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

Lawrence, C.J. (2014), "Robust Methods in Time Series Analysis," *Wiley StatsRef: Statistics Reference Online*.

Leeb, H., and Pötscher, B.M. (2006), "Can One Estimate the Conditional Distribution of Post–Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554-2591.

Ledolter, J. (1989), "The Effect of Additive Outliers on the Forecasts from ARIMA Models," *International Journal of Forecasting*, 5, 231-240.

Li, K.–C. (1987), "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics,* 15, 958-975.

Liu, J., Kumar, S., and Palomar, D.P. (2019), "Parameter Estimation of Heavy-Tailed AR Model with Missing Data Via Stochastic EM," *IEEE Transactions on Signal Processing*, 67, 2159-2172.

Lucas, A., Franses, P.H., and Van Dijk, D. (2009), *Outlier Robust Analysis of Economic Time Series*, Oxford University Press, Oxford, UK.

Ma, Y., and Genton, M.G. (2000), "Highly Robust Estimation of the Autocovariance Function," *Journal of Time Series Analysis,* 21, 663-684.

Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

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.

Muler, N., Peña, D., and Yohai, V. (2009), "Robust Estimation for ARMA Models," *The Annals of Statistics*, 37, 816-840.

Nishii, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics,* 12, 758-765.

Olive, D.J. (2008), *A Course in Statistical Theory*, online course notes at (http://parker.ad.siu.edu/Olive/infbook.htm).

Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

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

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

Olive, D.J. (2022a), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

Olive, D.J. (2022b), *Robust Statistics*, online course notes at (http://parker.ad.siu.edu/Olive/robbook.html).

Olive, D. J., and Hawkins, D. M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (http://parker.ad.siu.edu/Olive/pphbmld.pdf).

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

Pötscher, B.M. (1990), "Estimation of Autoregressive Moving-Average Order Given an Infinite Number of Models and Approximation of Spectral Sensities," *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.

R Core Team (2018), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019), "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference," *The Annals of Statistics*, 47, 3438-3469.

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

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: an Introduction with Applications,* Chapman & Hall, New York, NY.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486-494.

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

Stockinger, N., and Dutter, R. (1987), "Robust Times Series Analysis: a Survey," *Kybernetika*, 23, 3-88.

Tsay, R.S. (1986), "Time Series Model Specification in the Presence of Outliers," *Journal of the American Statistical Association*, 81, 132-141.

Tsay, R.S. (1988), "Outliers, Level Shifts, and Variance Changes in Time Series," *Journal of Forecasting*, 7, 1-20.

White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego, CA.

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

Yang, Y. (2003), "Regression with Multiple Candidate Models: Selecting or Mixing?" *Statistica Sinica*, 13, 783-809.

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.

Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation With Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.

Zhang, L., and Olive, D.J. (2022), "Data Splitting Inference," online at (http://parker.ad.siu.edu/Olive/ppdatsplittinf.pdf).