INFERENCE FOR TIME SERIES AFTER VARIABLE SELECTION

by

Mulubrhan G. Haile

B.Sc., University of Asmara, 2009
M.S., Southern Illinois University, 2017

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the
Doctor of Philosophy Degree

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale
May 2022

DISSERTATION APPROVAL

INFERENCE FOR TIME SERIES AFTER VARIABLE SELECTION

by

Mulubrhan G. Haile

A Dissertation Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in the field of Mathematics

Approved by:

Dr. David Olive, Chair

Dr. Seyed Yaser Samadi

Dr. Bhaskar Bhattacharya

Dr. Dashun Xu

Dr. Poopalasingam Sivakumar

Graduate School
Southern Illinois University Carbondale
March 29, 2022

AN ABSTRACT OF THE DISSERTATION OF

Mulubrhan G. Haile, for the Doctor of Philosophy degree in Mathematics, presented on March 29, 2022, at Southern Illinois University Carbondale.

TITLE: INFERENCE FOR TIME SERIES AFTER VARIABLE SELECTION

MAJOR PROFESSOR: Dr. David Olive

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 $\hat{\boldsymbol{\beta}}^*_{MIX}$. Prediction intervals for a wide variety of time series models are given, including prediction intervals after model selection.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr.David Olive for his invaluable assistance and insights leading to the writing of this paper. He was always available to meet and discuss when I've questions. My sincere thanks also goes to professors of my committee for their proper suggestions. Last but not least, I would like to thank the faculty and staff of the School of Mathematical and Statistical Sciences. I've enjoyed studying and working with you.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

A *time series* $Y_1, ..., Y_n$ consists of observations Y_t collected sequentially at times $1, ..., n$. Many time series models have the form

$$Y_t = \tau + \sum_i \psi_i Y_{t-ik_i} + \sum_j \nu_j e_{t-jk_j} + e_t \tag{1.1}$$

where the errors $\{e_t\}$ are independent and identically distributed (iid) unobserved random variables. Unless stated otherwise, assume the mean $E(e_t) = 0$ and the variance $V(e_t) = \sigma_e^2$. For example, the Box and Jenkins (1976) multiplicative seasonal ARIMA$(p, d, q) \times (P, D, Q)_s$ time series models have this form.

Next, several important time series models will be given. We will use the *R* software notation and write a moving average parameter θ and seasonal moving average parameter Θ with a positive sign. Some references and software will write the model with a negative sign for the moving average parameters. 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

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

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.

To describe ARIMA models, let the difference operator $\triangledown = (1-B)$. Let $X_t = \triangledown^d Y_t = (1-B)^d Y_t$ be the differenced time series. The first difference is $X_t = \triangledown Y_t = (1 - B)Y_t = Y_t - Y_{t-1}$. The second difference is $X_t = \triangledown^2 Y_t = \triangledown(\triangledown Y_t) = Y_t - 2Y_{t-1} + Y_{t-2}$. If Y_t follows an ARIMA(p, d, q) model, want X_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 $Y_t = \tau + (1 + \phi_1)Y_{t-1} + (\phi_2 - \phi_1)Y_{t-2} + \cdots + (\phi_p - \phi_{p-1})Y_{t-p} - \phi_p Y_{t-p-1} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$. The ARIMA$(p, d, q)$ model can be written compactly as $\phi(B) \triangledown^d Y_t = \tau + \theta(B)e_t$.

The multiplicative seasonal ARIMA models also have backshift and difference notation. Let $\Phi(B) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$. Let $\Theta(B) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$. Let s be the seasonal period. Hence $s = 4$ for quarterly data and $s = 12$ for monthly date. Then the multiplicative ARMA$(p, q) \times (P, Q)_s$ model satisfies $\phi(B)\Phi(B)Y_t = \tau + \theta(B)\Theta(B)e_t$. This model is an ARMA$(p+Ps, q+Qs)$ model where the nonzero coefficients are determined only by $p+P+q+Q$ coefficients, the AR characteristic polynomial is $\phi(B)\Phi(B)$ and the MA characteristic polynomial is $\theta(B)\Theta(B)$.

Let $\triangledown_s Y_t = (1 - B^s)Y_t = Y_t - Y_{t-s}$ and $\triangledown_s^D Y_t = (1 - B^s)^D Y_t$ where usually $d \leq 1$ and $D \leq 1$, $d = 2$ is rare and $D = 2$ is very rare. The differenced time series $X_t = \triangledown^d \triangledown_s^D Y_t$. Then $Y_t \sim$ ARIMA$(p, d, q) \times (P, D, Q)_s$ if $X_t \sim$ ARMA$(p, q) \times (P, Q)_s$. Also, $\phi(B)\Phi(B)\triangledown^d \triangledown_s^D Y_t = \tau + \theta(B)\Theta(B)e_t$ where the default is $\tau = 0$ if $d > 0$ or $D > 0$.

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 *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* $\rho_{t,s} = Corr(Y_t, Y_s) = \dfrac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}} = \dfrac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}$ 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 causal 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 $Z_t = \sum_{j=1}^{\infty} \pi_j Z_{t-j} + e_t$, which is an AR$(\infty)$ representation, where the $\pi_j \to 0$ rapidly as $j \to \infty$. 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.

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 $|\tau_k| < 1$.

Vector valued time series $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ are also common where $\boldsymbol{y}_t$ is a $k \times 1$ vector for $t \geq 1$.

CHAPTER 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).

The initial model to look at is the model I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. Similar I_I rules are used in Olive (2017a) and Olive and Hawkins (2005) for multiple linear regression and generalized linear models.

The aicmatrix computes $\Delta(I) = AIC(I) - AIC(I_{min})$ for ARIMA(p,d,q) models where d is fixed or for ARIMA$(p, d, q) \times (P, D, Q)_s$ models where d, P, D, Q and s are fixed, and p and q run from 0 to j for small $j = p_{max} = q_{max}$ such as $j = 5$. Here I_{min} is the ARIMA(p_m, d, q_m) model or the ARIMA$(p_m, d, q_m) \times (P, D, Q)_s$ model with the smallest AIC(I). This model will have a 0.00 in the aicmatrix. Look for model I_I with $p_I + q_I \leq p_m + q_m$ as small as possible such that the aicmatrix entry ≤ 2. It is possible that $I_I = I_{min}$. Also look at models I with $p + q \leq p_I + q_I$ with aicmatrix entries ≤ 7, especially models with entries ≤ 4. Check that the selected model I does not fail to reject H_0 for $H_0 : \phi_p = 0$ or $H_0 : \theta_q = 0$. Make the usual model checks of plotting the time series, ACF, PACF, response and residual plots, the ACF and PACF of the residuals, and the plot of the Box–Ljung pvalues.

Another useful concept is that of a submodel. If d, P, D, and Q are fixed and model I_i has p_i

and q_i for $i = 1, 2$, then I_1 is a submodel of I_2 if $p_1 \leq p_2$ and $q_1 \leq q_2$. If $\Delta(I_1) \leq \Delta(I) + 2$ where I_1 is a submodel of I, tentatively eliminate model I. Model I_1 will be a submodel of all models I with aicmatrix entries to the right and below the model I entry. Hence model I_1 is at the upper left corner of a block of models I such that I_1 is a submodel for each model I in the block.

These are rules of thumb: they do not always work but often lead to a good model. If I_I is the ARIMA(1,0,1) model, we might take an AR(3) or MA(3) model even though these have 1 more parameter.

Example 2.1. Shown below is the aicmatrix of $\Delta(I) = AIC(I) - AIC(I_{\min})$ for the R `WWWusage` 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 and I_I is the ARIMA(3,1,0) model. Interesting models have $p + q \leq 3$ with entries ≤ 7. These are the ARIMA(2,1,1), ARIMA(1,1,2), and ARIMA(1,1,1) models. Since the ARIMA(1,1,1) model is a submodel of the ARIMA(2,1,1) and ARIMA(1,1,2) models, look at the ARIMA(3,1,0) model I_I first, and then at the ARIMA(1,1,1) model.

```
aicmat(WWWusage,dd=1,pmax=5)
$aics            q
p      0     1    2     3    4    5  Find I_I by looking at models
0 119.86 38.67 8.74  9.13 8.24 7.72  on and above the diagonal
1  18.10  3.16 5.11  3.44 3.96 5.14  through (5,4) and (4,5) which have
2  11.04  5.15 6.22  4.63 2.10 6.95  p+q <= 9. Interesting models are on
3   0.85  2.80 4.48  3.27 3.62 5.29  or above the diagonal through (3,0),
4   2.79  1.74 5.04  7.94 4.26 6.99  (2,1), (1,2) and (0,3) since they
5   4.72  6.50 2.40 10.50 0.00 1.63  have p+q <= 3.
```

Suppose an ARMA(p_{max}, q_{max}) model is fit and then (model) variable selection is done where the true (optimal) model is an ARMA(p_o, q_o) model with $p_o \leq p_{max}$ and $q_o \leq q_{max}$. Let the selected model I be an ARMA(p_I, q_I) model. Then the model underfits unless $p_I \geq p_o$ and $q_I \geq q_o$. Let the weakly stationary and invertible AR(p) models have $q_{max} = 0$, and assume $p_o \leq p_{max}$. 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 for variable selection. See Hannan and Quinn (1979) and Shibata (1976).

For ARMA models, we may use $p_{max} = q_{max} = 5$. If (variable) model selection is restricted to MA models, we may use $q_{max} = 13$. 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 want to bootstrap time series variable selection estimators. 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

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{2.1}$$

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 (2.1) 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, $\boldsymbol{\beta}_O = \boldsymbol{0}$ 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.

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 variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, 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 $0 \le \pi_j \le 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \ge 2$, 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

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

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

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

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

Variable selection = model selection for ARMA time series will use similar notation. Note that S corresponds to the ARMA(p_o, q_o) model. Let $\boldsymbol{\beta}$ be an $m \times 1$ vector. Let $\boldsymbol{\beta} = (\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)^T = (\boldsymbol{\phi}^T, \boldsymbol{\theta}^T)^T$ with $m = p + q$. For an AR(p) model, let $\boldsymbol{\beta} = (\phi_1, ..., \phi_p)^T$ with $m = p$, and for an MA(q) model, let $\boldsymbol{\beta} = (\theta_1, ..., \theta_q)^T$ with $m = q$. If $\boldsymbol{\beta}_I = (\phi_1, ..., \phi_{p_I}, \theta_1, ..., \theta_{q_I})^T$, then $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, ..., \hat{\phi}_{p_I}, 0, .., 0, \hat{\theta}_1, ..., \hat{\theta}_{q_I}, 0, ..., 0)^T$. Sometimes we will use $\boldsymbol{\beta} = (\tau, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q)^T$, $\boldsymbol{\beta} = (\tau, \phi_1, ..., \phi_p)$, or $\boldsymbol{\beta} = (\tau, \theta_1, ..., \theta_q)^T$. For time series the number of submodels $J = (p_{max} + 1)(q_{max} + 1)$, $J = p_{max} + 1$, or $J = q_{max} + 1$, for ARMA, AR, or MA model selection. See Example 2.1 where there are 36 submodels. Note that the full model, e.g. the ARMA(p_{max}, q_{max}) model, is a submodel.

CHAPTER 3

PREDICTION INTERVALS

For forecasting, predict the test data $Y_{n+1}, ..., Y_{n+L}$ given the past training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ prediction interval (PI) for Y_{n+h} is $[L_n, U_n]$ where the coverage $P(L_n \leq Y_{n+h} \leq U_n) = 1 - \alpha_n$ is eventually bounded below by $1-\delta$ as $n \to \infty$. Often we want $1 - \alpha_n \to 1 - \delta$ as $n \to \infty$. By construction, some of the prediction intervals will have training data coverage $\approx 1 - \delta_n$ where $1 - \delta_n \geq 1 - \delta$, and $1 - \delta_n \to 1 - \delta$ as $n \to \infty$.

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

Example 3.1. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76
order data: 76 78 89 111 778
                  13 = 89 - 76
                      33 = 111 - 78
                         689 = 778 - 89
shorth(3) = [76,89]
```

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

$$[Z_{(s)}, Z_{(s+c-1)}] \text{ with } \mathrm{c} = \min(\mathrm{n}, \lceil \mathrm{n}[1-\delta+1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{3.1}$$

Some more notation is needed before deriving PIs for time series. Suppose the training data set is $Y_1, ..., Y_t$. The h-step ahead forecast for a future value Y_{t+h} is $\hat{Y}_t(h)$ and the h step ahead forecast residual is $\hat{e}_t(h) = Y_{t+h} - \hat{Y}_t(h)$. For example, a common choice for model (1.1) is

$$\hat{Y}_t(h) = \hat{\tau} + \sum_i \hat{\psi}_i Y^*_{t+h-ik_i} + \sum_j \hat{\nu}_j \hat{e}^*_{t+h-jk_j}$$

where $\hat{e}_t$ is the tth residual, $Y^*_{t+h-ik_i} = Y_{t+h-ik_i}$ if $h - ik_i \leq 0$, $Y^*_{t+h-ik_i} = \hat{Y}_t(h - ik_i)$ if $h - ik_i > 0$, $\hat{e}^*_{t+h-jk_j} = \hat{e}_{t+h-jk_j}$ if $h - jk_j \leq 0$, and $\hat{e}^*_{t+h-jk_j} = 0$ if $h - jk_j > 0$, and the forecasts $\hat{Y}_t(1), \hat{Y}_t(2), ..., \hat{Y}_t(L)$ are found recursively if there is data $Y_1, ..., Y_t$. Typically the residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ are the 1-step ahead forecast residuals and the fitted or predicted values $\hat{Y}_t = \hat{Y}_{t-1}(1)$ are the 1-step ahead forecasts.

Example 3.2 is useful to illustrate the forecasts. The *R* software produces $\hat{e}_t$ and $\hat{Y}_t = Y_t - \hat{e}_t$ for $t = m+1, ..., m+n_1$ where there are n_1 1-step ahead forecast residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ available, often with $m = 0$ and $n_1 = n$. In the examples, we get the formulas $\hat{Y}_n(h)$, and then replace n by t so that the test data formula is applied to the training data. Then the general formula for an ARMA(p, q) model is $\hat{Y}_t(h) = \hat{\tau} + \hat{\phi}_1 \hat{Y}_t(h-1) + \hat{\phi}_2 \hat{Y}_t(h-2) + \cdots + \hat{\phi}_{h-1} \hat{Y}_t(1) + \hat{\phi}_h Y_t + \cdots + \hat{\phi}_p Y_{t+h-p} + \hat{\theta}_h \hat{e}_t + \cdots + \hat{\theta}_q \hat{e}_{t+h-q}$ for $1 < h \leq min(p, q)$. Assume there are n_h forecast residuals $\hat{e}_t(h)$ available from the training data.

Example 3.2. a) Consider a moving average MA(2) = ARMA(0,2) = ARIMA(0,0,2) = ARIMA(0,0,2)$\times(0,0,0)_1$ model: $Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t$. Suppose data $Y_1, ..., Y_n$ from this model is available. The *R* software produces $\hat{e}_t$ and $\hat{Y}_t = Y_t - \hat{e}_t$ for $t = 1, ..., n$ where $\hat{Y}_t = \hat{Y}_{t-1}(1) = \hat{\tau} + \hat{\theta}_1 \hat{e}_{t-1} + \hat{\theta}_2 \hat{e}_{t-2}$ and $\hat{e}_t(1) = Y_{t+1} - \hat{Y}_t(1)$ for $t = 3, ..., n$. Also, $\hat{Y}_n(1) = \hat{\tau} + \hat{\theta}_1 \hat{e}_n + \hat{\theta}_2 \hat{e}_{n-1}$. Hence there are $n_1 = n$ 1-step ahead forecast residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ available. Similarly, $\hat{Y}_t(2) = \hat{\tau} + \hat{\theta}_2 \hat{e}_t$ for $t = 1, ..., n$. Hence the 2-step ahead forecast residuals are available for $t = 3, ..., n-2$. Now $\hat{Y}_t(h) = \hat{\tau} \approx \overline{Y}$ for $h > 2$. Hence there are n h-step ahead forecast residuals $Y_t - \overline{Y}$ for $h > 2$ and

$t = 1, ..., n$.

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

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

When $d > 0$ for an ARIMA(p, d, q) model, often $\tau = 0$.

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

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

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

The basic idea for getting prediction intervals for the test data is now given. Find the formulas for the test data $Y_{n+1}, ..., Y_{n+L}$, apply the formulas to the training data $Y_1, ..., Y_n$ to get forecast residuals. Apply the shorth to the n_h forecast residuals $\hat{e}_t(h)$ to get $[L_n(h), U_n(h)]$. Then the PI for Y_{n+h} is $[\hat{Y}_n(h) + L_n(h), \hat{Y}_n(h) + U_n(h)]$.

Often time series PIs assume normality and are similar to equation (3.2) below. The following normal PI is often used, but typically does not work well unless the h-step ahead forecast is normally distributed. For many time series models, a large sample normal $100(1-\delta)\%$ PI for Y_{t+h} is

$$[L_n, U_n] = \hat{Y}_t(h) \pm t_{1-\delta/2, n-p-q} SE(\hat{Y}_t(h)). \tag{3.2}$$

Suppose that as $n \to \infty$, $\hat{Y}_t(h) \to E(Y_{t+h}) = \mu_{t+h}$ and $SE(\hat{Y}_t(h)) \to SD(Y_{t+h}) = \sigma_{t+h}$. These quantities are conditional on the past, but the conditioning is suppressed. Then $P(Y_{t+h} \in [L_n, U_n]) \approx P(Y_{t+h} \in [\mu_{t+h} - z_{1-\delta/2}\sigma_{t+h}, \mu_{t+h} + z_{1-\delta/2}\sigma_{t+h}]) =$

$P[|Y_{t+h} - \mu_{t+h}| < z_{1-\delta/2}\sigma_{t+h}]$ " $\geq$ " $1 - \frac{1}{z^2_{1-\delta/2}}$ assuming Chebyshev's inequality holds to a good approximation. Hence a 95% PI could have coverage as low as 75% and a 99.7% PI could have coverage as low as 89%. If n is large, a 95% PI uses $t_{1-\delta/2,n-p-q} \approx 1.96$ while using $z_{1-\delta/2} = 5$ has coverage that is eventually bounded below by 96% as $n \to \infty$. The t cutoff tends to be too low while the Chebyshev cutoff tends to be too high.

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

$$[L_n, U_n] = [\overline{Y} + b_n L_n(h), \overline{Y} + b_n U_n(h)] \tag{3.3}$$

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

The following PI is new and takes into account the time series structure of the data. A similar idea in Masters (1995, p. 305) is to find the n_h h-step ahead forecast residuals and use percentiles to make PIs for Y_{t+h} for $h = 1, ..., L$. For ARIMA(p, d, q) models, let $c_2 = \lceil n_h(1 - \delta_n)\rceil$ and compute shorth(c_2) $= [L_n(h), U_n(h)]$ of the h-step ahead forecast residuals $\hat{e}_t(h)$. Let $a_h = \left(1 + \frac{15}{n_h}\right)\sqrt{\frac{n_h}{n_h - p - q}}$. Then a large sample $100(1-\delta)$% PI for Y_{t+h} is

$$[L_n, U_n] = [\hat{Y}_n(h) + a_h L_n(h), \hat{Y}_n(h) + a_h U_n(h)] \tag{3.4}$$

where $1 - \delta_n = \min(1 - \delta + 0.05, 1 - \delta + (p+q)/n_h)$ for $\delta > 0.1$ and $1 - \delta_n = \min(1 - \delta/2, 1 - \delta + 10(p+q)\delta/n_h)$ for $\delta \leq 0.1$. The correction factor helps compensate for undercoverage when $n_h \geq 20(p+q)$, and similar correction factors are used in Olive (2007, 2017b, 2018) and Pelawa Watagoda and Olive (2021b) to create prediction intervals for regression models and prediction regions for multivariate regression models. Note that for $h = 1$, an estimator for $\sigma^2 = V(e)$ is

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

suggesting that

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

Figure 3.1 shows a simulated MA(2) time series with $n = 100$, $L = 7$ and $U(-1, 1)$ errors. The horizontal lines correspond to the 95% PI (3.3). Two of the one hundred time series training data observations $Y_1, ...Y_{100}$ lie outside of the two lines. All seven of the future test data observations $Y_{101}, ..., Y_{107}$ lie within their large sample 95% PI (3.3).

Figure 3.1. PIs for an MA(2) Time Series with Uniform$(-1, 1)$ Errors

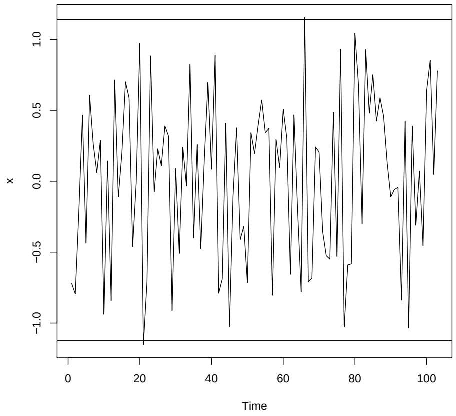


Table 3.1. Normal Errors

δ	n	PI	h=1	h=2	h=3	h=4	h=5	h=6	h=7
0.05	100	N	0.9354	0.9428	0.9526	0.9456	0.9496	0.9410	0.9442
0.05	100		3.900	4.087	4.214	4.214	4.214	4.214	4.214
0.05	100	A	0.9520	0.9652	0.9586	0.9518	0.9576	0.9510	0.9530
0.05	100		4.329	4.746	4.480	4.480	4.480	4.480	4.480
0.05	400	N	0.9444	0.9444	0.9506	0.9466	0.9536	0.9522	0.9442
0.05	400		3.913	4.077	4.182	4.182	4.182	4.182	4.182
0.05	400	A	0.9444	0.9480	0.9468	0.9464	0.9512	0.9460	0.9478
0.05	400		3.980	4.192	4.209	4.209	4.209	4.209	4.209
0.5	100	N	0.4888	0.4968	0.5004	0.4856	0.4966	0.4914	0.4948
0.5	100		1.326	1.388	1.431	1.431	1.431	1.431	1.431
0.5	100	A	0.5100	0.5162	0.5004	0.4898	0.4998	0.4892	0.4926
0.5	100		1.459	1.533	1.496	1.496	1.496	1.496	1.496
0.5	400	N	0.4940	0.49304	0.5028	0.5100	0.4884	0.4858	0.4924
0.5	400		1.344	1.399	1.435	1.435	1.435	1.435	1.435
0.5	400	A	0.4906	0.4902	0.4894	0.5020	0.4816	0.4808	0.4800
0.5	400		1.356	1.413	1.432	1.432	1.432	1.432	1.432

Table 3.2. t_5 Errors

δ	n	PI	h=1	h=2	h=3	h=4	h=5	h=6	h=7
0.05	100	N	0.9408	0.9492	0.9400	0.9412	0.9406	0.9390	0.9376
0.05	100		5.010	5.244	5.409	5.409	5.409	5.409	5.409
0.05	100	A	0.9542	0.9670	0.9484	0.9480	0.9466	0.9462	0.9472
0.05	100		5.665	6.320	5.753	5.753	5.753	5.753	5.753
0.05	400	N	0.9396	0.9494	0.9514	0.9502	0.9480	0.9482	0.9512
0.05	400		5.041	5.257	5.388	5.388	5.388	5.388	5.388
0.05	400	A	0.9448	0.9564	0.9488	0.9514	0.9476	0.9492	0.9494
0.05	400		5.192	5.484	5.456	5.456	5.456	5.456	5.456
0.5	100	N	0.5414	0.5452	0.5464	0.5508	0.5494	0.5606	0.5508
0.5	100		1.709	1.788	1.845	1.845	1.845	1.845	1.845
0.5	100	A	0.4988	0.5158	0.4888	0.4970	0.4978	0.5078	0.5008
0.5	100		1.603	1.710	1.679	1.679	1.679	1.679	1.679
0.5	400	N	0.5678	0.5690	0.5584	0.5650	0.5736	0.5518	0.5652
0.5	400		1.732	1.803	1.850	1.850	1.850	1.850	1.850
0.5	400	A	0.4884	0.5030	0.4956	0.4992	0.5026	0.4820	0.4930
0.5	400		1.468	1.561	1.597	1.597	1.597	1.597	1.597

Table 3.3. Uniform Errors

δ	n	PI	h=1	h=2	h=3	h=4	h=5	h=6	h=7
0.05	100	N	0.9902	0.9812	0.9822	0.9822	0.9850	0.9836	0.9816
0.05	100		2.256	2.361	2.435	2.435	2.435	2.435	2.435
0.05	100	A	0.9848	0.9848	0.9796	0.9708	0.9772	0.9726	0.9780
0.05	100		2.170	2.399	2.388	2.388	2.388	2.388	2.388
0.05	400	N	0.9994	0.9890	0.9816	0.9818	0.9812	0.9848	0.9864
0.05	400		2.263	2.357	2.416	2.416	2.416	2.416	2.416
0.05	400	A	0.9604	0.9590	0.9524	0.9504	0.9532	0.9546	0.9554
0.05	400		1.954	2.141	2.208	2.208	2.208	2.208	2.208
0.5	100	N	0.3752	0.3926	0.4078	0.4130	0.4214	0.4132	0.4128
0.5	100		0.769	0.806	0.831	0.831	0.831	0.831	0.831
0.5	100	A	0.4826	0.4876	0.4724	0.4692	0.4884	0.4780	0.4784
0.5	100		1.002	1.036	1.005	1.005	1.005	1.005	1.005
0.5	400	N	0.3908	0.3962	0.4158	0.4146	0.4216	0.4208	0.4064
0.5	400		0.777	0.809	0.830	0.830	0.830	0.830	0.830
0.5	400	A	0.4934	0.4796	0.4838	0.4840	0.4920	0.4890	0.4730
0.5	400		0.963	0.978	0.978	0.978	0.978	0.978	0.978

Table 3.4. EXP(1) - 1 Errors

δ	n	PI	h=1	h=2	h=3	h=4	h=5	h=6	h=7
0.05	100	N	0.9424	0.9516	0.9478	0.9458	0.9432	0.9422	0.9446
0.05	100		3.872	4.052	4.177	4.177	4.177	4.177	4.177
0.05	100	A	0.9590	0.9712	0.9562	0.9512	0.9492	0.9492	0.9550
0.05	100		3.726	4.389	4.047	4.047	4.047	4.047	4.047
0.05	400	N	0.9556	0.9442	0.9496	0.9446	0.9458	0.9414	0.9486
0.05	400		3.908	4.072	4.177	4.177	4.177	4.177	4.177
0.05	400	A	0.9598	0.9540	0.9496	0.9472	0.9462	0.9434	0.9504
0.05	400		3.224	3.689	3.8093	3.809	3.809	3.809	3.809
0.5	100	N	0.5250	0.5418	0.5528	0.5516	0.5620	0.5494	0.5546
0.5	100		1.323	1.382	1.425	1.425	1.425	1.425	1.425
0.5	100	A	0.5070	0.5068	0.5018	0.4920	0.4956	0.5012	0.5012
0.5	100		0.901	1.023	1.029	1.029	1.029	1.029	1.029
0.5	400	N	0.5358	0.5618	0.5620	0.5550	0.5604	0.5454	0.5568
0.5	400		1.342	1.397	1.432	1.432	1.432	1.432	1.432
0.5	400	A	0.5004	0.5042	0.4984	0.5028	0.4934	0.4842	0.4974
0.5	400		0.760	0.905	0.970	0.970	0.970	0.970	0.970

The simulations used the MA(2) model where the distribution of the white noise $\{e_t\}$ is N(0,1), t_5, $U(-1, 1)$ or (EXP(1) - 1). All these distributions have mean 0, but the fourth distribution is not symmetric. The simulation generates 5000 time series of length $n + L$ and PIs are found for $Y_{n+1}, ..., Y_{n+L}$. The simulations used $L = 7$ and 95% and 50% nominal PIs. The PIs used were the normal PI (3.2) and the alternative PI which uses PI (3.3) for Y_{t+h} where $h > 2$ and PI (3.4) for $h = 1, 2$. These PIs are denoted by N and A respectively in the tables. The simulated coverages and average lengths of the PI are shown.

With 5000 runs, coverages between 0.94 and 0.96 suggest that there is no reason to believe that the nominal coverage is not 0.95, while coverages between 0.48 and 0.52 suggest that there is no reason to believe that the nominal coverage is not 0.5.

From table 3.1 for normal errors, note that for $n = 100$, the coverages of PIs (3.3) and (3.4) were very similar to the those of PI (3.2). PIs (3.3) and (3.4) were longer than the normal PI (3.2) for $n = 100$ and normal errors. From table 3.2 for t_5 errors, the 95% normal PI (3.2) worked well, but the nominal 50% normal PI (3.2) had coverage that was too high and the average lengths were too large. The alternative PIs had coverage near 50% with shorter average lengths. From table 3.3 for uniform errors, the normal PIs (3.2) were too long and the coverage was too high for 95% PIs. The alternative PIs (3.3) and (3.4) had coverage closer to the nominal level with good coverage for $n = 100$. From table 3.4 with EXP(1) - 1 errors, for 95% PIs the normal PIs (3.2) were longer than the alternative PIs (3.3) and (3.4). For the 50% PIs, the normal PIs (3.2) were too long with coverage that was too high. The alternative PIs (3.3) and (3.4) were shorter with good coverage.

Next we consider time series PIs after variable selection. Simulations used the 1-step ahead PI with $d = 1$ for ease of programming. Let the full model be the ARMA(p_{max}, q_{max}) model. Let I_{min} be the ARMA(p_m, q_m) model that minimized a criterion such as AIC, AIC_C, or BIC. Find $\hat{Y}_n(h)$ and the forecast residuals $\hat{e}_t(h)$ for the selected model I_{min}. For $d = 1$ we will use the residuals $\hat{e}_t$. Let $k = p_m + q_m + 1$ and $\tilde{e}_t(h) = \sqrt{\frac{n}{n-k}}\hat{e}_t(h)$. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + k/n_h)$ for $\delta > 0.1$ and

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

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

$$c_{mod} = \min(n_h, \lceil n_h[q_n + 1.12\sqrt{\delta/n_h}\ ]\rceil). \tag{3.5}$$

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

$$[L_n, U_n] = [\hat{Y}_t(h) + \hat{L}_n(h), \hat{Y}_t(h) + \hat{U}_n(h)]. \tag{3.6}$$

Similar correction factors were used by Olive, Rathnayake, and Haile (2021) for prediction intervals for regression models, such as generalized linear models, after variable selection.

Why might PIs (3.3), (3.4), and (3.6) have good coverage? For both the test data and the training data, $Y_{t+h} = \hat{Y}_t(h) + \hat{e}_t(h) = \mu_{t+h} + e_t(h)$. First, consider the training data where J forecast residuals $\hat{e}_t(h)$ exist. Then the proportion of $Y_{t+h} \in [\hat{Y}_t(h) + L_n(h), \hat{Y}_t(h) + U_n(h)]$ = the proportion of the J forecast residuals $\hat{e}_t(h) \in [L_n(h), U_n(h)] \approx 1 - \delta_n \geq 1 - \delta$ by construction. Hence the training data coverage is good. If the selected fitted model is good, and the test data behaves like the training data, then we expect the test data coverage to be good. Hence we need consistent estimators and large n.

Second, assume the time series follow a weakly stationary ARMA model, and suppose $\hat{Y}_t(h)$ is a consistent estimator of μ_{t+h} and $\hat{e}_t(h)$ estimates $e_t(h)$ in that $\hat{e}_t(h) - e_t(h) \overset{D}{\to} 0$ as $n \to \infty$. Also assume that the percentiles of $\hat{e}_t(h)$ estimate the percentiles of $e_t(h)$ such that $P(e_t(h) \in [L_n(h), U_n(h)]) \to 1 - \delta$ as $n \to \infty$. Then $P(Y_{n+h} \in [\hat{Y}_n(h) + L_n(h), \hat{Y}_n(h) + U_n(h)]) \approx P(e_t(h) \in [L_n(h), U_n(h)]) \approx 1 - \delta$. These assumptions are roughly the assumptions made when normality is assumed, which makes the time series strictly stationary. For $h = 1$, the $\{\hat{e}_{t+1}\} = \{\hat{e}_t(1)\}$ estimate the iid $\{e_t\}$, and these assumptions may be reasonable if consistent estimators are used and n is large. For weakly stationary ARMA models, $\mu_{t+h} \to \mu$, $\hat{Y}_t(y) \to \mu$, and $\hat{e}_t(h)$ estimates $Y_{t+h} - \mu$ as $h \to \infty$. Lee and Scholtes (2014) discuss when the percentiles of forecast errors are consistent for ARMA models. For the MA(q) model, $e_t(h) = \theta_1 e_{t+h-1} + \theta_2 e_{t+h-2} + \cdots + \theta_{h-1} e_{t+1} + e_{t+h}$ for $h \leq q$,

$e_t(h) = Y_{t+h} - \mu$ for $h > q$, the $e_t(h)$ are identically distributed for fixed h, and the random variables $e_j(h), e_{j+h}(h), e_{j+2(h)}(h), ...$ are iid for fixed $h \leq q$. For $h \leq q$, there are h iid sequences starting at $j = 1, 2, ..., h$, respectively. For $h > q$ there are $q+1$ iid sequences starting at $j = 1, ..., (q+1)$. Since the sample percentiles of the iid sequences converge in probability to the population percentiles for fixed h, so do the sample percentiles of all of the data. Hence $P(e_t(h) \in [L_n(h), U_n(h)]) \approx 1 - \delta$ as $n \to \infty$ for the MA(q) model if consistent estimators are used. A weakly stationary, causal ARMA(p, q) time series follows an MA(∞) model which is approximately an MA(K) time series where K depends on the time series but not on n. Such time series tend to be ergodic: see White (1984, p. 46). For ergodic data from a unimodal distribution, Chen and Shao (1999) proved the sample shorth converges to the unique population shorth.

If the variable selection estimator is based on a consistent estimator and the probability that the variable selection estimator underfits goes to 0 as $n \to \infty$, then the variable selection estimator is consistent. For example, use the Yule Walker estimator and AIC for AR(p) variable selection. We recommend using PI (3.3) if $n_h < 50$. Tables 3.5 and 3.6 show simulation results, where PI 3.2 is the 95% normal PI, PI 3.3 is the 95% PI that ignores the time series structure of the data, PI 3.4 is the 95% PI that considers the time series structure of the data and a new large sample PI (PI 3.6) for Y_{n+h}. The coverages were high for uniform error types.

The following quote from Hyndman and Athanasopoulos (2018, last paragraph of $\oint$8.8) is important. *"As with most prediction interval calculations, ARIMA-based intervals tend to be too narrow. This occurs because only the variation in the errors has been accounted for. There is also variation in the parameter estimates, and in the model order, that has not been included in the calculation. In addition, the calculation assumes that the historical patterns that have been modelled will continue into the forecast period."* Also see Bhansali (1981) for the effects of estimating the order of the time series model.

There is a large literature on time series PIs, especially for AR(p) models. The bootstrap is often used. See Alonso, Peńa, and Romo (2002, 2003), Brockwell and Davis (2016), Clements and Kim (2007), deLuna (2000), Hyndman and Athanasopoulos (2018), Kabaila and He (2007), Masters (1995, p. 305), Pan and Politis (2016a), Pascual, Romo, and Ruiz (2001), Thombs and Schucany (1990), Vidoni (2009), and Wolf and Wunderli (2015) for references.

Some papers on the shorth include Chen and Shao (1999), Grübel (1988), and Einmahl and Mason (1992).

Table 3.5. one step PI after model selection, MA(2) is true

n	dist	PI 3.3	PI 3.6	PI F	PI 3.2
100	N	0.9582	0.9592	0.9442	0.9476
100		4.4553	4.3214	3.8857	3.9341
100	t5	0.9504	0.9550	0.9412	0.9434
100		5.7340	5.6747	5.0015	5.06377
100	U	0.9728	0.9776	0.9842	0.9860
100		2.3876	2.1992	2.2538	2.2819
100	sExp	0.9536	0.9540	0.9406	0.9424
100		4.0179	3.7989	3.8504	3.8983
400	N	0.9458	0.9500	0.9470	0.9476
400		4.2054	3.9990	3.9119	3.9239
400	t5	0.9432	0.9444	0.9404	0.9412
400		5.4640	5.2364	5.0455	5.0609
400	U	0.9518	0.9576	0.9988	0.9992
400		2.2084	1.9644	2.2593	2.2662
400	sExp	0.9558	0.9578	0.9508	0.9518
400		3.8057	3.2935	3.9047	3.9166
800	N	0.9516	0.9526	0.9514	0.9520
800		4.1704	3.9445	3.9147	3.9206
800	t5	0.9458	0.9480	0.9452	0.9456
800		5.4334	5.1604	5.0491	5.0568
800	U	0.9500	0.9524	0.9994	0.9994
800		2.1838	1.9255	2.2605	2.2640
800	sExp	0.9438	0.9438	0.9410	0.9410
800		3.7821	3.1842	3.9147	3.9207

Table 3.6. one step PI after model selection, AR(1) is true

n	dist	PI 3.3	PI 3.6	PI F	PI 3.2
100	N	0.9548	0.9562	0.9412	0.9436
100		4.3870	4.2758	3.8770	3.9250
100	t5	0.9486	0.9502	0.9402	0.9434
100		5.6495	5.5786	4.9980	5.0597
100	U	0.9744	0.9828	0.9904	0.9916
100		2.3104	2.1587	2.2479	2.2758
100	sExp	0.9556	0.9620	0.9482	0.9492
100		3.8550	3.6475	3.8582	3.90609
400	N	0.9502	0.9506	0.9490	0.9494
400		4.1343	3.9811	3.9103	3.9222
400	t5	0.9442	0.9452	0.9432	0.9440
400		5.3588	5.1925	5.0425	5.0579
400	U	0.9614	0.9616	0.9990	0.9990
400		2.1382	1.9554	2.2603	2.2672
400	sExp	0.9518	0.9504	0.9452	0.9456
400		3.6109	3.2302	3.9048	3.9167
800	N	0.9504	0.9480	0.9490	0.9494
800		4.1063	3.9413	3.9203	3.9262
800	t5	0.9462	0.9512	0.9474	0.9478
800		5.3277	5.1354	5.0384	5.0461
800	U	0.9584	0.9616	0.9998	0.9998
800		2.1148	1.9218	2.2610	2.2645
800	sExp	0.9502	0.9552	0.9484	0.9484
800		3.5783	3.1373	3.9129	3.9189

CHAPTER 4

PREDICTION INTERVALS AND REGIONS FOR THE RANDOM WALK

Now consider a random walk (with drift) $Y_t = Y_{t-1} + e_t$ where the e_t are iid. Suppose there is a sample $Y_1, ..., Y_n$ and we want a PI for Y_{n+j}. Then $Y_t = Y_{t-2} + e_{t-1} + e_t = Y_{t-j} + e_{t-j+1} + \cdots + e_t = Y_0 + e_1 + \cdots + e_t$, or $Y_{n+j} = Y_n + e_{n+1} + e_{n+2} + \cdots + e_{n+j} = Y_n + \epsilon_{n,j}$. Let $e_j = Y_j - Y_{j-1}$ for $j = 2, ..., n$. Divide $e_2, ..., e_n$ into blocks of length j and let ϵ_i be the sum of the e_i in each block. Hence $\epsilon_1 = e_2 + \cdots + e_{j+1}$, $\epsilon_2 = e_{j+2} + \cdots + e_{2j+1}$, and $\epsilon_i = e_{(i-1)j+2} + e_{(i-1)j+3} + \cdots + e_{(i-1)j+j+1}$ for $i = 1, ..., m = \lfloor n/j \rfloor$. These ϵ_i are iid from the same distribution as $\epsilon_{n,j}$. Assume $n \geq 50j$ and let $[L, U]$ be the shorth(c) PI for a future value of ϵ_f based on $\epsilon_1, ..., \epsilon_m$ with $m \geq 50$. Then the large sample $100(1 - \delta)\%$ PI for Y_{n+j} is $[Y_n + L, Y_n + U]$. Note that $\epsilon_j = \epsilon_{n,j} \approx N(j\mu, j\sigma^2)$ for large j by the central limit theorem if $E(e_t) = \mu$ and $V(e_t) = \sigma^2$.

The random walk can be written as $Y_t = Y_0 + \sum_{i=1}^t e_i$ where $Y_0 = y_0$ is often a constant. Pankratz (1983, p. 106) notes that the random walk model has been found to be a good model for many stock price time series. A stochastic process $\{N(t) : t \geq 0\}$ is a counting process if $N(t)$ counts the total number of events that occurred in time interval $(0, t]$. Let e_n be the interarrival time or waiting time between the $(n - 1)$th and nth events counted by the process, $n \geq 1$. If the nonnegative e_i are iid, then $\{N(t), t \geq 0\}$ is a *renewal process*. Let $Y_n = \sum_{i=1}^n e_i$ = the time of occurrence of the nth event = waiting time until the nth event. Then Y_n is a random walk with $Y_0 = y_0 = 0$. Let $E(e_i) = \mu > 0$. Then $E(Y_n) = n\mu$ and $V(Y_n) = nV(e_i)$ if $V(e_i)$ exists. A Poisson process with rate λ is a renewal process where the e_i are iid EXP(λ) with $E(e_i) = 1/\lambda$. See Ross (2014) for the Poisson process and renewal process. Given $Y_1, ..., Y_n$, then n evnts have occurred, and the 1-step ahead PI is for the time until the next event, the 2-step ahead PI is for the time until the next 2 events, and the d-step ahead PI is for the time for the next d events.

The R code below gives the h-step ahead 95% PI for the time until the next h events for $h = 1, 2, 3$ and 4 if the e_i are iid EXP(1) with $n = 1000000$, which corresponds to a Poisson process with $\lambda = 1$. The 1-step ahead large sample 96% PI is [0.000,3.003] with length 3.003.

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

times<-rexp(1000000)

renewalpi(times)

$onepi

[1] 9.639189e-08 3.002906

$twopi

[1] 0.04208788 4.77489907

$threepi

[1] 0.3227718 6.4326904

$fourpi

[1] 0.6858191 7.9398378
```

Some sample output for 100 runs is shown below. The coverage and average length of the h-step ahead 95% PIs is computed for $h = 1, 2, 3$ and 4.

```
rwpisim(n=10000,nruns=100,type=2,tdf=1)

$onepimnlen   #C(1,1)

[1] 26.62644  #25.41

$onecov

[1] 0.94

$twopimnlen   #C(2,2)

[1] 54.35798  #50.82

$twocov

[1] 0.93

$threepimnlen #C(3,3)

[1] 82.4097   #76.24

$threecov

[1] 0.94

$fourpimnlen #C(4,4)
```

```
[1] 111.4    #101.65

$fourcov

[1] 0.95
```

Table 4.1. Random walk PI, nruns=5000,δ=0.05

n	dist	h=1	h=2	h=3	h=4
100	N	0.9554	0.9618	0.9432	0.9208
100		4.1675	6.32058	7.2125	7.7710
100	C	0.9594	0.9602	0.9394	0.9204
100		47.2301	570.4022	578.6495	562.9710
100	EXP	0.9602	0.9588	0.9466	0.9238
100		3.6581	6.2683	7.1028	7.6876
100	U	0.9496	0.9610	0.9436	0.9232
100		1.9027	3.2920	3.9957	4.3782
400	N	0.9576	0.9562	0.9590	0.9578
400		4.0667	5.7800	7.2466	8.3284
400	C	0.9562	0.9556	0.9648	0.9580
400		32.8266	72.3425	133.9639	189.4935
400	EXP	0.9632	0.9576	0.9604	0.9578
400		3.3091	5.1449	6.7579	7.9262
400	U	0.9532	0.9480	0.9554	0.9548
400		1.9035	3.1644	4.0582	4.7017
800	N	0.9466	0.9528	0.9532	0.9568
800		4.0192	5.7505	7.0041	8.1543
800	C	0.9536	0.9576	0.9522	0.9536
800		29.7084	65.2802	98.3195	142.1259
800	EXP	0.9592	0.9594	0.9570	0.9540
800		3.1997	5.0468	6.4145	7.6673
800	U	0.9498	0.9506	0.9546	0.9572
800		1.9013	3.1659	3.9642	4.6309

A small random walk simulation was done for the large sample 95% PIs using 5000 runs with $Y_0 = 1$. So an observed coverage in [0.94, 0.96] gives no reason to doubt that the PI has the nominal coverage of 0.95. The errors e_i were iid from four distributions: i) N(1,1), ii) $t_1 \sim Cauchy(1,1)$, iii) EXP(1), and iv) uniform$(0,2)$. Only distribution iii) is not symmetric. We computed the d-step ahead 95% PIs for $d = 1,2,3,4 = J$. We want $n \geq 50J$, but simulatations may use smaller n such as $n = 25J$. The asymptotic optimal lengths are i) 3.92, 5.54, 6.79, 7.84, ii) 25.41, 50.82, 76.24, 101.65, iii) 3.00, 4.72, 6.11, 7.22, iv) 1.90, 3.11, 3.87, 4.48. The *tspack* function `rwpisim` was used for the simulation.

Let the population forecast error be $e(h)$. For type 1, the asymptotic optimal lengths of the large sample 95% PIs are $3.92\sqrt{h}$ where $e(h) \sim N(h, \sigma^2 = h)$. For type 2, $e(h) \sim C(h, \sigma = h)$: a Cauchy distribution. For type 3, $e(h) \sim G(h, 1)$: a Gamma distribution. For type 4, $e(2) \sim$ triangular(0,4). The distribution of the sum of n iid U(0,1) random variables is known as the Irwin-Hall distribution. See Gray and Odell (1966), Marengo, Farnsworth, and Stefanic (2017), and Roach (1963).

If $\boldsymbol{Y}_t = \boldsymbol{Y}_{t-1} + \boldsymbol{e}_t$, use the same idea but apply the Olive (2017b) prediction regions. To describe these prediction regions, Mahalanobis distances will be useful. Let the $g \times 1$ column vector T be a multivariate location estimator, and let the $g \times g$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = D_{\boldsymbol{w}_i}^2(T, \boldsymbol{C}) = (\boldsymbol{w}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{w}_i - T) \tag{4.1}$$

for each observation $\boldsymbol{w}_i$, where $i = 1, ..., n$. Notice that the Euclidean distance of $\boldsymbol{w}_i$ from the estimate of center T is $D_i(T, \boldsymbol{I}_g)$ where $\boldsymbol{I}_g$ is the $g \times g$ identity matrix. The classical Mahalanobis distance D_i uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{w}}, \boldsymbol{S})$, the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{w}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{w}_i \text{ and } \boldsymbol{S} = \frac{1}{\mathrm{n}-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{w}})(\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{w}})^{\mathrm{T}}. \tag{4.2}$$

Consider predicting a future test value $\boldsymbol{w}_f$, given past training data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ with $J = n$ where $\boldsymbol{w}_1, ..., \boldsymbol{w}_n, z_f$ are iid. Prediction intervals are a special case of prediction regions with $g = 1$ so the $\boldsymbol{w}_i$ are random variables.

Definition 4.1. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{w}_f \in \mathcal{A}_n) \geq 1-\delta$ asymptotically. A prediction region is *asymptotically optimal* if its volume converges in probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{w}_f$.

Like prediction intervals, prediction regions need correction factors. For iid data from a distribution with a $g \times g$ nonsingular covariance matrix, it was found that the simulated maximum undercoverage of prediction region (4.4) without the correction factor was about 0.05 when $n = 20g$. Hence the correction factor (4.3) is used to give better coverage for small n. Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{4.3}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Let $D_{(U_n)}$ be the $100q_n$th sample quantile of the D_i where $i = 1, ..., n$.

Definition 4.2. The large sample $100(1-\delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{w}_f$ given iid data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}, \tag{4.4}$$

while the large sample $100(1-\delta)\%$ *classical prediction region* is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) \leq \chi^2_{g,1-\delta}\}. \tag{4.5}$$

For the classical prediction region, see Chew (1966) and Johnson and Wichern (1988, pp. 134, 151). The nonparametric prediction region is due to Olive (2013). Also see Olive (2017b: pp. 151-153, 2018). The classical prediction region is a large sample prediction region if the iid $\boldsymbol{w}_i$ are iid $N_g(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is nonsingular. The nonparametric prediction region is a large sample prediction region if the iid $\boldsymbol{w}_i$ have a nonsingular covariance matrix, and is asymptotically optimal for a large class of elliptically contoured distribution, including multivariate normal distributions with nonsingular covariance matrices. Regions with smaller asymptotic volumes can exist if the distribution is not elliptically contoured. From Olive (2018, p. 161), simulated coverage was often near the nominal for $n \geq 20g$, but simulated volumes behaved better for $n \geq 50g$. Figure 5.1 shows population 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of Y_f for two multivariate normal distributions.

Some prediction intervals for stochastic processes include Pan and Politis (2016b), Vidoni (2004), and Vit (1973). Mykland (2003) describes how to convert prediction regions into investment strategies. Tables 4.2-4.7 show simulation results for the random walk prediction region. Prediction regions do not have lengths, where as prediction intervals should have lengths as well as coverages.The coverages were good for $h = 1$ and 2, and became good for $h = 3$ and 4 as the sample size n increased.

Table 4.2. Random walk PR, nruns=5000, p=2

n	ψ	type	h=1	h=2	h=3	h=4
100	0	1	0.9434	0.9314	0.9208	0.9014
100	0	2	0.9412	0.9370	0.9296	0.9100
100	0	3	0.9478	0.9446	0.9378	0.9196
100	0	4	0.9394	0.9258	0.9202	0.8964
100	0.707	1	0.9470	0.9436	0.9302	0.9096
100	0.707	2	0.9430	0.9372	0.9282	0.9122
100	0.707	3	0.9476	0.9480	0.9388	0.9176
100	0.707	4	0.9450	0.9308	0.9208	0.9016
100	0.9	1	0.9420	0.9386	0.9258	0.9106
100	0.9	2	0.9432	0.9388	0.9242	0.9116
100	0.9	3	0.9458	0.9474	0.9388	0.9182
100	0.9	4	0.9396	0.9356	0.9206	0.8996
200	0	1	0.9458	0.9420	0.9416	0.9412
200	0	2	0.9546	0.9468	0.9408	0.9416
200	0	3	0.9568	0.9444	0.9424	0.9412
200	0	4	0.9450	0.9410	0.9316	0.9324
200	0.707	1	0.9444	0.9372	0.9346	0.9350
200	0.707	2	0.9472	0.9458	0.9410	0.9388
200	0.707	3	0.9482	0.9482	0.9488	0.9472
200	0.707	4	0.9404	0.9386	0.9340	0.9374
200	0.9	1	0.9478	0.9436	0.9378	0.9352
200	0.9	2	0.9498	0.9486	0.9450	0.9450
200	0.9	3	0.9506	0.9514	0.9482	0.9460
200	0.9	4	0.9416	0.9358	0.9374	0.9344

Table 4.3. Random walk PR, nruns=5000, p=2

n	ψ	type	h=1	h=2	h=3	h=4
400	0	1	0.9440	0.9464	0.9428	0.9456
400	0	2	0.9490	0.9442	0.9408	0.9374
400	0	3	0.9546	0.9516	0.9472	0.9458
400	0	4	0.9474	0.9438	0.9396	0.9382
400	0.707	1	0.9476	0.9460	0.9486	0.9482
400	0.707	2	0.9508	0.9468	0.9450	0.9492
400	0.707	3	0.9472	0.9512	0.9490	0.9482
400	0.707	4	0.9464	0.9442	0.9426	0.9456
400	0.9	1	0.9482	0.9518	0.9486	0.9426
400	0.9	2	0.9464	0.9472	0.9496	0.9420
400	0.9	3	0.9512	0.9500	0.9500	0.9450
400	0.9	4	0.9510	0.9426	0.9418	0.9412

Table 4.4. Random walk PR, nruns=5000, p=4

n	ψ	type	h=1	h=2	h=3	h=4
200	0	1	0.9464	0.9460	0.9342	0.9172
200	0	2	0.9498	0.9454	0.9378	0.9216
200	0	3	0.9440	0.9414	0.9384	0.9220
200	0	4	0.9444	0.9384	0.9304	0.9116
200	0.5	1	0.9448	0.9424	0.9294	0.9124
200	0.5	2	0.9450	0.9500	0.9408	0.9192
200	0.5	3	0.9452	0.9478	0.9436	0.9304
200	0.5	4	0.9422	0.9474	0.9308	0.9084
200	0.9	1	0.9554	0.9476	0.9380	0.9186
200	0.9	2	0.9510	0.9476	0.9430	0.9190
200	0.9	3	0.9480	0.9494	0.9436	0.9294
200	0.9	4	0.9442	0.9350	0.9294	0.9098
400	0	1	0.9532	0.9516	0.9508	0.9496
400	0	2	0.9492	0.9492	0.9468	0.9442
400	0	3	0.9500	0.9438	0.9494	0.9500
400	0	4	0.9452	0.9420	0.9378	0.9412
400	0.5	1	0.9476	0.9468	0.9480	0.9376
400	0.5	2	0.9458	0.9490	0.9454	0.9454
400	0.5	3	0.9494	0.9450	0.9456	0.9466
400	0.5	4	0.9444	0.9490	0.9438	0.9352
400	0.9	1	0.9534	0.9498	0.9516	0.9496
400	0.9	2	0.9514	0.9488	0.9494	0.9458
400	0.9	3	0.9566	0.9554	0.9534	0.9520
400	0.9	4	0.9438	0.9418	0.9378	0.9418

Table 4.5. nruns=5000, p=4

n	ψ	type	h=1	h=2	h=3	h=4
400	0.9	4	0.9438	0.9418	0.9378	0.9418
800	0	1	0.9522	0.9486	0.9462	0.9434
800	0	2	0.9450	0.9452	0.9464	0.9472
800	0	3	0.9524	0.9540	0.9526	0.9524
800	0	4	0.9478	0.9456	0.9490	0.9498
800	0.5	1	0.9468	0.9452	0.9478	0.9482
800	0.5	2	0.9450	0.9478	0.9560	0.9514
800	0.5	3	0.9500	0.9482	0.9480	0.9420
800	0.5	4	0.9472	0.9502	0.9502	0.9464
800	0.9	1	0.9426	0.9480	0.9504	0.9514
800	0.9	2	0.9460	0.9474	0.9458	0.9432
800	0.9	3	0.9488	0.9510	0.9490	0.9472
800	0.9	4	0.9490	0.9480	0.9448	0.9480

Table 4.6. Random walk PR, nruns=5000, p=8

n	ψ	type	h=1	h=2	h=3	h=4
400	0	1	0.9426	0.9438	0.9370	0.9214
400	0	2	0.9490	0.9502	0.9444	0.9270
400	0	3	0.9466	0.9530	0.9476	0.9392
400	0	4	0.9416	0.9446	0.9388	0.9216
400	0.354	1	0.9514	0.9446	0.9456	0.9186
400	0.354	2	0.9450	0.9572	0.9460	0.9290
400	0.354	3	0.9556	0.9546	0.9496	0.9314
400	0.354	4	0.9416	0.9412	0.9340	0.9182
400	0.9	1	0.9484	0.9462	0.9424	0.9198
400	0.9	2	0.9524	0.9502	0.9480	0.9310
400	0.9	3	0.9482	0.9576	0.9546	0.9392
400	0.9	4	0.9458	0.9376	0.9346	0.9228
800	0	1	0.9458	0.9450	0.9460	0.9484
800	0	2	0.9516	0.9554	0.9514	0.9506
800	0	3	0.9494	0.9508	0.9480	0.9544
800	0	4	0.9432	0.9408	0.9438	0.9418
800	0.354	1	0.9456	0.9464	0.9478	0.9450
800	0.354	2	0.9474	0.9550	0.9540	0.9488
800	0.354	3	0.9534	0.9516	0.9532	0.9536
800	0.354	4	0.9494	0.9466	0.9480	0.9518
800	0.9	1	0.9436	0.9482	0.9478	0.9450
800	0.9	2	0.9500	0.9494	0.9512	0.9514
800	0.9	3	0.9552	0.9520	0.9514	0.9484
800	0.9	4	0.9474	0.9450	0.9494	0.9464

Table 4.7. Random walk PR, nruns=5000, p=8

n	ψ	type	h=1	h=2	h=3	h=4
1600	0	1	0.9506	0.9516	0.9476	0.9464
1600	0	2	0.9522	0.9534	0.9532	0.9514
1600	0	3	0.9496	0.9530	0.9524	0.9522
1600	0	4	0.9418	0.9428	0.9414	0.9430
1600	0.354	1	0.9506	0.9472	0.9504	0.9502
1600	0.354	2	0.9440	0.9520	0.9488	0.9502
1600	0.354	3	0.9506	0.9572	0.9574	0.9570
1600	0.354	4	0.9488	0.9418	0.9444	0.9462
1600	0.9	1	0.9510	0.9496	0.9476	0.9458
1600	0.9	2	0.9492	0.9500	0.9532	0.9474
1600	0.9	3	0.9524	0.9558	0.9548	0.9540
1600	0.9	4	0.9450	0.9508	0.9452	0.9500

CHAPTER 5

THE BOOTSTRAP

This chapter follows Olive (2022, ch. 2) closely, and argues that, under regularity conditions, applying the nonparametric prediction region of chapter 4 to a bootstrap sample results in a confidence region. The volume of a confidence region → 0 as $n \to 0$, while the volume of a prediction region goes to that of a population region that would contain a new $\boldsymbol{w}_f$ with probability $1 - \delta$. The nominal coverage is $100(1 - \delta)$. If the actual coverage $100(1 - \alpha_n) > 100(1 - \delta)$, then the region is *conservative*. If $100(1 - \alpha_n) < 100(1 - \delta)$, then the region is *liberal*. A region that is 5% conservative is considered "much better" than a region that is 5% liberal.

When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations ($2SD(\overline{Y}_n) = 2\sigma/\sqrt{n}$) of $\theta = \mu$ is about 95%. Hence the probability that θ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\theta - 1.96S/\sqrt{n}, \theta + 1.96S/\sqrt{n}]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if θ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}]$ is a large sample 95% confidence interval for the population mean θ. Note that the lengths of the two intervals are the same. Where the interval is centered, at the parameter θ or the statistic $\overline{Y}_n$, determines whether the interval is a prediction or a confidence interval. See Theorem 5.2 for a similar relationship between confidence regions and prediction regions. We often want $P(\boldsymbol{\theta} \in \mathcal{A}_n) \to 1 - \delta$ as $n \to \infty$.

Definition 5.1. A *large sample* $100(1 - \delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

There are several methods for obtaining a bootstrap sample $T_1^*, ..., T_B^*$ where the sample size n is suppressed: $T_i^* = T_{in}^*$. The parametric bootstrap, nonparametric bootstrap, and residual bootstrap will be discussed in this Section.

When $g = 1$, a confidence interval is a special case of a confidence region. Again we often want the probability to converge to $1 - \delta$ if the confidence interval is based on a statistic with an asymptotic distribution that has a probability density function.

Definition 5.2. The interval $[L_n, U_n]$ is a large sample $100(1-\delta)\%$ *confidence interval* for θ if $P(L_n \le \theta \le U_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

Using the notation from Section 3 for the shorth PI, let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1-\delta/2) \rceil$. Then a common nonparametric large sample $100(1-\delta)\%$ PI for Z_f is

$$[Z_{(k_1)}, Z_{(k_2)}] \tag{5.1}$$

where $0 < \delta < 1$. See Frey (2013) for references.

Next we discuss bootstrap confidence intervals (5.2) and (5.3) that are obtained by applying prediction intervals (5.1) and (3.1) to the bootstrap sample with B used instead of n. Some additional bootstrap CIs are obtained from bootstrap confidence regions from Section 5.2 when $g = 1$. See Efron (1982) and Chen (2016) for the percentile method CI. Let T_n be an estimator of a parameter θ such as $T_n = \overline{Z} = \sum_{i=1}^n Z_i/n$ with $\theta = E(Z_1)$. Let $T_1^*, ..., T_B^*$ be a bootstrap sample for T_n. Let $T_{(1)}^*, ..., T_{(B)}^*$ be the order statistics of the the bootstrap sample.

Definition 5.3. The bootstrap percentile method large sample $100(1-\delta)\%$ confidence interval for θ is an interval $[T^*_{(k_L)}, T^*_{(K_U)}]$ containing $\approx \lceil B(1-\delta) \rceil$ of the T_i^*. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. A common choice is

$$[T^*_{(k_1)}, T^*_{(k_2)}]. \tag{5.2}$$

Definition 5.4. The large sample $100(1-\delta)\%$ *shorth*(c) *CI*

$$[T^*_{(s)}, T^*_{(s+c-1)}] \tag{5.3}$$

uses the interval $[T^*_{(1)}, T^*_{(c)}], [T^*_{(2)}, T^*_{(c+1)}], ..., [T^*_{(B-c+1)}, T^*_{(B)}]$ of shortest length. Here

$$c = \min(B, \lceil B[1-\delta + 1.12\sqrt{\delta/B}\,]\,\rceil). \tag{5.4}$$

The shorth CI can be regarded as the shortest percentile method confidence interval, asymptotically. Hence the shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples. See Remark 5.1 for some theory for bootstrap CIs such as (5.2) and (5.3). Olive (2014: p. 238, 2017b: p. 168, 2018) recommended using the shorth CI for the percentile CI.

5.1 THE NONPARAMETRIC BOOTSTRAP

This section illustrates the nonparametric bootstrap with some examples. Suppose a statistic T_n is computed from a data set of n cases. The nonparametric bootstrap draws n cases with replacement from that data set. Then T_1^* is the statistic T_n computed from the sample. This process is repeated B times to produce the bootstrap sample $T_1^*, ..., T_B^*$. Sampling cases with replacement uses the empirical distribution.

Definition 5.5. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 5.1. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 5.5. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \dfrac{n-1}{n}\boldsymbol{S} \equiv \dfrac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^n \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\text{Cov}(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \ \square$$

Example 5.2. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function (widely used in robust statistics). The *nonparametric bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 5.3. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

The bootstrap has been widely used to estimate the population covariance matrix of the statistic Cov(T_n), for testing hypotheses, and for obtaining confidence regions (often confidence intervals). An iid sample $T_{1n}, ..., T_{Bn}$ of size B of the statistic would be very useful for inference, but typically we only have one sample of data and one value $T_n = T_{1n}$ of the statistic, where n is often suppressed. Often $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, and the bootstrap sample $T^*_{1n}, ..., T^*_{Bn}$ is formed where $T^*_{jn} = t(\boldsymbol{w}^*_{j1}, ..., \boldsymbol{w}^*_{jn})$. Results summarized in Remark 5.2 imply that $T^*_{1n} - T_n, ..., T^*_{Bn} - T_n$ is pseudodata for $T_{1n} - \boldsymbol{\theta}, ..., T_{Bn} - \boldsymbol{\theta}$ when n is large in that $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T^* - T_n) \xrightarrow{D} \boldsymbol{u}$.

Suppose there is a statistic T_n that is a $g \times 1$ vector. Let

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T^*_i \text{ and } \boldsymbol{S}^*_T = \frac{1}{B-1}\sum_{i=1}^{B}(T^*_i - \overline{T}^*)(T^*_i - \overline{T}^*)^T \tag{5.5}$$

be the sample mean and sample covariance matrix of the bootstrap sample $T^*_1, ..., T^*_B$ where $T^*_i = T^*_{i,n}$. Fix n, and let $E(T^*_{i,n}) = \boldsymbol{\theta}_n$ and Cov($T^*_{i,n}$) = $\boldsymbol{\Sigma}_n$.

We will often assume that Cov(T_n) = $\boldsymbol{\Sigma}_T$, and $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ where $\boldsymbol{\Sigma}_A > 0$ is positive definite and nonsingular. Often $n\hat{\boldsymbol{\Sigma}}_T \xrightarrow{P} \boldsymbol{\Sigma}_A$. For example, using least squares and the residual bootstrap for the multiple linear regression model, $\boldsymbol{\Sigma}_n = \frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, $T_n = \boldsymbol{\theta}_n = \hat{\boldsymbol{\beta}}$,

$\boldsymbol{\theta} = \boldsymbol{\beta}$, $\hat{\boldsymbol{\Sigma}}_T = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\boldsymbol{\Sigma}_A = \sigma^2 \lim_{n\to\infty}(\boldsymbol{X}^T\boldsymbol{X}/n)^{-1}$. See Example 7.1.

5.2 BOOTSTRAP CONFIDENCE REGIONS FOR HYPOTHESIS TESTING

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for a $g\times 1$ parameter vector $\boldsymbol{\theta}$ is a set $\mathcal{A}_n = \mathcal{A}_{n,B}$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_{n,B})$ is eventually bounded below by $1-\delta$ as $n, B \to \infty$. The B is often suppressed. 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. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$.

For a confidence region, let the $g\times 1$ vector T_n be an estimator of the $g\times 1$ parameter vector $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $\boldsymbol{A}$ be a full rank $g\times p$ constant matrix. For variable selection, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \boldsymbol{0}$. 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 VS or MIX. See chapter 6. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$. See Equation (5.5). A useful fact for the F and chi-square distributions is $d_n F_{g,d_n,1-\delta} \to \chi^2_{g,1-\delta}$ as $d_n \to \infty$. Here $P(X \leq \chi^2_{g,1-\delta}) = 1 - \delta$ if $X \sim \chi^2_g$, and $P(X \leq F_{g,d_n,1-\delta}) = 1 - \delta$ if $X \sim F_{g,d_n}$. Let $k_B = \lceil B(1 - \delta)\rceil$. Confidence region (5.6) needs $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A > 0$ as $n, B \to \infty$. See Machado and Parente (2005) for regularity conditions for this assumption.

Definition 5.6. a) The standard bootstrap large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{1-\delta}\} \tag{5.6}$$

where $D^2_{1-\delta} = \chi^2_{g,1-\delta}$ or $D^2_{1-\delta} = d_n F_{g,d_n,1-\delta}$ where $d_n \to \infty$ as $n \to \infty$. b) The Bickel and Ren (2001) large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\hat{\boldsymbol{\Sigma}}_A/n]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{(k_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_A/n) \leq D^2_{(k_B,T)}\} \tag{5.7}$$

where the cutoff $D^2_{(k_B,T)}$ is the $100k_B$th sample quantile of the
$D_i^2 = (T_i^* - T_n)^T[\hat{\boldsymbol{\Sigma}}_A/n]^{-1}(T_i^* - T_n) = n(T_i^* - T_n)^T[\hat{\boldsymbol{\Sigma}}_A]^{-1}(T_i^* - T_n)$.

The following three confidence regions will be used for inference after variable selection. The Olive (2017ab, 2018) prediction region method applies prediction region (4.4) to the bootstrap sample. Olive (2017ab, 2018) also gave the modified Bickel and Ren confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$. The hybrid confidence region is due to Pelawa Watagoda and Olive (2021a). For prediction region (4.4), the correction factor (4.3) was used to give better coverage for small n. When applied to a bootstrap sample of size B, the correction factor (5.8) gives better coverage when $B \geq 50g$. This result is useful because the bootstrap confidence regions can be slow to simulate. Hence we want to use small values of $B \geq 50g$. Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \text{ otherwise.} \tag{5.8}$$

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 quantile of the D_i. If B is large enough, $D_{(U_B)}$ is the $100(1 - \delta)$th quantile.

Definition 5.7. a) The prediction region method large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} \tag{5.9}$$

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 $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. (This procedure is basically the one sample Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi^2_{g,1-\delta}$ cutoff by $D^2_{(U_B)}$.) b) The modified Bickel and Ren (2001) large sample $100(1 - \delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\} \tag{5.10}$$

where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if

$(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}^*_T]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$.

c) Shift region (5.9) to have center T_n, or equivalently, change the cutoff of region (5.10) to $D^2_{(U_B)}$ to get the hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w}-T_n)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w}-T_n) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}^*_T) \leq D^2_{(U_B)}\}. \tag{5.11}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if
$(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}^*_T]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Hyperellipsoids (5.9) and (5.11) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (5.9) and (5.10) is

$$\frac{|\boldsymbol{S}^*_T|^{1/2}}{|\boldsymbol{S}^*_T|^{1/2}} \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g. \tag{5.12}$$

The volume of confidence region (5.10) tends to be greater than that of (5.9) since the T^*_i are closer to $\overline{T}^*$ than T_n on average.

If $g = 1$, then a hyperellipsoid is an interval, and confidence intervals are special cases of confidence regions. Suppose the parameter of interest is θ, and there is a bootstrap sample $T^*_1, ..., T^*_B$ where the statistic T_n is an estimator of θ based on a sample of size n. The percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T^*_i. Let $a_i = |T^*_i - \overline{T}^*|$. Let $\overline{T}^*$ and S^{2*}_T be the sample mean and variance of the T^*_i. Then the squared Mahalanobis distance $D^2_\theta = (\theta - \overline{T}^*)^2 / S^{*2}_T \leq D^2_{(U_B)}$ is equivalent to $\theta \in [\overline{T}^* - S^*_T D_{(U_B)}, \overline{T}^* + S^*_T D_{(U_B)}] = [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$, which is an interval centered at $\overline{T}^*$ just long enough to cover U_B of the T^*_i. Hence the prediction region method is a special case of the percentile method if $g = 1$. See Definition 5.3. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The CI $[T_n - a_{(U_B,T)}, T_n + a_{(U_B,T)}]$ corresponding to (5.10) is defined similarly, and $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$ is the CI for (24). Note that the three CIs corresponding to (5.9)–(5.11) can be computed without finding S^*_T or $D_{(U_B)}$ even if $S^*_T = 0$. The shorth(c) CI (5.3) computed from the T^*_i can be much shorter than the Efron (2014) or prediction region method confidence intervals.

Remark 5.1. Under regularity conditions, Olive (2017b, 2018) proved that (5.9) is a large sample confidence region. See Bickel and Ren (2001) for (5.10), while Pelawa Watagoda and Olive (2021a) gave simpler proofs, and proved that the shorth(c) interval applied to a bootstrap sample of a random variable gives a large sample confidence interval. If $g = 1$, if $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$, and if $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$ where U has a unimodal probability density function symmetric about zero, then the confidence intervals from the three confidence regions (5.9)–(5.11), the shorth confidence interval (5.3), and the "usual" percentile method confidence interval (5.2) are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically).

Remark 5.2. Note that if (5.10) is a large sample confidence regions, then so are (5.9) and (5.11) if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$ as $n \to \infty$. Pelawa Watagoda and Olive (2021a) showed that this condition holds if $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ where $E(\boldsymbol{u}) = \boldsymbol{0}$ and Cov$(\boldsymbol{u}) = \boldsymbol{\Sigma_u} \neq \boldsymbol{0}$. Thus $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. In addition, assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{C}$ 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. If n and B are large enough and $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned," then the confidence regions (5.9), (5.10), and (5.11) should still have coverage near $1 - \delta$. The regularity conditions for (5.9)–(5.11) are weaker when $g = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

Remark 5.3. For bootstrapping the $m \times 1$ vector $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$, we will often want $n \geq 20m$ and $B \geq \max(100, n, 50m)$. If T_n is $g \times 1$, we might replace m by g or replace m by df if df is the model degrees of freedom. Sometimes much larger n is needed to avoid undercoverage. We want $B \geq 50g$ so that $\boldsymbol{S}_T^*$ is a good estimator of $Cov(T_n^*)$. Prediction region theory uses correction factors like (3.1) and (4.3) to compensate for finite n. The bootstrap confidence regions (5.9)–(5.11) and the shorth CI use the correction factors (5.8) and (5.4) to compensate for finite $B \geq 50g$. Note that the correction factors make the volume of the confidence region larger as B decreases. Hence a test with larger B will have more power.

5.3 LARGE SAMPLE THEORY FOR VARIABLE SELECTION ESTIMATORS

This section gives the large sample theory for $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$, and follows Rathnayake and Olive (2020) closely. Pelawa Watagoda and Olive (2021ab) gave theory for $\hat{\boldsymbol{\beta}}_{MIX}$ and $\hat{\boldsymbol{\beta}}_{VS}$ for the multiple linear regression model, and Rathnayke and Olive (2021) extended the theory to many other models, including GLMs, some time series models, and some survival regression models.

Assume 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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \tag{5.13}$$

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

Theorem 5.1 for $\hat{\boldsymbol{\beta}}_{MIX}$, due to Rathnayake and Olive (2021), generalizes the Pelawa Watagoda and Olive (2021b) theorem for multiple linear regression, and is useful for understanding Theorem 5.3 for $\hat{\boldsymbol{\beta}}_{VS}$. The first assumption in Theorem 5.1 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the variable selection estimator corresponding to I_{min} underfits with probability going to zero, and the assumption holds under regularity conditions for GLMs if BIC or AIC is used. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). For multiple linear regression with Mallows (1973) C_p or AIC, see Li (1987), Nishii (1984), and Shao (1993). For AR(p) variable selection with AIC, AIC_C, and BIC, see Hannan and Quinn (1979) and Shibata (1976). For MA(q) and ARMA(p, q) variable selection, the assumption has perhaps not yet been proved. However, the condition is necessary for the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$ to be a consistent estimator of $\boldsymbol{\beta}$. See Rathnayake and Olive (2021). The assumption on $\boldsymbol{u}_{jn}$ in Theorem 5.1

is reasonable by (5.13) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection.

Theorem 5.1. *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* $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0})$. *a) Then*

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{5.14}$$

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* $g \times p$ *full rank matrix with* $1 \le g \le p$. *Then*

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{5.15}$$

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

d) If $\pi_a = 1$, *then* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{a,0})$ *where* SEL *is* VS *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 $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ 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 J 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). □

The following subscript notation is useful. Subscripts before the MIX are used for subsets of $\hat{\boldsymbol{\beta}}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. 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 $\hat{\boldsymbol{\beta}}_{VS}$. The subscript 0 is still used for zero padding. We may use $FULL$ to denote the full model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$.

Theorem 5.1 has several other applications. First, the theory gives the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{MIX}$ corresponding to many variable selection estimators. Second, the theory is useful for explaining why $\hat{\boldsymbol{\beta}}_{I_{min}}$ should not be used, but $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ is a good estimator. For a random quantity to be a $k \times 1$ random vector, the dimension of the random quantity needs to be k (with probability one). Since the dimension of $\hat{\boldsymbol{\beta}}_{I_{min}}$ is a random variable, the random quantity $\hat{\boldsymbol{\beta}}_{I_{min}}$ is neither a random vector nor a statistic. Then $\hat{\boldsymbol{\beta}}_{I_{min}}$ is not a consistent estimator for any parameter vector $\boldsymbol{\beta}_{I_j}$, and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min}} - \boldsymbol{\beta}_{I_j})$ can not be used as an asymptotic pivot even if $I_{min} = I_j$ is observed. Compare Leeb and Pötscher (2006). A third application is bootstrap inference for hypothesis testing. Fourth, the theory can be used to justify prediction intervals after variable selection. See chapter 8, Pelawa Watagoda and Olive (2021b) and Olive, Rathnayake, and Haile (2021).

The following Pelawa Watagoda and Olive (2021a) theorem is useful for bootstrapping variable 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. Let $D^2_{(U_B)}$ be the cutoff computed from the $D^2_i(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. Note that $T_i = T_{in}$. The hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(T_n, \boldsymbol{C})$ is centered at T_n, while the hyperellipsoid corresponding to $D^2(\overline{T}, \boldsymbol{C})$ is centered at $\overline{T}$. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = (\overline{T} - T_n)^T \boldsymbol{C}^{-1}(\overline{T} - T_n) = (T_n - \overline{T})^T \boldsymbol{C}^{-1}(T_n - \overline{T}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \le D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \le D^2_{(U_B)}$. In Theorem 5.2, 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 ("not too ill conditioned") for each $n \ge 20g$, say. This condition is weaker than the stated assumption $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$ where $\boldsymbol{\Sigma}_A$ is some nonsingular matrix. Often $\boldsymbol{\Sigma}_A = \lim_{n\to\infty} n\boldsymbol{\Sigma}_{T_n}$.

Theorem 5.2: Geometric Argument. *Suppose* $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ *with* $E(\boldsymbol{u}) = \boldsymbol{0}$ *and* $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} \ne \boldsymbol{0}$. *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* $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \le D^2_{(U_B)}\}$ *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* $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \le D^2_{(U_B)}\}$ *is a large sample* $100(1 - \delta)\%$ *confidence region for* $\boldsymbol{\theta}$ *where* T_n *is a randomly selected* T_i.

Proof. The region R_c centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$

which is eventually bounded below by $1-\delta$ as $B \to \infty$. Since the $\sqrt{n}(T_i - \boldsymbol{\theta})$ are iid,

$$\begin{bmatrix} \sqrt{n}(T_1 - \boldsymbol{\theta}) \\ \vdots \\ \sqrt{n}(T_B - \boldsymbol{\theta}) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. For fixed B, the average of these random vectors is

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma_u}}{B}\right)$$

where AN_g denotes an approximate multivariate normal distribution. Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Thus R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. □

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 (5.11) is a confidence region by the geometric argument (as is region (5.10) which tends to use a larger cutoff), and (5.9) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$.

Let the random selection estimator $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$. Here $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$. We have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (5.15) where $E(\boldsymbol{v}) = \boldsymbol{0}$, and $\boldsymbol{\Sigma_v} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. Hence the above geometric argument holds: 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 (5.9) to (5.11). This condition holds only under strong regularity conditions such as $\pi_a = 1$. Chapter 7 will explain why the bootstrap confidence regions are still useful.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Let $W = W_{VS} = k$ if $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ where $P(W_{VS} = k) = \pi_{kn}$ for $k = 1, ..., J$. Then $(\hat{\boldsymbol{\beta}}_{VS:n}, W_{VS:n}) = (\hat{\boldsymbol{\beta}}_{VS}, W_{VS})$ has a joint distribution where the sample size n is usually suppressed. Note that $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_W,0}$. 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 $\hat{\boldsymbol{\beta}}_{I_k,0}|(W_{VS} = k)$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(W_{VS} = k) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \le t_1, ..., z_p \le t_p)$ by $P(\boldsymbol{z} \le \boldsymbol{t})$. Then

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]\pi_{kn}$$

$$= \sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}]\pi_{kn} = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

Hence $\hat{\boldsymbol{\beta}}_{VS}$ 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}.

Charkhi and Claeskens (2018) showed that $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$ if $S \subseteq I_j$ for the MLE with AIC, and gave a forward selection example. Here $\boldsymbol{w}_j$ is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\boldsymbol{0}$. Hence $E(\boldsymbol{w}_j) = 0$, and $\text{Cov}(\boldsymbol{w}_j) = \boldsymbol{\Sigma}_j$ exits. Note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ 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 $\boldsymbol{u}_{jn}$ to that of $\boldsymbol{w}_{jn}$. Similarly, selection bias does change the distribution of the selected $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_j$. Let $W = W_{VS,\infty}$ where $P(W = k) = \pi_k$. In Theorem 5.1, the assumption $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j$ is a mild assumption that is made for large sample tests for whether a reduced model I_j is good. The reasonable Theorem 5.3 assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild. Regularity conditions for the $\boldsymbol{w}_j$ to have $E(\boldsymbol{w}_j) = 0$ may be strong. The proof for Equation (5.16) is the same as that for (5.14). Theorem 5.3 is due to Rathnayake and Olive (2021), and Pelawa Watagoda and Olive (2021b) have a similar theorem for multiple linear regression.

Theorem 5.3. *Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ 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*

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{5.16}$$

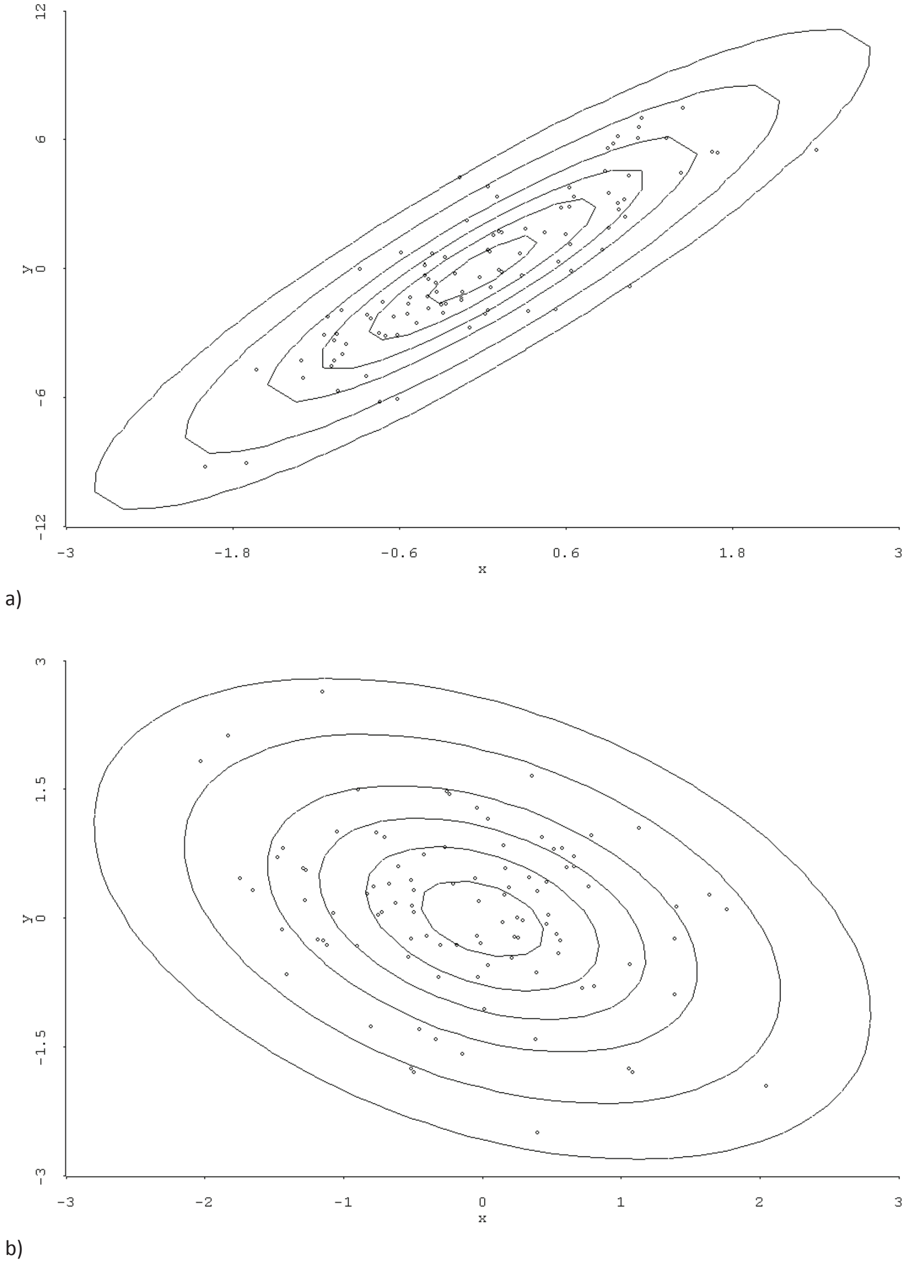
where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$. Thus $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

The estimator $\hat{\boldsymbol{\beta}}_{MIX}$ is the random vector with selection probabilities $P(W_{MIX} = k) = \pi_{kn} = P(W_{VS} = k)$ where W_{MIX} is independent of the $\hat{\boldsymbol{\beta}}_{I_k,0}$. Simulating $\hat{\boldsymbol{\beta}}_{MIX}$ and $\hat{\boldsymbol{\beta}}_{VS}$ is informative. We consider $\boldsymbol{X} = \boldsymbol{X}_n$ fixed or condition on $\boldsymbol{X}_n$. The probabilities π_{kn} depend on $\boldsymbol{X}_n$, n, p, the variable selection estimator, and the population model that generates $\boldsymbol{Y}$. Consider the experiment of generating $\boldsymbol{Y}$ from the model. (For example, i) for a parametric regression model, generate $Y_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{\gamma})$ for $i = 1, ..., n$ to form $\boldsymbol{Y}$, and ii) for multiple linear regression, generate $\boldsymbol{e}$ from the population of $\boldsymbol{e}$, and form $\boldsymbol{Y} = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{e}$.) Then regress $\boldsymbol{Y}$ on $\boldsymbol{X}_n$ with variable selection to generate $(\hat{\boldsymbol{\beta}}_{VS}, W_{VS})$. Generate another $\boldsymbol{Y}$ from the model, and generate $(\hat{\boldsymbol{\beta}}_{MIX}, W_{VS})$. Then $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$whenever $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}^C_{I_k,0}$, but I_k was chosen for $\hat{\boldsymbol{\beta}}_{MIX}$ before generating the new $\boldsymbol{Y}$, so there is no selection bias. Repeat to get the sample $(\hat{\boldsymbol{\beta}}^C_{I_{k_1},0}, \hat{\boldsymbol{\beta}}_{I_{k_1},0}), ..., (\hat{\boldsymbol{\beta}}^C_{I_{k_B},0}, \hat{\boldsymbol{\beta}}_{I_{k_B},0})$.

The hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(T_n, \boldsymbol{C})$ is centered at T_n, while the hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(\overline{T}, \boldsymbol{C})$ is centered at $\overline{T}$. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = (\overline{T} - T_n)^T\boldsymbol{C}^{-1}(\overline{T} - T_n) = (T_n - \overline{T})^T\boldsymbol{C}^{-1}(T_n - \overline{T}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \le D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \le D^2_{(U_B)}$.

The prediction region method will often simulate well even if B is rather small. If the ellipses are centered at T_n or $\overline{T}^*$, Figure 5.1 shows confidence regions if the plotted points are $T^*_1, ..., T^*_B$ where the T^*_i are approximately multivariate normal. If the ellipses are centered at $\overline{T}$, Figure 5.1 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of T_f for two multivariate normal statistics. Then the plotted points are iid $T_1, ..., T_B$. If $nCov(T) \xrightarrow{P} \boldsymbol{\Sigma}_A$, and the

Figure 5.1. Confidence Regions for 2 Statistics with MVN Distributions



T_i^* are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. By Theorem 5.2, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near T_n and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths of this covering region are about one tenth of those in Figure 2. Hence when T_n falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If T_n is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

Hence B does not need to be large provided that n and B are large enough so that $S_T^* \approx \text{Cov}(T^*) \approx \boldsymbol{\Sigma}_A/n$. However, we need $B \to \infty$ for the coverage to go to $1-\delta$. There is undercoverage for finite B. Using (5.8) increases the training data coverage and hence reduces undercoverage. The price to pay is that the prediction region method confidence region is inflated to have better coverage, so the power of the hypothesis test is decreased if moderate B is used instead of larger B. If n is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \geq Jg$ where $J = 20$ or 50. For small g, using $B = 1000$ often led to good simulations for GLMs and multiple linear regression, but $B = \max(50g, 100)$ may work well.

CHAPTER 6

SOME LARGE SAMPLE THEORY FOR TIME SERIES

In the this chapter and chapter 8, assume that the AR(p), MA(q), and ARMA(p, q) models are weakly stationary, causal, and invertible. Let the ARIMA(p, d, q) models have known d and apply the results to the weakly stationary, causal, and invertible ARMA model from the differenced time series. Such a time series has both an AR(∞) and MA(∞) representation where the magnitude of the parameters decreases to zero rapidly. We will use different formulas for $\boldsymbol{\beta}$ as in chapter 2.

Large sample theory is useful for the bootstrap. The estimator in Theorem 6.1 is sometimes used to estimate σ_e^2. See Granger and Newbold (1977, p. 85) and Pankratz (1983, p. 206).

Remark 6.1. We often use the phrase “under regularity conditions” if we have not found a clear statement of the regularity conditions. In the theorem below, we can replace $n - p - q$ by $n - c$ where, for example, $c = 0$ or $p + q + 1$.

Theorem 6.1. Let $Y_1, ..., Y_n$ be an ARMA time series with $V(E_t) = \sigma_e^2$, and let the r_i be the (one step ahead) residuals. Under regularity conditions,

$$\tilde{\sigma}^2 = \frac{\sum r_i^2}{n - p - q} \tag{6.1}$$

is a consistent estimator of σ_e^2.

Let $Y_1, ..., Y_n$ be an AR(p) time series with $\gamma_k = Cov(Y_t, Y_{t-k})$. Let

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

To describe the least squares model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ for an AR(p) time series let $\phi_0 = \tau$ and let $Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t$. Let $\boldsymbol{\beta} = (\phi_0, ..., \phi_p)^T$. 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 $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ 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 $p+1$. Then the least squares estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. By Theorem 6.2 a), $\hat{\boldsymbol{\beta}} \approx N_{p+1}(\boldsymbol{\beta}, \hat{\sigma}_e^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. So tests from ordinary multiple linear regression can be applied to AR(p) time series, and $SE(\hat{\boldsymbol{\beta}}_i) = \sqrt{\hat{\sigma}_e^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}$.

The least squares estimator can be computed by plugging in sample covariance matrices. Let $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$, and let $\boldsymbol{\beta}^T = (\beta_0, \boldsymbol{\beta}_2^T)$ where β_0 is the intercept and the slopes vector $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{YW} = (\beta_1, ..., \beta_p)^T$. Let the population covariance matrices

$$\text{Cov}(\boldsymbol{u}) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u} - E(\boldsymbol{u}))^T] = \boldsymbol{\Sigma_u}, \text{ and}$$

$$\text{Cov}(\boldsymbol{u}, Y) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(Y - E(Y))] = \boldsymbol{\Sigma_u}_Y.$$

Then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ (even if a linear model does not hold) are

$$\beta_1 = E(Y) - \boldsymbol{\beta}_2^T E(\boldsymbol{u}) \text{ and } \boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}\mathrm{Y}}.$$

Let the sample covariance matrices be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T \text{ and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y}).$$

Let the method of moments or maximum likelihood estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} =$

$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i Y_i - \overline{\boldsymbol{u}}\,\overline{Y}$. Then it can be shown that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ satisfies $\hat{\beta}_1 = \overline{Y} - \hat{\boldsymbol{\beta}}_S^T\overline{\boldsymbol{u}}$ and

$$\hat{\boldsymbol{\beta}}_2 = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}.$$

The Yule Walker (Durbin Levinson) equations can be written in two equivalent forms using $\rho_k = \gamma_k/\gamma_0$. Let $\boldsymbol{\beta} = \boldsymbol{\phi} = (\phi_1, ..., \phi_p)^T$. Let $\boldsymbol{x} = (Y_{t-1}, Y_{t-2}, ..., Y_{t-p})^T$ and $Y = Y_t$. (Note that $\boldsymbol{x}$ is $\boldsymbol{u}$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_2$ in the above paragraph.) The Yule Walker equations are

$\rho_1 = \phi_1 + \phi_2\rho_1 + \phi_3\rho_2 + \cdots + \phi_p\rho_{p-1}$ or $\gamma_1 = \phi_1\gamma_0 + \phi_2\gamma_1 + \phi_3\gamma_2 + \cdots + \phi_p\gamma_{p-1}$

$\rho_2 = \phi_1\rho_1 + \phi_2 + \phi_3\rho_1 + \cdots + \phi_p\rho_{p-2}$ or $\gamma_2 = \phi_1\gamma_1 + \phi_2 + \phi_3\gamma_1 + \cdots + \phi_p\gamma_{p-2}$

$\vdots$

$\rho_p = \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \phi_3\rho_{p-3} + \cdots + \phi_p$ or $\gamma_p = \phi_1\gamma_{p-1} + \phi_2\gamma_{p-2} + \phi_3\gamma_{p-3} + \cdots + \phi_p\gamma_0$.

In matrix form

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} \quad \text{or} \quad \boldsymbol{\rho}_{\boldsymbol{x},\mathrm{Y}} = \boldsymbol{\rho}_{\boldsymbol{x}}\boldsymbol{\phi}$$

or

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} \quad \text{or} \quad \boldsymbol{\Sigma}_{\boldsymbol{x},\mathrm{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\phi}.$$

Hence $\boldsymbol{\beta} = \boldsymbol{\phi}_{YW} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\rho}_{\boldsymbol{x}}^{-1}\boldsymbol{\rho}_{\boldsymbol{x},Y}$. Then $\hat{\boldsymbol{\phi}}_{YW} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y} = \tilde{\boldsymbol{\rho}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{\rho}}_{\boldsymbol{x},Y}$. Here the estimators are found by replacing ρ_k by $r_k = \hat{\rho}_k$ and by replacing γ_k by $\hat{\gamma}_k$. Plugging in estimators of $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and

$\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is like a method of moments estimator. Note that $\boldsymbol{\rho}_{\boldsymbol{x}}$, $\boldsymbol{\rho}_{\boldsymbol{x},Y}$, $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ have the desired form since $\boldsymbol{x} = (Y_{t-1}, ..., Y_{t-p})^T$ and $Y = Y_t$. Hence $Cov(X_i, Y) = Cov(Y_{t-i}, Y_t) = \gamma_i$, $Cov(X_i, X_j) = Cov(Y_{t-i}, Y_{t-j}) = \gamma_{|i-j|}$, $corr(X_i, X_j) = corr(Y_{t-i}, Y_{t-j}) = \rho_{|i-j|}$, and $corr(X_i, Y) = corr(Y_{t-i}, Y_t) = \rho_i$.

Note that $\boldsymbol{\beta} = (\phi_1, ..., \phi_p)^T = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{YW}$, but $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{YW}$ use different plug in estimators of $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$. If an AR($p_{max}$) model is fit but the true model is AR(p_o) with $p_o \leq p_{max}$, then $\boldsymbol{\beta} = (\phi_1, ..., \phi_{p_o}, 0, ..., 0)^T$. From Theorem 6.2 a), the asymptotic covariance matrix corresponding to $\hat{\boldsymbol{\beta}}$ is $\sigma_e^2 \gamma_{p_{max}}^{-1}$ while the asymptotically efficient covariance matrix corresponding to $\hat{\boldsymbol{\beta}}_{p_o}$ is $\sigma_e^2 \gamma_{p_o}^{-1}$.

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).

Theorem 6.2. 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

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}) \tag{6.2}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}$. Equation (6.2) 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,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_q(\boldsymbol{0}, \boldsymbol{V}). \tag{6.3}$$

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,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{V}). \tag{6.4}$$

The main point of Theorem 6.2 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 6.2a). 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).

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.

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 $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j$ is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}_{j,0}) \tag{6.5}$$

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 6.3 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. For AR model selection, the probability of underfitting goes to 0 if the AIC, BIC, or 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). 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 $\hat{\boldsymbol{\beta}}_{MS}$ to be a consistent estimator of $\boldsymbol{\beta}$. See Rathnayake and Olive (2021). The assumption on $\boldsymbol{u}_{jn}$ in Theorem 6.3 is reasonable by (6.5) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection. The proofs of Theorems 6.3 and 6.4 are exactly as in Rathnayake and Olive (2021).

Theorem 6.3. 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

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{6.6}$$

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

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{6.7}$$

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 $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ 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). □

Theorem 6.3 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 6.3d) is useful for *variable selection consistency* where $\pi_a = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to \infty$. See Claeskens and Hjort (2008) for references.

Theorem 6.4. *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*

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{6.8}$$

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.

CHAPTER 7

BOOTSTRAPPING VARIABLE SELECTION ESTIMATORS

This chapter follows Rathnayake and Olive (2021) closely. Obtaining the bootstrap samples for $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ is simple. Generate $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ that would be used to produce $\hat{\boldsymbol{\beta}}^*$ if the full model estimator $\hat{\boldsymbol{\beta}}$ was being bootstrapped. (The nonparametric bootstrap, parametric bootstrap, and residual bootstrap using the residuals from the full OLS MLR model are discussed below. See chapter 8 for time series results.) Often $\boldsymbol{X}^* = \boldsymbol{X}_n$. Instead of generating $\hat{\boldsymbol{\beta}}^*$, compute the variable selection estimator $\hat{\boldsymbol{\beta}}^*_{VS,1} = \hat{\boldsymbol{\beta}}^{*C}_{I_{k_1},0}$. Then generate another $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ and compute $\hat{\boldsymbol{\beta}}^*_{MIX,1} = \hat{\boldsymbol{\beta}}^*_{I_{k_1},0}$ (using the same subset I_{k_1}). This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$. Note that the $\boldsymbol{\epsilon}^*_i = \hat{\boldsymbol{\beta}}^*_{VS,i} - \hat{\boldsymbol{\beta}}^*_{MIX,i}$ are iid with respect to the bootstrap distribution for $i = 1, .., B$. Let $T = T_n = \hat{\boldsymbol{\beta}}$. Then $\sqrt{n}(\overline{T}^*_{VS} - \overline{T}^*_{MIX}) = \sqrt{n} \sum_{i=1}^B \boldsymbol{\epsilon}^*_i / B$, and the two bagging estimators may be asymptotically equivalent. Let the selection probabilities for the bootstrap variable selection estimator be ρ_{kn}. Then this bootstrap procedure bootstraps both $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ with $\pi_{kn} = \rho_{kn}$.

The key idea is to show that the bootstrap data cloud is slightly more variable than the iid data cloud, so confidence region (5.9) 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 $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} -$

$\hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$. Then by Equation (6.2),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \ \text{ and } \ \sqrt{n}(\hat{\boldsymbol{\beta}}^*_{\mathrm{I_j},0} - \hat{\boldsymbol{\beta}}_{\mathrm{I_j},0}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{V}_{\mathrm{j},0}). \tag{7.1}$$

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 (5.10) 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 $n \geq 20p$), and centering the bootstrap data cloud at T_n results in the confidence region (5.10) having slightly higher asymptotic coverage than applying (5.11) to the iid data cloud. Also, (5.10) tends to have higher coverage than (5.11) since the cutoff for (5.10) tends to be larger than the cutoff for (5.11). Region (5.9) has the same volume as region (5.11), 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. See Breiman (1996) and Yang (2003). 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 Rathnayake and Olive (2021) for $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{\beta}_S = \boldsymbol{\theta}_0$ with $n \geq 20p$, the coverage tended to get close to $1 - \delta$ for $B \geq \max(200, 50p)$ so that $\boldsymbol{S}^*_T$ is a good estimator of $\mathrm{Cov}(T^*)$. 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 $n \geq 20p$ and $B \geq 50p$. Pelawa Watagoda and Olive (2021b) had similar results for multiple linear regression using forward selection with C_p.

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 $c = 0.01$.

A confidence interval $[L_n, U_n]$ can be computed without $\boldsymbol{S}_T^*$ for (5.9), (5.10), and (5.11). 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.

To see the effect of zero padding, consider $H_0 : \boldsymbol{A\beta} = \boldsymbol{\beta}_O = \boldsymbol{0}$ where $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$ and $O \subseteq E$ in (5.4) so that H_0 is true. Suppose a nominal 95% confidence region is used and U_B is the 96th percentile. Hence the confidence region (5.9) or (5.10) covers at least 96% of the bootstrap sample. If $\hat{\boldsymbol{\beta}}^*_{O,j} = \boldsymbol{0}$ for more than 4% of the $\hat{\boldsymbol{\beta}}^*_{O,1}, ..., \hat{\boldsymbol{\beta}}^*_{O,B}$, then $\boldsymbol{0}$ is in the confidence region and the bootstrap test fails to reject H_0. If this occurs for each run in the simulation, then the observed coverage will be 100%.

Now suppose $\hat{\boldsymbol{\beta}}^*_{O,j} = \boldsymbol{0}$ for $j = 1, ..., B$. Then $\boldsymbol{S}_T^*$ is singular, but the singleton set $\{\boldsymbol{0}\}$ is the large sample $100(1 - \delta)\%$ confidence region (5.9), (5.10), or (5.11) for $\boldsymbol{\beta}_O$ and $\delta \in (0, 1)$, and the pvalue for $H_0 : \boldsymbol{\beta}_O = \boldsymbol{0}$ is one. (This result holds since $\{\boldsymbol{0}\}$ contains 100% of the $\hat{\boldsymbol{\beta}}^*_{O,j}$ in the bootstrap sample.) For large sample theory tests, the pvalue estimates the population pvalue. Let I denote the other predictors in the model so $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. For the I_{min} model from variable selection, there may be strong evidence that $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_I$ is in the model if the "100%" confidence region is $\{\boldsymbol{0}\}$, $n \geq 20p$, and $B \geq 50p$. (Since the pvalue is one, this technique may be useful for data snooping: applying regression model theory to submodel I may have negligible selection bias.)

Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and that $S \subseteq I_j$. We want to examine when Equation (7.1) holds or when

$$\text{Cov}(\hat{\boldsymbol{\beta}}^*_I) - \text{Cov}(\hat{\boldsymbol{\beta}}_I) \to \boldsymbol{0} \tag{7.2}$$

as $n, B \to \infty$. Then the component clouds of the iid data for $\hat{\boldsymbol{\beta}}_{MIX}$ and the bootstrap data clouds for $\hat{\boldsymbol{\beta}}^*_{MIX}$ have the same asymptotic variability, and the bootstrap confidence regions may give good results.

For multiple linear regression with the residual bootstrap $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}$ that uses residuals from the full OLS model, Pelawa Watagoda and Olive (2021b) showed that Equation (7.2) holds for OLS variable selection. The nonparametric bootstrap (also called the empirical bootstrap, naive bootstrap, and the pairs bootstrap) draws a sample n cases $(Y_i^*, \boldsymbol{x}_i^*)$ with replacement from the n cases $(Y_i, \boldsymbol{x}_i)$, and regresses the Y_i^* on the $\boldsymbol{x}_i^*$ to get $\hat{\boldsymbol{\beta}}^*_{VS,1}$, and then draw another sample to get $\hat{\boldsymbol{\beta}}^*_{MIX,1}$. This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$. Under regularity conditions, Equation (7.1) holds. See, for example, Freedman (1981). Assumptions for the nonparametric bootstrap tend to be rather strong: often one assumption is that the n cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some population.

Next, consider the parametric regression model $Y_i|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{\gamma})$, and the parametric bootstrap. Suppose $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, and that $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ as $n \to \infty$. These assumptions tend to be mild for a parametric regression model where the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ is used. Then $\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$, the inverse Fisher information matrix. If $\boldsymbol{I}_n(\boldsymbol{\beta})$ is the Fisher information matrix based on a sample of size n, then $\boldsymbol{I}_n(\boldsymbol{\beta})/n \xrightarrow{P} \boldsymbol{I}(\boldsymbol{\beta})$. For GLMs, see, for example, Sen and Singer (1993, p. 309). For the parametric regression model, we regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ where the $n \times 1$ vector $\boldsymbol{Y} = (Y_i)$ and the ith row of the $n \times p$ design matrix $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$.

Remark 7.1. For bootstrap theory, we will use the following result several times. Suppose the estimator has theory $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$ under regularity conditions where $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} V(\boldsymbol{\beta})$ if $\hat{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}$ as $n \to \infty$. Assume the bootstrap model satisfies the theory for fixed n such that $\sqrt{m}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_n) \xrightarrow{D} N_k(\boldsymbol{0}, V(\hat{\boldsymbol{\beta}}_n))$ as $m \to \infty$. Then $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_n) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$ as $n \to \infty$. (Think of using a triangular array.)

The parametric bootstrap uses $\boldsymbol{Y}_j^* = (Y_i^*)$ where $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ for $i = 1, ..., n$ where $D(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is a parametric distribution. Regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., B$. The large sample theory for $\hat{\boldsymbol{\beta}}^*$ is simple by Remark 7.1. Note that if $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$ where $\boldsymbol{b}$ does not depend on n, then $(\boldsymbol{Y}^*, \boldsymbol{X})$ follows the parametric regression model with parameters $(\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{b}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{b}))$. Now fix large integer n_0, and let $\boldsymbol{b} = \hat{\boldsymbol{\beta}}_{n_o}$. Then $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{n_o}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}}_{n_o}))$. Since

$N_p(\mathbf{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}})) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta})) \tag{7.3}$$

as $n \to \infty$.

Now suppose $S \subseteq I$. Without loss of generality, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}(I)^T, \hat{\boldsymbol{\beta}}(O)^T)^T$. Then $(\boldsymbol{Y}, \boldsymbol{X}_I)$ follows the parametric regression model with parameters $(\boldsymbol{\beta}_I, \boldsymbol{\gamma})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}_I))$. Now $(\boldsymbol{Y}^*, \boldsymbol{X}_I)$ only follows the parametric regression model asymptotically, since $\hat{\boldsymbol{\beta}}(O) \neq \mathbf{0}$. However, under regularity conditions, $E(\hat{\boldsymbol{\beta}}_I^*) \approx \hat{\boldsymbol{\beta}}_I$ and Cov$(\hat{\boldsymbol{\beta}}_I^*)-$ Cov$(\hat{\boldsymbol{\beta}}_I) \to \mathbf{0}$ as $n, B \to \infty$. See the following example.

Example 7.1. Consider the multiple linear regression model: $Y_i = \beta_0 + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$ where the random variables e_i are iid with variance $V(e_i) = \sigma_e^2$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume the maximum leverage $\max_{i=1,...,n} \boldsymbol{x}_{iI}^T(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{x}_{iI} \to 0$ in probability as $n \to \infty$ for each I with $S \subseteq I$. For the OLS model with $S \subseteq I$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}_I)$ where $\boldsymbol{V}_I = \sigma_e^2\boldsymbol{W}_I$ and $(\boldsymbol{X}_I^T\boldsymbol{X}_I)/n \xrightarrow{P} \boldsymbol{W}_I^{-1}$ by Theorem 6.2. Assume a constant β_0 is in each submodel. Under mild conditions for a) and c), $\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}_I)$ if $S \subseteq I$. For MLR, this example shows that the residual bootstrap, parametric bootstrap, and nonparametric booststrap for OLS are robust to the unknown error distribution of the iid e_i.

a) Consider the (MLR) parametric bootstrap for the above model with
$\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2\boldsymbol{I}) \sim N_n(\boldsymbol{H}\boldsymbol{Y}, \hat{\sigma}_n^2\boldsymbol{I})$ where **we are not assuming** that the $e_i \sim N(0, \sigma^2)$, and

$$\hat{\sigma}_n^2 = MSE = \frac{1}{n-p-1}\sum_{i=1}^n r_i^2$$

where the residuals are from the full OLS model. Then MSE is a $\sqrt{n}$ consistent estimator of σ_e^2 under mild conditions by Su and Cook (2012). Thus $\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^* \sim N_{a_I}(\hat{\boldsymbol{\beta}}_I, \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1})$ since $E(\hat{\boldsymbol{\beta}}_I^*) = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_I$ because $\boldsymbol{H}\boldsymbol{X}_I = \boldsymbol{X}_I$, and Cov$(\hat{\boldsymbol{\beta}}_I^*) = \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \sim N_{a_I}(\mathbf{0}, n\hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$ if $S \subseteq I$. Note that $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_n + \boldsymbol{e}^*$ where the $e_i^* = e_{i,n}^*$ are iid with $E(e_i^*) = 0$ and $V(e_i^*) = \hat{\sigma}_n^2$.

b) For the (MLR) residual bootstrap using residuals from the full OLS model, $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_n + \boldsymbol{e}^*$ where the $e_i^* = e_{i,n}^*$ are sampled with replacement from the residuals. With respect to the bootstrap distribution, the e_i^* are iid with $E(e_i^*) = \overline{r} = 0$ since a constant is in the model, and $V(e_i^*) = \frac{n-p-1}{n}\hat{\sigma}_n^2$. Pelawa Watagoda and Olive (2021a) showed that $E(\hat{\boldsymbol{\beta}}_I^*) = \hat{\boldsymbol{\beta}}_I$ and $\text{Cov}(\hat{\boldsymbol{\beta}}_I^*) = [(n - p - 1)/n]\hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1} \xrightarrow{P} \boldsymbol{V}_I$ for $S \subseteq I$.

c) For the (MLR) nonparametric bootstrap, the cases $(Y_i, \boldsymbol{x}_i)$ are drawn with replacement from the n cases. For the full model, $\boldsymbol{Y}^* = \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_n + \boldsymbol{e}^*$ where the $(Y_i, \boldsymbol{x}_i, r_i)$ are sampled with replacement. With respect to the bootstrap distribution, the e_i^* are iid with $E(e_i^*) = 0$ and $V(e_i^*) = \frac{n-p-1}{n}\hat{\sigma}_n^2$. The e_i^* depend on $\boldsymbol{x}_i^*$ since $e_i^* = r_j$ if $\boldsymbol{x}_i^* = \boldsymbol{x}_j$. For a submodel I with $S \subseteq I$, $\boldsymbol{Y}^* = \boldsymbol{X}_I^*\hat{\boldsymbol{\beta}}_{I,n} + \boldsymbol{e}_I^*$ where the $(Y_i, \boldsymbol{x}_{i,I}, r_{i,I})$ are sampled with replacement. Freedman (1981) showed that under mild conditions, $n(\boldsymbol{X}_I^{*T}\boldsymbol{X}_I^*)^{-1} \xrightarrow{P} \boldsymbol{W}_I$ when $n(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1} \xrightarrow{P} \boldsymbol{W}_I$, and under stronger regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$ if $S \subseteq I$. Remark 7.1 does not hold since we can not fix n and let $m \to \infty$.

The bootstrap component clouds for $\hat{\boldsymbol{\beta}}_{VS}^*$ are again separated compared to the iid clouds for $\hat{\boldsymbol{\beta}}_{VS}$, which are centered about $\boldsymbol{\beta}$. Heuristically, most of the selection bias is due to predictors in E, not to the predictors in S. Hence $\hat{\boldsymbol{\beta}}_{S,VS}^*$ is roughly randomly selected and similar to $\hat{\boldsymbol{\beta}}_{S,MIX}^*$. Typically the distributions of $\hat{\boldsymbol{\beta}}_{E,VS}^*$ and $\hat{\boldsymbol{\beta}}_{E,MIX}^*$ are not similar, but use the same zero padding. These two results make $\hat{\boldsymbol{\beta}}_{VS}^*$ simulate well.

One problem with the bootstrap methods is that $\boldsymbol{S}_T^*$ can be singular due to one or more columns of zeros in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables 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 $c = 0.01$. Let $\boldsymbol{S}_A^*$ be the covariance matrix from the augmented bootstrap sample. Then apply the confidence regions to the augmented bootstrap sample or plug in $\boldsymbol{S}_A^*$ for $\boldsymbol{S}_T^*$ for the confidence regions from the unaugmented bootstrap sample. Augmentation changes the probabilities ρ_{kn} to $\rho_{kn}/1.01$ except the full model selection probability changes from ρ_{fn} to $0.01 + \rho_{fn}/1.01$. A confidence interval $[L_n, U_n]$ can be computed without $\boldsymbol{S}_T^*$ for (5.9), (5.10), and (5.11). Using the confidence interval $[\max(L_n, T^*_{(1)}), \min(U_n, T^*_{(B)})]$ can give a shorter covering region.

One of the best methods for inference after variable or model selection is “data splitting.” Data splitting uses a training set to find a model, e.g. $I_{min} = I_j$. Then I_j is used as the full model for the validation set, avoiding selection bias so valid inference can be done. See, for example, Rinaldo et al. (2019). For time series, the training set might be the first J cases and the validation set the last $n - J$ cases.

CHAPTER 8

BOOTSTRAPPING TIME SERIES

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 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 (6.1) 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 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(∞) 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 8.1. Assume the time series is such that Theorem 6.2 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 6.2. 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 $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_n) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$ as $n \to \infty$. □

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 8.1 are rather rare in the time series literature. Bühlmann (1994) has such a result for the AR(p) model.

If Equation (7.1.) holds so $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}_{j,0})$, we would like to show that $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j,0} - \hat{\boldsymbol{\beta}}_{I_j,0}) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}_{j,0})$ if I_j was selected with random selection. This result holds for the full model by Theorem 8.1. 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 $\hat{\boldsymbol{\beta}}_{MX}^*$.

Remark 8.1. The above result also holds if the least squares estimator or normal MLE estimator is used since these estimators are consistent. Hence the convergence in the proof holds on a set of probability converging to one.

The AR(p) least squares model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ can be bootstrapped as in Example 7.1, except the models selected are I_j corresponding to the AR($j-1$) model and to the first j columns of $\boldsymbol{X}$ for $j = 1, ..., p = p_{max}$. With respect to the bootstrap distribution, $\boldsymbol{X}$ is a constant matrix, so $\boldsymbol{Y}^*$ follows an MLR model, not an AR(p) model. However, the large sample theory $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_{I_j})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j}^* - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_{I_j})$ from Example 7.1 a) is the same as that given by Theorem 6.2 b). Hence the resulting bootstrap confidence regions should be useful, and may give more precise inference that using to full AR(p_{max}) model.

There is a large literature for bootstrapping time series. Often the bootstrap is used for prediction intervals: find $Y_{n+1}^{*(j)}, ..., Y_{n+L}^{*(j)}$ for $j = 1, ..., B$. Then use percentiles of the $Y_{n+k}^{*(j)}$ to make the prediction interval. We do not recommend using the parametric bootstrap for prediction intervals since typically the iid e_t do not follow a $N(0, \sigma_e^2)$ distribution. Also bootstrap prediction intervals are computationally expensive compared to the new prediction intervals described in this dissertation.

For bootstrapping time series, see, for example, Bühlmann (1997, 2002), Härdle, Horowitz, and Kreiss (2003), Kreiss and Lahiri (2012), Kreiss, Paparoditis, and Politis (2011), Lahiri (2003), Politis (2003).

8.1 SIMULATION

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_{max} and q_{max} were rather small to make the simulation time shorter. For time series, let the full model be the AR(p_{max}), MA(q_{max}), or ARMA(p_{max}, q_{max}) model. Let $k = p_{max} + q_{max}$. Let $\boldsymbol{\beta} = (\phi_1, ..., \phi_{p_{max}})^T$, $\boldsymbol{\beta} = (\theta_1, ..., \theta_{q_{max}})^T$, or $\boldsymbol{\beta} = (\phi_1, ..., \phi_{p_{max}}, \theta_1, ..., \theta_{q_{max}})^T$. 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 is a white noise. Let $I = \{1, ..., p_I, p_{max} + 1, ..., p_{max} + q_I\}$ index the ARMA(p_I, q_I) model. Then $\boldsymbol{\beta}_I = (\phi_1, ..., \phi_{p_I}, \theta_1, ..., \theta_{q_I})^T$ and $\boldsymbol{\beta}_{I,0} = (\phi_1, ..., \phi_{p_I}, 0, ..., 0, \theta_1, ..., \theta_{q_I}, 0, ..., 0)^T$. Let the model selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I,0}$ with probabilities $\pi_{I,n}$ where p_I runs from 0 to p_{max} and q_I runs from 0 to q_{max} for a total of $(1 + p_{max})(1 + q_{max})$ possible models. If $I = \emptyset$, then β_I does not exist, but $\hat{\boldsymbol{\beta}}_{I,0} = \mathbf{0}$, the $k \times 1$ vector of zeroes. For example, if $p_{max} = q_{max} = 5$, $S = \{1, 6, 7\}$ 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 *tspack* function `msarsim` simulates AR model selection with AIC. Let $k = p_{max}$. We recommend $n \geq 10k$ and $B \geq 20k$. 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 $\phi = (0.5, 0.33)$ 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), corrsponding 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 $B \approx 1.01BB$ where there was 1% augmentation from the bootstrapped full model. The simulations bootstrapped the full model $\hat{\boldsymbol{\beta}} = \hat{\phi}$, the model selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, 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. The term "full" is for the AR(p_{max}) full model, the term "VS" 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.

There was a convergence problem when trying to get the simulation for the mixed ARMA models. Simulation results in the last table are an average of five 1000 runs

Table 8.1. AR(p) Model Selection, n=100, tstype=1, BB=100, pmax=5, btype=1

e	ϕ_1	ϕ_2	$\phi_{p_{max}-1}$	$\phi_{p_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9486	0.9536	0.9662	0.9724	0.9588	0.9610	0.9724	0.9344	0.9480	0.9578
len	0.4274	0.4611	0.4443	0.3965	3.1913	3.1913	3.8924	1.9595	1.9595	2.0673
N,VS	0.9434	0.9982	0.9998	1.0000	0.9976	0.9960	0.9980	0.9590	0.9660	0.9728
len	0.4327	0.4410	0.3661	0.2718	4.1515	4.1515	4.4179	1.9633	1.9633	2.0642
N,MIX	0.9426	0.9984	0.9998	1.0000	0.9992	0.9980	0.9988	0.9576	0.9660	0.9760
len	0.4251	0.3885	0.2853	0.1842	4.6250	4.6250	4.8664	1.9650	1.9650	2.0758
t,full	0.9438	0.9568	0.9632	0.9710	0.9556	0.9558	0.9690	0.9358	0.9484	0.9558
len	0.4267	0.4611	0.4426	0.3974	3.1926	3.1926	3.8849	1.9587	1.9587	2.0634
t,VS	0.9446	0.9960	0.9996	1.0000	0.9964	0.9938	0.9966	0.9582	0.9642	0.9702
len	0.4322	0.4413	0.3634	0.2724	4.1520	4.1520	4.4218	1.9636	1.9636	2.0635
t,MIX	0.9442	0.9966	0.9996	1.0000	0.9986	0.9960	0.9984	0.9610	0.9648	0.9750
len	0.4246	0.3894	0.2845	0.1835	4.6370	4.6370	4.8772	1.9640	1.9640	2.0736
U,full	0.9472	0.9566	0.9652	0.9684	0.9542	0.9548	0.9740	0.9370	0.9514	0.9564
len	0.4279	0.4615	0.4419	0.3974	3.1915	3.1915	3.8985	1.9620	1.9620	2.0680
U,VS	0.9478	0.9958	1.0000	1.0000	0.9968	0.9950	0.9980	0.9610	0.9682	0.9744
len	0.4330	0.4412	0.3660	0.2734	4.1333	4.1333	4.4082	1.9667	1.9667	2.0676
U,MIX	0.9452	0.9962	1.0000	1.0000	0.9998	0.9978	0.9996	0.9602	0.9678	0.9776
len	0.4243	0.3893	0.2853	0.1870	4.6210	4.6210	4.8647	1.9647	1.9647	2.0750
E,full	0.9460	0.9598	0.9738	0.9742	0.9688	0.9652	0.9776	0.9350	0.9488	0.9568
len	0.4266	0.4606	0.4425	0.3982	3.1876	3.1876	3.8922	1.9576	1.9576	2.0647
E,VS	0.9438	0.9986	1.0000	1.0000	0.9990	0.9970	0.9996	0.9584	0.9654	0.9726
len	0.4324	0.4394	0.3648	0.2717	4.1751	4.1751	4.4426	1.9630	1.9630	2.0645
E,MIX	0.9414	0.9984	1.0000	1.0000	0.9998	0.9986	0.9998	0.9574	0.9630	0.9762
len	0.4252	0.3857	0.2806	0.1824	4.6552	4.6552	4.9004	1.9670	1.9670	2.0755

Table 8.2. AR(p) Model Selection,n=100,tstype=1,BB=100, pmax=5,btype=2

e	ϕ_1	ϕ_2	$\phi_{p_{max}-1}$	$\phi_{p_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9532	0.9218	0.9768	0.9814	0.9716	0.9500	0.9656	0.9168	0.9004	0.9346
len	0.4356	0.4657	0.4393	0.3861	2.8605	2.8605	3.4370	2.4902	2.4902	2.7112
N,VS	0.9554	0.9172	0.9998	1.0000	0.9994	0.9958	0.9994	0.9074	0.9056	0.9314
len	0.4591	0.4643	0.3620	0.2620	3.7975	3.7975	4.0539	2.5027	2.5027	2.7535
N,MIX	0.9458	0.8660	0.9998	1.0000	0.9996	0.9964	0.9996	0.8412	0.9006	0.9144
len	0.4547	0.4492	0.2787	0.1750	4.1940	4.1940	4.4183	2.5094	2.5094	2.8405
t,full	0.9552	0.9284	0.9778	0.9814	0.9690	0.9476	0.9664	0.9162	0.9044	0.9350
len	0.4368	0.4667	0.4387	0.3854	2.8586	2.8586	3.4250	2.4862	2.4862	2.7113
t,VS	0.9558	0.9276	1.0000	1.0000	1.0000	0.9966	1.0000	0.9024	0.9088	0.9322
len	0.4612	0.4660	0.3611	0.2604	3.7991	3.7991	4.0556	2.4975	2.4975	2.7548
t,MIX	0.9484	0.8746	1.0000	1.0000	1.0000	0.9978	1.0000	0.8450	0.9070	0.9250
len	0.4560	0.4511	0.2787	0.1729	4.2042	4.2042	4.4271	2.5022	2.5022	2.8450
U,full	0.9518	0.9190	0.9778	0.9812	0.9722	0.9552	0.9716	0.9192	0.8970	0.9294
len	0.4342	0.4663	0.4412	0.3856	2.8605	2.8605	3.4326	2.4854	2.4854	2.7174
U,VS	0.9534	0.9172	0.9998	1.0000	0.9990	0.9960	0.9988	0.8954	0.9012	0.9272
len	0.4574	0.4620	0.3623	0.2592	3.8220	3.8220	4.0751	2.4940	2.4940	2.7529
U,MIX	0.9364	0.8612	1.0000	1.0000	0.9996	0.9974	0.9992	0.8310	0.8926	0.9130
len	0.4526	0.4480	0.2761	0.1719	4.2084	4.2084	4.4331	2.5076	2.5076	2.8487
E,full	0.9574	0.9336	0.9812	0.9810	0.9794	0.9584	0.9748	0.9306	0.9084	0.9432
len	0.4379	0.4657	0.4397	0.3848	2.8620	2.8620	3.4305	2.4905	2.4905	2.7205
E,VS	0.9584	0.9342	1.0000	1.0000	0.9994	0.9980	0.9994	0.9124	0.9112	0.9372
len	0.4612	0.4637	0.3604	0.2587	3.8329	3.8329	4.0841	2.5044	2.5044	2.7616
E,MIX	0.9488	0.8782	1.0000	1.0000	0.9998	0.9976	0.9996	0.8508	0.9088	0.9256
len	0.4558	0.4498	0.2754	0.1695	4.2269	4.2269	4.4536	2.5082	2.5082	2.8458

Table 8.3. AR(p) Model Selection,n=100,tstype=2,BB=100, pmax=5, btype=1

e	ϕ_1	ϕ_2	$\phi_{p_{max}-1}$	$\phi_{p_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9372	0.9524	0.9666	0.9748	0.9518	0.9534	0.9684	0.9266	0.9432	0.9508
len	0.4275	0.4616	0.4436	0.3980	3.1853	3.1853	3.8961	1.9603	1.9603	2.0631
N,VS	0.9376	0.9974	1.0000	1.0000	0.9980	0.9950	0.9986	0.9530	0.9622	0.9696
len	0.4335	0.4414	0.3668	0.2726	4.1296	4.1296	4.4040	1.9659	1.9659	2.0641
N,MIX	0.9312	0.9978	1.0000	1.0000	0.9994	0.9976	0.9992	0.9508	0.9620	0.9736
len	0.4255	0.3896	0.2839	0.1855	4.6164	4.6164	4.8599	1.9650	1.9650	2.0734
t,full	0.9484	0.9524	0.9704	0.9698	0.9554	0.9562	0.9696	0.9326	0.9466	0.9534
len	0.4258	0.4586	0.4397	0.3942	3.2030	3.2030	3.9050	1.9633	1.9633	2.0682
t,VS	0.9434	0.9980	1.0000	1.0000	0.9980	0.9958	0.9988	0.9562	0.9612	0.9706
len	0.4310	0.4377	0.3608	0.2679	4.1741	4.1741	4.4407	1.9684	1.9684	2.0705
t,MIX	0.9372	0.9974	1.0000	1.0000	0.9996	0.9980	0.9994	0.9542	0.9600	0.9716
len	0.4229	0.3854	0.2789	0.1818	4.6629	4.6629	4.8987	1.9680	1.9680	2.0792
U,full	0.9438	0.9640	0.9634	0.9730	0.9588	0.9602	0.9728	0.9318	0.9460	0.9544
len	0.4283	0.4627	0.4434	0.3980	3.1906	3.1906	3.8899	1.9617	1.9617	2.0671
U,VS	0.9418	0.9974	1.0000	1.0000	0.9980	0.9960	0.9986	0.9604	0.9690	0.9734
len	0.4338	0.4427	0.3676	0.2724	4.1388	4.1388	4.4092	1.9677	1.9677	2.0669
U,MIX	0.9358	0.9974	1.0000	1.0000	0.9990	0.9980	0.9988	0.9544	0.9642	0.9752
len	0.4271	0.3911	0.2850	0.1862	4.6103	4.6103	4.8529	1.9656	1.9656	2.0762
E,full	0.9596	0.9632	0.9690	0.9720	0.9620	0.9620	0.9712	0.9426	0.9572	0.9610
len	0.4246	0.4644	0.4387	0.3935	3.2164	3.2164	3.9035	1.9630	1.9630	2.0574
E,VS	0.9600	0.9986	1.0000	1.0000	0.9980	0.9956	0.9972	0.9676	0.9720	0.9788
len	0.4291	0.4421	0.3577	0.2620	4.2165	4.2165	4.4794	1.9664	1.9664	2.0547
E,MIX	0.9556	0.9986	1.0000	1.0000	0.9990	0.9974	0.9986	0.9624	0.9682	0.9790
len	0.4143	0.3837	0.2753	0.1766	4.6755	4.6755	4.9142	1.9616	1.9616	2.0560

Table 8.4. AR(p) Model Selection, n=100,tstype=2,BB=100, pmax=5, btype=2

e	ϕ_1	ϕ_2	$\phi_{p_{max}-1}$	$\phi_{p_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9498	0.9232	0.9740	0.9822	0.9688	0.9462	0.9660	0.9188	0.8996	0.9364
len	0.4375	0.4648	0.4396	0.3852	2.8633	2.8633	3.4299	2.4891	2.4891	2.7115
N,VS	0.9548	0.9204	1.0000	1.0000	1.0000	0.9954	1.0000	0.9008	0.9048	0.9302
len	0.4611	0.4647	0.3612	0.2594	3.8157	3.8157	4.0681	2.5046	2.5046	2.7600
N,MIX	0.9450	0.8684	1.0000	1.0000	1.0000	0.9972	1.0000	0.8340	0.8966	0.9166
len	0.4564	0.4504	0.2777	0.1710	4.2037	4.2037	4.4889	2.5104	2.5104	2.8443
t,full	0.9516	0.9266	0.9736	0.9822	0.9738	0.9536	0.9710	0.9136	0.9062	0.9354
len	0.4376	0.4651	0.4366	0.3833	2.8722	2.8722	3.4386	2.4963	2.4963	2.7155
t,VS	0.9552	0.9252	1.0000	1.0000	0.9998	0.9956	0.9998	0.9018	0.9128	0.9352
len	0.4617	0.4621	0.3569	0.2582	3.8377	3.8377	4.0902	2.5045	2.5045	2.7566
t,MIX	0.9472	0.8702	1.0000	1.0000	1.0000	0.9974	0.9996	0.8394	0.9056	0.9218
len	0.4557	0.4481	0.2728	0.1689	4.2207	4.2207	4.4470	2.5067	2.5067	2.8434
U,full	0.9460	0.9198	0.9794	0.9820	0.9762	0.9542	0.9706	0.9192	0.8948	0.9300
len	0.4375	0.4691	0.4417	0.3849	2.8610	2.8610	3.4277	2.4861	2.4861	2.7164
U,VS	0.9488	0.9186	1.0000	1.0000	0.9998	0.9966	0.9998	0.8984	0.8984	0.9254
len	0.4610	0.4638	0.3618	0.2587	3.8402	3.8402	4.0911	2.4953	2.4953	2.7541
U,MIX	0.9314	0.8590	1.0000	1.0000	1.0000	0.9968	1.0000	0.8326	0.8920	0.9100
len	0.4552	0.4480	0.2758	0.1699	4.2255	4.2255	4.4478	2.5061	2.5061	2.8469
E,full	0.9528	0.9330	0.9802	0.9814	0.9756	0.9542	0.9694	0.9254	0.9038	0.9364
len	0.4347	0.4704	0.4370	0.3817	2.8764	2.8764	3.4385	2.5096	2.5096	2.7249
E,VS	0.9542	0.9302	1.0000	1.0000	1.0000	0.9962	0.9998	0.9054	0.9122	0.9328
len	0.4580	0.4651	0.3563	0.2527	3.8609	3.8609	4.1087	2.5191	2.5191	2.7688
E,MIX	0.9412	0.8674	1.0000	1.0000	1.0000	0.9976	0.9998	0.8284	0.9098	0.9228
len	0.4518	0.4501	0.2716	0.1654	4.2470	4.2470	4.4644	2.5121	2.5121	2.8558

Table 8.5. AR(p) Model Selection, n=400,ϕ = 0.5,BB=100, pmax=5, btype=1

e	ϕ_1	ϕ_2	$\phi_{p_{max}-1}$	$\phi_{p_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9522	0.9508	0.9574	0.9642	0.9438	0.9596	0.9720	0.9442	0.9502	0.9574
len	0.2104	0.2329	0.2304	0.2066	3.1743	3.1743	3.9237	1.9582	1.9582	2.0517
N,VS	0.9562	0.9976	0.9998	1.0000	0.9954	0.9938	0.9966	0.9486	0.9526	0.9580
len	0.2102	0.2248	0.1957	0.1496	4.0079	4.0079	4.2904	1.9592	1.9592	2.0364
N,MIX	0.9578	0.9974	0.9998	1.0000	0.9984	0.9972	0.9988	0.9494	0.9494	0.9538
len	0.2043	0.2015	0.1551	0.1067	4.5015	4.5015	4.7525	1.9574	1.9574	2.0193
t,full	0.9544	0.9558	0.9578	0.9598	0.9478	0.9624	0.9718	0.94352	0.9554	0.9624
len	0.2100	0.2328	0.2308	0.2066	3.1733	3.1733	3.9342	1.9572	1.9572	2.0534
t,VS	0.9596	0.9972	0.9996	0.9998	0.9966	0.9938	0.9976	0.9510	0.9576	0.9622
len	0.2096	0.2247	0.1969	0.1501	4.0100	4.0100	4.2937	1.9551	1.9551	2.0337
t,MIX	0.9622	0.9980	0.9996	0.9998	0.9984	0.9966	0.9986	0.9516	0.9534	0.9596
len	0.2041	0.2012	0.1556	0.1071	4.5124	4.5124	4.7637	1.9587	1.9587	2.0245
U,full	0.9482	0.9528	0.9578	0.9548	0.9470	9602	0.9712	0.9338	0.9504	0.9572
len	0.2102	0.2333	0.2317	0.2066	3.1737	3.1737	3.9198	1.9558	1.9558	2.0538
U,VS	0.9524	0.9960	1.0000	1.0000	0.9970	0.9952	0.9976	0.9406	0.9530	0.9596
len	0.2101	0.2252	0.1973	0.1491	4.0063	4.0063	4.2884	1.9566	1.9566	2.0369
U,MIX	0.9526	0.9964	1.0000	1.0000	0.9980	0.9966	0.9984	0.9436	0.9490	0.9544
len	0.2043	0.2012	0.1558	0.1059	4.5100	4.5100	4.7564	1.9605	1.9605	2.0200
E,full	0.9544	0.9510	0.9632	0.9630	0.9480	0.9616	0.9720	0.9446	0.9536	0.9574
len	0.2104	0.2331	0.2311	0.2064	3.1723	3.1723	3.9228	1.9579	1.9579	2.0511
E,VS	0.9574	0.9958	0.9994	1.0000	0.9960	0.9930	0.9964	0.9502	0.9556	0.9602
len	0.2101	0.2250	0.1964	0.1492	4.0136	4.0136	4.2979	1.9573	1.9573	2.0359
E,MIX	0.9556	0.9960	0.9998	1.0000	0.9988	0.9960	0.9980	0.9520	0.9476	0.95522
len	0.2040	0.2013	0.1551	0.1058	4.5080	4.5080	4.7575	1.9611	1.9611	2.0232

Table 8.6. AR(p) Model Selection, n=400,$\phi = 0.5$,BB=200, pmax=5, btype=2

e	ϕ_1	ϕ_2	$\phi_{p_{max}-1}$	$\phi_{p_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9552	0.9464	0.9606	0.9662	0.9554	0.9554	0.9674	0.9380	0.9532	0.9630
len	0.2152	0.2343	0.2299	0.2036	2.8531	2.8531	3.4531	2.4793	2.4793	2.6524
N,VS	0.9610	0.9528	0.9996	0.9998	0.9966	0.9892	0.9974	0.9492	0.9592	0.9652
len	0.2152	0.2333	0.2009	0.1506	3.5670	3.5670	3.8436	2.4932	2.4932	2.6497
N,MIX	0.9578	0.9480	0.9996	0.9998	0.9984	0.9948	0.9976	0.9460	0.9518	0.9636
len	0.2124	0.2233	0.1604	0.1076	3.9967	3.9967	4.2330	2.5055	2.5055	2.6487
t,full	0.96222	0.9496	0.9558	0.9660	0.9492	0.9486	0.9628	0.9410	0.9558	0.9634
len	0.2160	0.2346	0.2293	0.2045	2.8560	2.8560	3.4481	2.4763	2.4763	2.6536
t,VS	0.9650	0.9550	0.9994	1.0000	0.9968	0.9906	0.9968	0.9548	0.9628	0.9690
len	0.2159	0.2332	0.2008	0.1520	3.5573	3.5573	3.8353	2.4951	2.4951	2.6495
t,MIX	0.9676	0.9496	0.9996	1.0000	0.9984	0.9938	0.9978	0.9530	0.9586	0.9686
len	0.2129	0.2233	0.1612	0.1089	3.9789	3.9789	4.2201	2.5098	2.5098	2.6502
U,full	0.9522	0.9452	0.9594	0.9618	0.9520	0.9516	0.9632	0.9392	0.9504	0.9636
len	0.2147	0.2346	0.2307	0.2044	2.8577	2.8577	3.4605	2.4769	2.4769	2.6483
U,VS	0.9560	0.9496	0.9996	1.0000	0.9974	0.9930	0.9972	0.9516	0.9584	0.9684
len	0.2149	0.2334	0.2023	0.1529	3.5394	3.5394	3.8185	2.4935	2.4935	2.6474
U,MIX	0.9576	0.9488	0.9996	1.0000	0.9990	0.9954	0.9988	0.9500	0.9544	0.9660
len	0.2124	0.2240	0.1621	0.1108	3.9636	3.9636	4.2020	2.5105	2.5105	2.6526
E,full	0.9558	0.9494	0.9650	0.9622	0.9556	0.9550	0.9634	0.9444	0.9578	0.9670
len	0.2158	0.2346	0.2293	0.2037	2.8552	2.8552	3.4555	2.4769	2.4769	2.6482
E,VS	0.9584	0.9552	0.9996	1.0000	0.9962	0.9934	0.9964	0.9554	0.9638	0.9728
len	0.2158	0.2332	0.2005	0.1511	3.5638	3.5638	3.8415	2.4952	2.4952	2.6462
E,MIX	0.9598	0.9506	0.9994	1.0000	0.9980	0.9948	0.9974	0.9536	0.9622	0.9710
len	0.2132	0.2238	0.1602	0.1083	3.9920	3.9920	4.2298	2.5099	2.5099	2.6502

Table 8.7. AR(p) Model Selection, n=400,$\phi = (0.5, 0.33)$,BB=200, pmax=5, btype=1

e	ϕ_1	ϕ_2	$\phi_{p_{max}-1}$	$\phi_{p_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9516	0.9528	0.9584	0.9656	0.9478	0.9600	0.9714	0.9388	0.9490	0.9560
len	0.2106	0.2330	0.2306	0.2068	3.1742	3.1742	3.9235	1.9584	1.9584	2.0527
N,VS	0.9522	0.9970	1.0000	1.0000	0.9968	0.9952	0.9978	0.9442	0.9520	0.9568
len	0.2103	0.2249	0.1953	0.1496	4.0195	4.0195	4.3052	1.9607	1.9607	2.0393
N,MIX	0.9552	0.9970	1.0000	1.0000	0.9984	0.9978	0.9986	0.9470	0.9478	0.9552
len	0.2042	0.2005	0.1551	0.1055	4.5174	4.5174	4.7682	1.9609	1.9609	2.0242
t,full	0.9534	0.9504	0.9602	0.9644	0.9480	0.9612	0.9704	0.9436	0.9550	0.9584
len	0.2098	0.2325	0.2303	0.2066	3.1835	3.1835	3.9338	1.9592	1.9592	2.0522
t,VS	0.9566	0.9960	0.9998	1.0000	0.9956	0.9936	0.9960	0.9460	0.9564	0.9582
len	0.2094	0.2236	0.1951	0.1486	4.0387	4.0387	4.3175	1.9599	1.9599	2.0369
t,MIX	0.9544	0.9970	0.9996	1.0000	0.9980	0.9958	0.9982	0.9478	0.9522	0.9544
len	0.2031	0.1996	0.1533	0.1044	4.5395	4.5395	4.7851	1.9604	1.9604	2.0210
U,full	0.9560	0.9632	0.9576	0.9638	0.9488	0.9622	0.9718	0.9474	0.9590	0.9642
len	0.2102	0.2329	0.2311	0.2068	3.1749	3.1749	3.9185	1.9614	1.9614	2.0564
U,VS	0.9600	0.9980	1.0000	0.9998	0.9970	0.9952	0.9978	0.9530	0.9604	0.9646
len	0.2097	0.2240	0.1955	0.1495	4.0204	4.0204	4.3047	1.9606	1.9606	2.0396
U,MIX	0.9594	0.9986	1.0000	0.9998	0.9992	0.9978	0.9992	0.9506	0.9552	0.9600
len	0.2038	0.2008	0.1544	0.1054	4.5302	4.5302	4.7798	1.9605	1.9605	2.0213
E,full	0.9610	0.9536	0.9622	0.9618	0.9504	0.9620	0.9712	0.9488	0.9602	0.9646
len	0.2088	0.2333	0.2302	0.2054	3.1872	3.1872	3.9355	1.9545	1.9545	2.0486
E,VS	0.9648	0.9960	0.9998	1.0000	0.9980	0.9974	0.9984	0.9548	0.9616	0.9654
len	0.2088	0.2246	0.1944	0.1473	4.0401	4.0401	4.3196	1.9567	1.9567	2.0329
E,MIX	0.9668	0.9968	0.9998	1.0000	0.9990	0.9976	0.9992	0.9548	0.9554	0.9618
len	0.2019	0.1994	0.1527	0.1038	4.5330	4.5330	4.7819	1.9585	1.9585	2.0170

Table 8.8. AR(p) Model Selection, n=400,$\phi = (0.5, 0.33)$,BB=200, pmax=5, btype=2

e	ϕ_1	ϕ_2	$\phi_{p_{max}-1}$	$\phi_{p_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9586	0.9498	0.9638	0.9624	0.9550	0.9546	0.9658	0.9404	0.9530	0.9648
len	0.2145	0.2342	0.2297	0.2039	2.8507	2.8507	3.4645	2.4748	2.4748	2.6479
N,VS	0.9588	0.9548	0.9998	0.9998	0.9982	0.9942	0.9988	0.9538	0.9618	0.9684
len	0.2145	0.2332	0.2014	0.1516	3.5445	3.5445	3.8268	2.4930	2.4930	2.6475
N,MIX	0.9620	0.9474	1.0000	0.9998	0.9986	0.9966	0.9990	0.9536	0.9590	0.9696
len	0.2129	0.2238	0.1614	0.1106	3.9697	3.9697	4.2116	2.5086	2.5086	2.6551
t,full	0.9500	0.9482	0.9586	0.9682	0.9568	0.9534	0.9656	0.9434	0.9578	0.9676
len	0.2154	0.2343	0.2288	0.2038	2.8566	2.8566	3.4629	2.4791	2.4791	2.6574
t,VS	0.9546	0.9524	0.9994	1.0000	0.9974	0.9922	0.9978	0.9552	0.9634	0.9726
len	0.2153	0.2330	0.2005	0.1516	3.5599	3.5599	3.8371	2.4988	2.4988	2.6560
t,MIX	0.9548	0.9462	0.9994	1.0000	0.9984	0.9958	0.9982	0.9542	0.9610	0.9690
len	0.2133	0.2234	0.1599	0.1082	3.9920	3.9920	4.2306	2.5106	2.5106	2.6526
U,full	0.9524	0.9492	0.9618	0.9634	0.9496	0.9522	0.9632	0.9374	0.9536	0.9626
len	0.2148	0.2354	0.2304	0.2040	2.8542	2.8542	3.4598	2.4766	2.4766	2.6409
U,VS	0.9558	0.9512	0.9996	1.0000	0.9966	0.9912	0.9970	0.9526	0.9594	0.9664
len	0.2149	0.2345	0.2019	0.1516	3.5394	3.5394	3.8197	2.4950	2.4950	2.6114
U,MIX	0.9590	0.9434	0.9998	1.0000	0.9986	0.9948	0.9978	0.9520	0.9578	0.9670
len	0.2130	0.2240	0.1619	0.1107	3.9678	3.9678	4.20801	2.5077	2.5077	2.6497
E,full	0.9558	0.9542	0.9582	0.9616	0.9530	0.9530	0.9652	0.9436	0.9540	0.9660
len	0.2152	0.2360	0.2289	0.2033	2.8653	2.8653	3.4664	2.4847	2.4847	2.6608
E,VS	0.9610	0.9572	0.9996	0.9998	0.9976	0.9930	0.9980	0.9542	0.9614	0.9730
len	0.2154	0.2350	0.2006	0.1500	3.5634	3.5634	3.8412	2.5013	2.5013	2.6534
E,MIX	0.9558	0.9504	0.9996	0.9998	0.9988	0.9962	0.9992	0.9510	0.9592	0.9678
len	0.2131	0.2247	0.1604	0.1073	3.9840	3.9840	4.2199	2.5187	2.5187	2.6633

Table 8.9. MA(q) Model Selection, n=100, tstype=1,BB=100, qmax=5, btype=1

e	θ_1	θ_2	$\theta_{q_{max}-1}$	$\theta_{q_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9340	0.9484	0.9422	0.9304	0.8630	0.9920	0.9976	0.9180	0.9610	0.9736
len	0.4790	0.5454	0.5978	0.5690	3.3725	3.3725	4.5228	1.9611	1.96111	2.2120
N,VS	0.9784	0.9998	1.0000	1.0000	0.9998	0.9998	1.0000	0.9802	0.9870	0.9894
len	0.5689	0.4942	0.2545	0.1897	5.4980	5.4980	5.6905	1.9926	1.99263	2.1369
N,MIX	0.9772	0.9996	1.0000	1.0000	1.0000	0.9998	1.0000	0.9766	0.9868	0.9880
len	0.5922	0.4291	0.2278	0.1794	5.8106	5.8106	5.9888	1.9974	1.9974	2.1292
t,full	0.9286	0.9468	0.9472	0.9340	0.8716	0.9918	0.9978	0.9110	0.9532	0.9714
len	0.4787	0.5447	0.5990	0.5687	3.3681	3.3681	4.4906	1.9657	1.9657	2.2122
t,VS	0.9732	0.9998	1.0000	1.0000	0.9998	0.9994	0.9998	0.9730	0.9822	0.9850
len	0.5658	0.4926	0.2526	0.1902	5.5154	5.5154	5.7063	1.9869	1.9869	2.1263
t,MIX	0.9748	0.9998	1.0000	1.0000	1.0000	0.9992	1.0000	0.9742	0.9846	0.9858
len	0.5898	0.4303	0.2251	0.1797	5.8076	5.8076	5.9857	1.9919	1.9919	2.1198
U,full	0.9282	0.9418	0.9402	0.9294	0.8640	0.9908	0.9960	0.9152	0.9534	0.9676
len	0.4800	0.5444	0.5994	0.5698	3.3696	3.3696	4.5138	1.9629	1.9629	2.2137
U,VS	0.9752	0.9992	1.0000	1.0000	1.0000	1.0000	1.0000	0.9712	0.9814	0.9828
len	0.5657	0.4933	0.2571	0.1936	5.4919	5.4919	5.6861	1.9865	1.9865	2.1314
U,MIX	0.9746	0.9996	1.0000	1.0000	1.0000	0.9998	1.0000	0.9724	0.9822	0.9840
len	0.5889	0.4313	0.2296	0.1822	5.7952	5.7952	5.9739	1.9922	1.9922	2.1212
E,full	0.9338	0.9508	0.9462	0.9358	0.8700	0.9930	0.9976	0.9224	0.9612	0.9728
len	0.4798	0.5452	0.5970	0.5686	3.3718	3.3718	4.4758	1.9626	1.9626	2.2022
E,VS	0.9756	0.9992	1.0000	1.0000	1.0000	0.9998	1.0000	0.9760	0.9832	0.9854
len	0.5661	0.4920	0.2528	0.1898	5.5171	5.5171	5.7071	1.9910	1.9910	2.1312
E,MIX	0.9734	0.9990	1.0000	1.0000	1.0000	0.9998	1.0000	0.9722	0.9852	0.9852
len	0.5901	0.4291	0.2260	0.1789	5.8349	5.8349	6.0110	1.9962	1.9962	2.1234

Table 8.10. MA(q) Model Selection, n=100,tstype=1,BB=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	0.9390	0.9478	0.9366	0.9296	0.8594	0.9934	0.9968	0.9216	0.9618	0.9742
len	0.4808	0.5462	0.5988	0.5695	3.3739	3.3739	4.5373	1.9632	1.9632	2.2130
N,VS	0.9786	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000	0.9798	0.9868	0.9900
len	0.5728	0.4956	0.2567	0.1936	5.4872	5.4872	5.6805	1.9979	1.9979	2.1401
N,MIX	0.9768	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000	0.9772	0.9872	0.9896
len	0.5970	0.4321	0.2311	0.1820	5.7890	5.7890	5.9680	2.0104	2.0104	2.1378
t,full	0.9244	0.9510	0.9424	0.9380	0.9694	0.9914	0.9978	0.9076	0.9522	0.9662
len	0.4786	0.5454	0.5959	0.5670	3.3845	3.3845	4.5081	1.9671	1.9671	2.2166
t,VS	0.9750	0.9996	1.0000	1.0000	0.9996	0.9996	1.0000	0.9748	0.9830	0.9850
len	0.5681	0.4926	0.2509	0.1896	5.5024	5.5024	5.6934	1.9898	1.9898	2.1292
t,MIX	0.9758	0.9996	1.0000	1.0000	0.9996	0.9998	1.0000	0.9740	0.9844	0.9846
len	0.5920	0.4271	0.2246	0.1792	5.8234	5.8234	5.9993	1.9989	1.9989	2.1246
U,full	0.9346	0.9492	0.9344	0.9306	0.8484	0.9916	0.9968	0.9168	0.9614	0.9772
len	0.4795	0.5470	0.6002	0.5710	3.3731	3.3731	4.5339	1.9588	1.9588	2.2159
U,VS	0.9774	0.9998	1.0000	1.0000	0.9998	0.9994	0.9998	0.9760	0.9852	0.9890
len	0.5686	0.4987	0.2582	0.1944	5.4833	5.4833	5.6744	1.9962	1.9962	2.1439
U,MIX	0.9778	0.9998	1.0000	1.0000	0.9998	0.9994	1.0000	0.9770	0.9864	0.9896
len	0.5938	0.4344	0.2331	0.1839	5.8019	.8019	5.9800	2.0031	2.0031	2.1341
E,full	0.9386	0.9508	0.9404	0.9360	0.8714	0.9946	0.9978	0.9180	0.9606	0.9720
len	0.4764	0.5392	0.5925	0.5635	3.3942	3.3942	4.5112	1.9714	1.9714	2.1984
E,VS	0.9770	0.9996	1.0000	1.0000	1.0000	0.9996	1.0000	0.9770	0.9832	0.9864
len	0.5595	0.4865	0.2478	0.1888	5.5183	5.5183	5.7084	1.9954	1.9954	2.1238
E,MIX	0.9734	0.9998	1.0000	1.0000	1.0000	0.9998	1.0000	0.9744	0.9848	0.9870
len	0.5867	0.4209	0.2208	0.1782	5.8227	5.8227	6.0027	2.0093	2.0093	2.1230

Table 8.11. MA(q) Model Selection, n=100, tstype=2,BB=100, qmax=5, btype=1

e	θ_1	θ_2	$\theta_{q_{max}-1}$	$\theta_{q_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9302	0.9458	0.9402	0.9300	0.8916	0.9818	0.9898	0.9190	0.9682	0.9798
len	0.4696	0.5559	0.5991	0.5762	3.0437	3.0437	3.9726	2.5169	2.5169	2.8582
N,VS	0.9576	0.9802	1.0000	1.0000	0.9998	0.9994	1.0000	0.9726	0.9806	0.9858
len	0.4835	0.6004	0.3660	0.2454	4.2819	4.2819	4.5142	2.5671	2.5671	2.7344
N,MIX	0.9618	0.9756	1.0000	1.0000	1.0000	0.9996	1.0000	0.9708	0.9836	0.9884
len	0.4774	0.5974	0.3147	0.2210	4.5326	4.5326	4.7355	2.5766	2.5766	2.7250
t,full	0.9410	0.9450	0.9446	0.9320	0.9066	0.9864	0.9926	0.9168	0.9696	0.9808
len	0.4701	0.5568	0.5997	0.5754	3.0376	3.0376	3.9507	2.5189	2.5189	2.8530
t,VS	0.9644	0.9756	1.0000	1.0000	1.0000	0.9998	1.0000	0.9742	0.9800	0.9864
len	0.4826	0.5984	0.3604	0.2410	4.2793	4.2793	4.5116	2.5609	2.5609	2.7280
t,MIX	0.9662	0.9736	1.0000	1.0000	1.0000	1.0000	1.0000	0.9666	0.9820	0.9868
len	0.4759	0.5975	0.3115	0.2172	4.5435	4.5435	4.7450	2.5732	2.5732	2.7232
U,full	0.9326	0.9418	0.9424	0.9292	0.8990	0.9844	0.9922	0.9130	0.9624	0.9780
len	0.4709	0.5556	0.5991	0.5776	3.0371	3.0371	3.9621	2.5144	2.5144	2.8532
U,VS	0.9598	0.9740	1.0000	1.0000	0.9994	0.9994	0.9996	0.9686	0.9782	0.9850
len	0.4850	0.6011	0.3629	0.2464	4.2798	4.2798	4.5134	2.5601	2.5601	2.7321
U,MIX	0.9606	0.9744	1.0000	1.0000	0.9994	0.9994	0.9998	0.9640	0.9792	0.9832
len	0.4778	0.5987	0.3128	0.2206	4.5352	4.5352	4.7403	2.5740	2.5740	2.7245
E,full	0.9454	0.9512	0.9394	0.9364	0.9112	0.9878	0.9938	0.9286	0.9700	0.9806
len	0.4704	0.5567	0.5989	0.5765	3.0441	3.0441	3.9451	2.5177	2.5177	2.8371
E,VS	0.9638	0.9784	1.0000	1.0000	1.0000	0.9998	1.0000	0.9776	0.9846	0.9872
len	0.4815	0.5976	0.3588	0.2425	4.2862	4.2862	4.5192	2.5609	2.5609	2.7253
E,MIX	0.9688	0.9748	1.0000	1.0000	1.0000	0.9998	1.0000	0.9726	0.9844	0.98800
len	0.4743	0.5950	0.3101	0.2177	4.5533	4.5533	4.7529	2.5732	2.5732	2.7223

Table 8.12. MA(q) Model Selection, n=100, tstype=2,BB=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	0.9380	0.9464	0.9382	0.9298	0.8912	0.9818	0.9906	0.9154	0.9664	0.9808
len	0.4701	0.5561	0.6006	0.5754	3.0421	3.0421	3.9717	2.5199	2.5199	2.8599
N,VS	0.9660	0.9766	1.0000	1.0000	1.0000	0.9996	1.0000	0.9728	0.9824	0.9906
len	0.4838	0.6026	0.3657	0.2446	4.2746	4.2746	4.5122	2.5617	2.5617	2.7340
N,MIX	0.9630	0.9784	1.0000	1.0000	1.0000	0.9996	1.0000	0.9696	0.9820	0.9878
len	0.4778	0.6004	0.3164	0.2210	4.5293	4.5293	4.7343	2.5745	2.5745	2.7302
t,full	0.9350	0.9494	0.9454	0.9316	0.9086	0.9862	0.9920	0.9174	0.9630	0.9782
len	0.4689	0.5530	0.5974	0.5745	3.0517	3.0517	3.9582	2.5251	2.5251	2.8581
t,VS	0.9606	0.9780	1.0000	1.0000	1.0000	0.9998	1.0000	0.9700	0.9798	0.9844
len	0.4802	0.5926	0.3562	0.2381	4.3025	4.3025	4.5315	2.5665	2.5665	2.7273
t,MIX	0.9596	0.9730	1.0000	1.0000	1.0000	0.9998	1.0000	0.9668	0.9804	0.9836
len	0.4738	0.5916	0.3068	0.2135	4.5619	4.5619	4.7606	2.5806	2.5806	2.7214
U,full	0.9318	0.9438	0.9404	0.9246	0.8922	0.9860	0.9936	0.9096	0.9630	0.9766
len	0.4726	0.5579	0.6053	0.5777	3.0393	3.0393	3.9862	2.5178	2.5178	2.8646
U,VS	0.9612	0.9784	1.0000	1.0000	0.9996	0.9990	0.9996	0.9704	0.9818	0.9858
len	0.4864	0.6016	0.3668	0.2477	4.2745	4.2745	4.5087	2.5628	2.5628	2.7302
U,MIX	0.9620	0.9780	1.0000	1.0000	1.0000	0.9996	1.0000	0.9674	0.9808	0.9852
len	0.4803	0.6006	0.3184	0.2240	4.5218	4.5218	4.7282	2.5756	2.5756	2.7261
E,full	0.9430	0.9506	0.9428	0.9356	0.9046	0.9868	0.9926	0.9258	0.9660	0.9806
len	0.4661	0.5486	0.5919	0.5689	3.0476	3.0476	3.9639	2.5296	2.5296	2.8509
E,VS	0.9640	0.9794	1.0000	1.0000	1.0000	0.9998	1.0000	0.9772	0.9832	0.9874
len	0.4794	0.5870	0.3538	0.2380	4.2892	4.2892	4.5236	2.5642	2.5642	2.7267
E,MIX	0.9666	0.9800	1.0000	1.0000	1.0000	1.0000	1.0000	0.9730	0.9820	0.9880
len	0.4732	0.5842	0.3072	0.2145	4.5534	4.5534	4.7543	2.5735	2.5735	2.7165

Table 8.13. MA(q) Model Selection, n=400,tstype=1,BB=100, qmax=5,btype=1

e	θ_1	θ_2	$\theta_{q_{max}-1}$	$\theta_{q_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9530	0.9536	0.9516	0.9522	0.9320	0.9896	0.99420	0.9380	0.9542	0.9628
len	0.2155	0.2414	0.2470	0.2230	3.2005	3.2005	4.1791	1.9568	1.9568	2.0809
N,VS	0.9572	0.9990	1.0000	1.0000	0.9998	0.9998	1.0000	0.9470	0.9546	0.9614
len	0.2156	0.2139	0.1110	0.0795	5.5321	5.5321	5.7162	1.9595	1.9595	2.0456
N,MIX	0.9518	0.9990	1.0000	1.0000	0.9998	0.9998	1.0000	0.9432	0.9452	0.9502
len	0.2060	0.1787	0.0967	0.0727	5.8620	5.8620	6.0320	1.9605	1.9605	2.0303
t,full	0.9498	0.9532	0.9558	0.9506	0.9338	0.9934	0.9972	0.9398	0.9528	0.9626
len	0.2154	0.2421	0.2471	0.2238	3.2020	3.2020	4.1682	1.9540	1.9540	2.0796
t,VS	0.9586	0.9988	1.0000	1.0000	1.0000	1.0000	1.0000	0.9472	0.9558	0.9638
len	0.2157	0.2140	0.1117	0.0797	5.5106	5.5106	5.6947	1.9588	1.9588	2.0446
t,MIX	0.9534	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000	0.9412	0.9492	0.9560
len	0.2064	0.1786	0.0968	0.0731	5.8309	5.8309	6.0033	1.9623	1.9623	2.0277
U,full	0.9500	0.9528	0.9518	0.9516	0.9276	9902	0.9964	0.9404	0.9552	0.9628
len	0.2155	0.2418	0.2471	0.2237	3.1975	3.1975	4.1800	1.9596	1.9596	2.0781
U,VS	0.9570	0.9990	1.0000	1.0000	0.9998	0.9998	0.9998	0.9508	0.9582	0.9636
len	0.2157	0.2142	0.1118	0.0799	5.5266	5.5266	5.7095	1.9611	1.9611	2.0435
U,MIX	0.9544	0.9994	1.0000	1.0000	0.9998	1.0000	0.9998	0.9422	0.9482	0.9558
len	0.2059	0.1784	0.0975	0.0731	5.8406	5.8406	6.0096	1.9605	1.9605	2.0251
E,full	0.9568	0.9552	0.9490	0.9524	0.9298	0.9894	0.9962	0.9394	0.9578	0.9640
len	0.2155	0.2419	0.2466	0.2227	3.2002	3.2002	4.1789	1.9589	1.9589	2.0807
E,VS	0.9660	0.9994	1.0000	1.0000	0.9998	0.9998	1.0000	0.9550	0.9604	0.9668
len	0.2154	0.2136	0.1115	0.0807	5.5234	5.5234	5.7042	1.9606	1.9606	2.0425
E,MIX	0.9598	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000	0.9482	0.9502	0.9564
len	0.2052	0.1774	0.0973	0.0739	5.8561	5.8561	6.0250	1.9583	1.9583	2.0257

Table 8.14. MA(q) Model Selection, n=400,tstype=1,BB=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	0.9514	0.9538	0.9536	0.9470	0.9292	0.9926	0.9968	0.9408	0.9582	0.9654
len	0.2153	0.2417	0.2463	0.2227	3.2006	3.2006	4.1834	1.9581	1.9581	2.0805
N,VS	0.9620	0.9986	1.0000	1.0000	1.0000	1.0000	1.0000	0.9558	0.9598	0.9656
len	0.2154	0.2136	0.1107	0.0803	5.5342	5.5342	5.7191	1.9600	1.9600	2.0464
N,MIX	0.9550	0.9988	1.0000	1.0000	1.0000	1.0000	1.0000	0.9450	0.9530	0.9558
len	0.2061	0.1780	0.0958	0.0736	5.8565	5.8565	6.0277	1.9624	1.9624	2.0307
t,full	0.9558	0.9570	0.9516	0.9532	0.9346	0.9912	0.9952	0.9420	0.9596	0.9636
len	0.2148	0.2416	0.2459	0.2225	3.2086	3.2086	4.1666	1.9597	1.9597	2.0813
t,VS	0.9630	0.9994	1.0000	1.0000	1.0000	1.0000	1.0000	0.9528	0.9622	0.9650
len	0.2147	0.2138	0.1104	0.0787	5.5490	5.5490	5.7289	1.9623	1.9623	2.0472
t,MIX	0.9592	0.9992	1.0000	1.0000	1.0000	1.0000	1.0000	0.9468	0.9540	0.9602
len	0.2052	0.1768	0.0962	0.0722	5.8743	5.8743	6.0424	1.9586	1.9586	2.0262
U,full	0.9536	0.9546	0.9580	0.9524	0.9336	0.9904	0.9950	0.9398	0.9558	0.9648
len	0.2157	0.2425	0.2473	0.2237	3.1997	3.1997	4.1545	1.9560	1.9560	2.0769
U,VS	0.9614	0.9992	1.0000	1.0000	0.9998	0.9998	1.0000	0.9534	0.9604	0.9648
len	0.2160	0.2148	0.1115	0.0807	5.5294	5.5294	5.7105	1.9598	1.9598	2.0440
U,MIX	0.9604	0.9992	1.0000	1.0000	1.0000	1.0000	1.0000	0.9486	0.9522	0.9572
len	0.2063	0.1781	0.0966	0.0734	5.8550	5.8550	6.0222	1.9600	1.9600	2.0263
E,full	0.9492	0.9550	0.9546	0.9596	0.9342	0.9910	0.9952	0.9370	0.9510	0.9592
len	0.2138	0.2400	0.2451	0.2222	3.2139	3.2139	4.1663	1.9565	1.9565	2.0707
E,VS	0.9582	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000	0.9494	0.9548	0.9630
len	0.2142	0.2122	0.1095	0.0782	5.5507	5.5507	5.7303	1.9594	1.9594	2.0400
E,MIX	0.9588	0.9994	1.0000	1.0000	1.0000	1.0000	1.0000	0.9472	0.9480	0.9530
len	0.2053	0.1764	0.0951	0.0718	5.8724	5.8724	6.0403	1.9669	1.9669	2.0257

Table 8.15. MA(q) Model Selection, n=400,tstype=2,BB=100, qmax=5, btype=1

e	θ_1	θ_2	$\theta_{q_{max}-1}$	$\theta_{q_{max}}$	pr0	hyb0	br0	pr1	hyb1	br1
N,full	0.9526	0.9548	0.9550	0.9536	0.9388	0.9862	0.9928	0.9418	0.9672	0.9764
len	0.2151	0.2419	0.2463	0.2232	2.8727	2.8727	3.6766	2.4795	2.4795	2.7066
N,VS	0.9622	0.9680	1.0000	1.0000	0.9994	0.9992	0.9998	0.9624	0.9672	0.9756
len	0.2146	0.2309	0.1550	0.1029	4.3233	4.3233	4.5433	2.4985	2.4985	2.6085
N,MIX	0.9570	0.9660	1.0000	1.0000	0.9998	0.9994	0.9998	0.9538	0.9558	0.9616
len	0.2045	0.2174	0.1283	0.0877	4.6013	4.6013	4.7932	2.4978	2.4978	2.5728
t,full	0.9548	0.9608	0.9560	0.9602	0.9454	0.9862	0.9926	0.9448	0.9722	0.9790
len	0.2147	0.2427	0.2466	0.2234	2.8750	2.8750	3.6609	2.4757	2.4757	2.7010
t,VS	0.9612	0.9690	1.0000	1.0000	0.9988	0.9988	0.9994	0.9646	0.9708	0.9750
len	0.2138	0.2312	0.1538	0.1018	4.3133	4.3133	4.5338	2.4952	2.4952	2.6009
t,MIX	0.9622	0.9658	1.0000	1.0000	0.9992	0.9992	0.9998	0.9614	0.9620	0.9670
len	0.2038	0.2170	0.1270	0.0867	4.5911	4.5911	4.7824	2.4969	2.4969	2.5735
U,full	0.9484	0.9536	0.9548	0.9550	0.9232	0.9818	0.9898	0.9350	0.9612	0.9728
len	0.2143	0.2424	0.2465	0.2232	2.8778	2.8778	3.6696	2.4766	2.4766	2.7053
U,VS	0.9540	0.9684	1.0000	1.0000	0.9988	0.9988	0.9992	0.9536	0.9604	0.9684
len	0.2138	0.2318	0.1561	0.1041	4.2986	4.2986	4.5203	2.4946	2.4946	2.6027
U,MIX	0.9550	0.9650	1.0000	1.0000	0.9996	0.9994	0.9996	0.9524	0.9500	0.9588
len	0.2039	0.2179	0.1300	0.0886	4.5785	4.5785	4.7709	2.4969	2.4969	2.5748
E,full	0.9536	0.9550	0.9510	0.9548	0.9416	0.9892	0.9922	0.9444	0.9694	0.9792
len	0.2148	0.2423	0.2462	0.2232	2.8717	2.8717	3.6732	2.4759	2.4759	2.7095
E,VS	0.9638	0.9706	1.0000	1.0000	0.9994	0.9994	0.9996	0.9640	0.9664	0.9730
len	0.2143	0.2311	0.1549	0.1026	4.3259	4.3259	4.5460	2.4954	2.4954	2.6045
E,MIX	0.9600	0.9666	1.0000	1.0000	1.0000	0.9996	1.0000	0.9606	0.9590	0.9666
len	0.2038	0.2168	0.1281	0.0881	4.6094	4.6094	4.8001	2.4942	2.4942	2.5688

Table 8.16. MA(q) Model Selection, n=400,tstype=2,BB=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	0.9488	0.9534	0.9518	0.9518	0.9310	0.9822	0.9898	0.9318	0.9604	0.9726
len	0.2149	0.2421	0.2460	0.2226	2.8762	2.8762	3.6775	2.4779	2.4779	2.7089
N,VS	0.9550	0.9614	1.0000	1.0000	0.9988	0.9984	0.9992	0.9504	0.9566	0.9664
len	0.2146	0.2307	0.1549	0.1030	4.3214	4.3214	4.5400	2.4989	2.4989	2.6063
N,MIX	0.9524	0.9614	1.0000	1.0000	0.9994	0.9992	0.9998	0.9476	0.9500	0.9588
len	0.2040	0.2172	0.1287	0.0877	4.5978	4.5978	4.7907	2.4964	2.4964	2.5722
t,full	0.9518	0.9556	0.9580	0.9558	0.9392	0.9866	0.9922	0.9402	0.9624	0.9732
len	0.2144	0.2416	0.2459	0.2224	2.8808	2.8808	3.6508	2.4824	2.4824	2.7037
t,VS	0.9598	0.9660	1.0000	1.0000	0.9990	0.9990	0.9996	0.9560	0.9640	0.9700
len	0.2137	0.2301	0.1535	0.1020	4.3284	4.3284	4.5457	2.5020	2.5020	2.6072
t,MIX	0.9574	0.9606	1.0000	1.0000	0.9998	0.9994	1.0000	0.9542	0.9544	0.9618
len	0.2030	0.2160	0.1271	0.0869	4.6006	4.6006	4.7908	2.4954	2.4954	2.5711
U,full	0.9508	0.9570	0.9534	0.9424	0.9378	0.9868	0.9924	0.9434	0.9666	0.9776
len	0.2149	0.2425	0.2463	0.2230	2.8735	2.8735	3.6806	2.4793	2.4793	2.7110
U,VS	0.9558	0.9712	1.0000	1.0000	0.9996	0.9994	0.9998	0.9608	0.9672	0.9724
len	0.2147	0.2318	0.1554	0.1042	4.3270	4.3270	4.5486	2.5037	2.5037	2.6110
U,MIX	0.9578	0.9648	1.0000	1.0000	0.9998	0.9996	0.9998	0.9566	0.9546	0.9620
len	0.2042	0.2172	0.1288	0.0887	4.5980	4.5980	4.7890	2.4950	2.4950	2.5705
E,full	0.9592	0.9524	0.9568	0.9544	0.9354	0.9852	0.9926	0.9402	0.9642	0.9748
len	0.2128	0.2411	0.2453	0.2217	2.8778	2.8778	3.6727	2.4790	2.4790	2.7010
E,VS	0.9674	0.9636	1.0000	1.0000	0.9996	0.9996	0.9998	0.9566	0.9624	0.9690
len	0.2129	0.22940	0.1529	0.1019	4.3291	4.3291	4.5502	2.4972	2.4972	2.6035
E,MIX	0.9620	0.9600	1.0000	1.0000	1.0000	0.9998	1.0000	0.9492	0.9554	0.9620
len	0.2039	0.2153	0.1261	0.0870	4.5984	4.5984	4.7908	2.4964	2.4964	2.5685

Table 8.17. ARMA(p,q) Model Selection, n=100,tstype=1,BB=100, pmax=3,qmax=3

e	ϕ_1	ϕ_2	θ_1	θ_2	pr0	hyb0	br0	pr1	hyb1	br1
N,full	1.0000	0.9960	0.9987	0.9710	0.9446	0.9380	0.9690	0.9750	0.8930	0.9850
len	1.7815	1.7972	1.8944	1.5646	2.6425	2.6425	3.0200	3.6293	3.6293	4.1632
N,VS	0.9984	0.9880	0.9990	0.9996	0.9940	0.9370	0.9970	0.9700	0.9360	0.9740
len	1.7041	1.5181	1.8711	1.2561	2.9957	2.9957	3.3141	3.5298	3.5298	3.9126
N,MIX	1.0000	0.9954	1.0000	1.0000	0.9996	0.9400	0.9980	0.9640	0.9230	0.9684
len	1.670	1.3870	1.8374	1.2145	3.0949	3.0949	3.4249	3.5820	3.5820	4.0066
t,full	0.9990	0.9970	0.9980	0.9720	0.9470	0.9370	0.9690	0.9780	0.9030	0.9880
len	1.7409	1.7680	1.8459	1.4605	2.6431	2.6431	2.9968	3.6073	3.6073	4.1112
t,VS	1.0000	0.9910	0.9990	0.9990	0.9920	0.9500	0.9800	0.9750	0.9440	0.9790
len	1.6468	1.5045	1.7933	1.2020	3.0119	3.0119	3.3208	3.5367	3.5367	3.8934
t,MIX	1.0000	0.9960	1.0000	0.9990	0.9940	0.9530	0.9950	0.9710	0.9340	0.9710
len	1.6167	1.3712	1.7657	1.1535	3.1006	3.1006	3.4084	3.5951	3.5951	3.9902
U,full	1.0000	0.9970	0.9990	0.9720	0.9390	0.9300	0.9690	0.9720	0.8834	0.9820
len	1.7525	1.7836	1.8674	1.5279	2.6469	2.6469	3.0434	3.6338	3.6338	4.1503
U,VS	0.9990	0.9890	0.9980	0.9990	0.9920	0.9370	0.9920	0.9640	0.9302	0.9664
len	1.6953	1.5371	1.8588	1.2459	2.9955	2.9955	3.3181	3.5468	3.5468	3.9228
U,MIX	1.0000	0.9960	0.9990	1.0000	0.9950	0.9410	0.9960	0.9630	0.9200	0.9640
len	1.6517	1.3937	1.8284	1.2077	3.0910	3.0910	3.4208	3.6057	3.6057	4.0249
E,full	0.9990	0.9980	0.9980	0.9770	0.9530	0.9414	0.9720	0.9764	0.9004	0.9840
len	1.7786	1.8009	1.8947	1.5615	2.6464	2.6464	3.0157	3.6401	3.6401	4.1597
E,VS	0.9990	0.9910	0.9980	1.0000	0.9960	0.9450	0.9960	0.9760	0.9410	0.9750
len	1.7005	1.5247	1.8655	1.2601	3.0076	3.0076	3.3226	3.5313	3.5313	3.9129
E,MIX	0.9990	0.9980	0.9990	0.9990	0.9980	0.9510	0.9980	0.9700	0.9330	0.9700
len	1.6757	1.3942	1.8468	1.2200	3.1133	3.1133	3.4248	3.5935	3.5935	4.0066

CHAPTER 9

REAL DATA EXAMPLES

Common examples of random walk are stock prices. Consider the daily closing prices of major European stock indices: Germany DAX (Ibis), Switzerland SMI, France CAC, and UK FTSE. The data are sampled in business time, i.e., weekends and holidays are omitted. The EuStock-Markets dataset is a multivariate time series with 1860 observations on 4 variables. If we consider DAX the second indice of EuStockMarkets, the random walk looks good up to 1450. Since we want our errors to be scattered around y=0 need to consider the DAX data upto 1450 only, see below for the plot of the errors. The prediction interval for 1451st from the past 1 to 1450 is also given below.

Figure 9.1. plot of the errors for the rw dataset

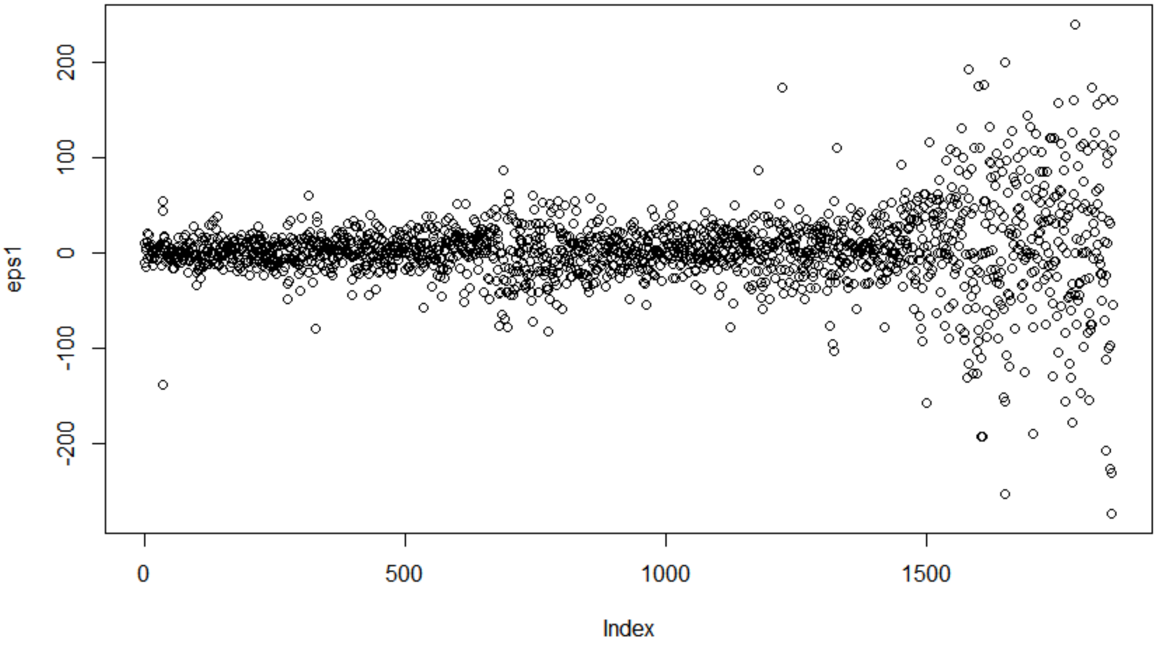
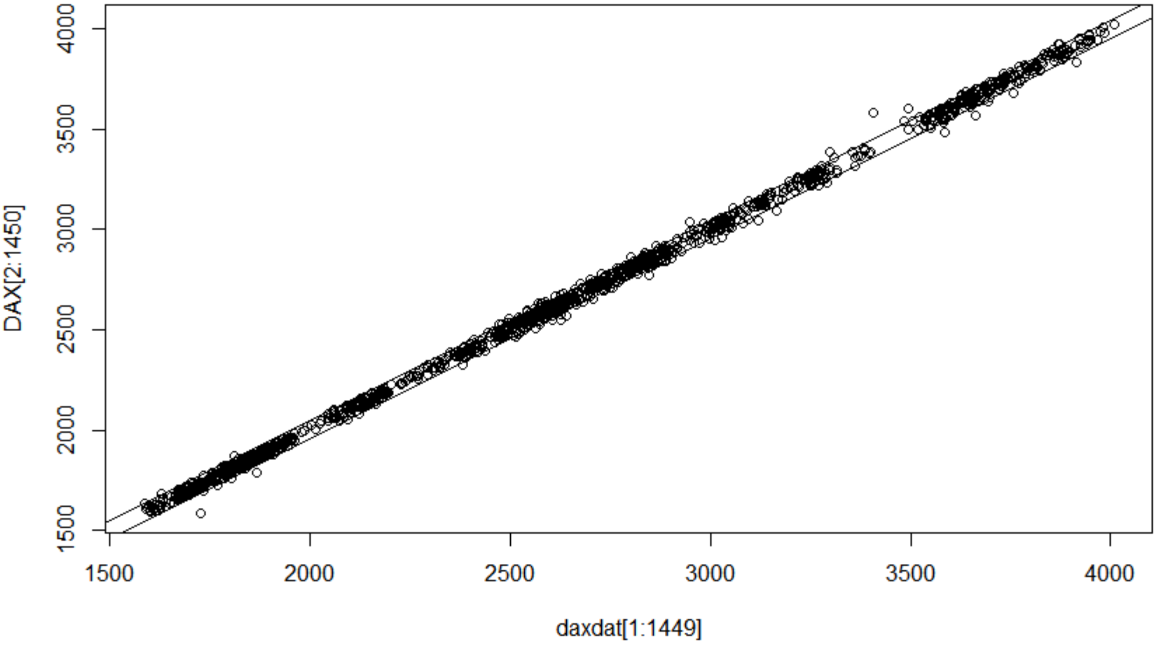


Figure 9.2. PI plot of rw dataset



The presidents and LakeHuron datasets in R can be used for Prediction Interval Illustrations. After careful analysis of model fitting: model specification, parameter estimation and model diagnostics for both data sets, it can be shown that presidents can be best fitted with AR(1) model and LakeHuron dataset can be best fitted with AR(2)model, below are ACF and PACF plots for these two datasets. The prediction interval obtained using locpi ignores the time series structure of the datasets, these are given by parallel lines. These intervals are wider than the ones produced using locpi2 which considers the time series structure of the datasets. To get the shorth of the residuals, missing values were omitted for the presidents dataset. In both cases the prediction intervals produced using locpi2 were shorter than the ones produced using locpi. Both datasets were divided into training data and test data, for presidents dataset I took the first 119 observations(1 left out) as my training data and for the LakeHuron the first 96 observations(2 left out).

Figure 9.3. ACF plot of presidents dataset

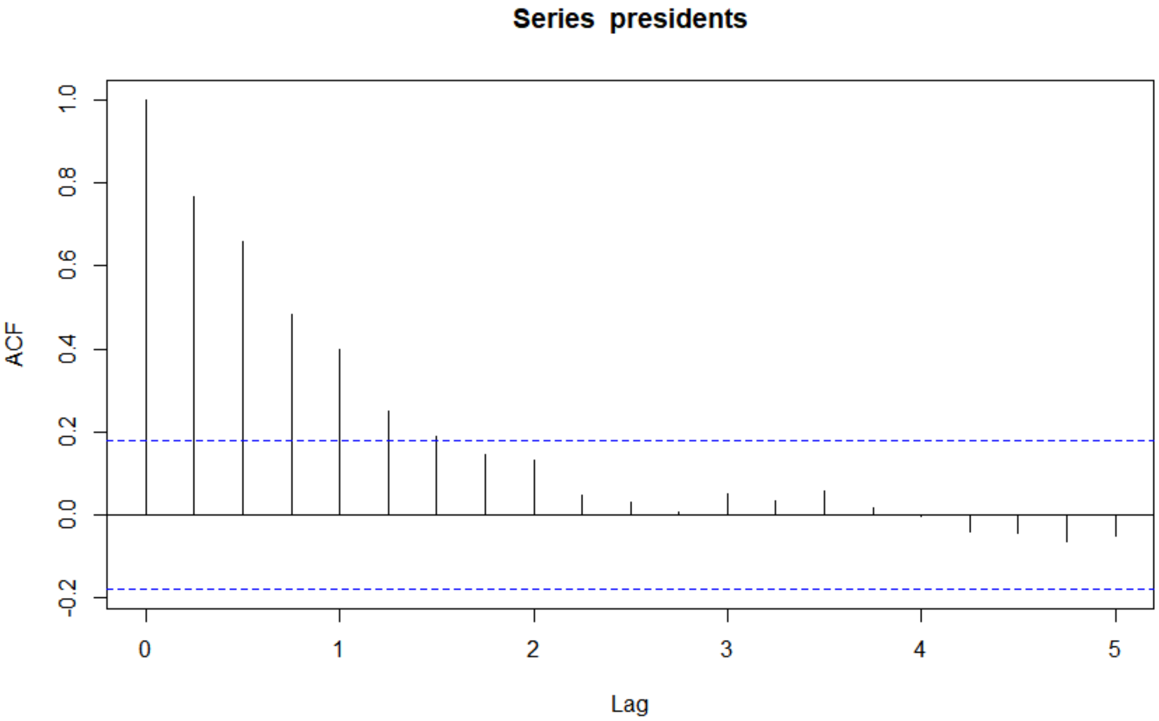


Figure 9.4. PACF plot of presidents dataset

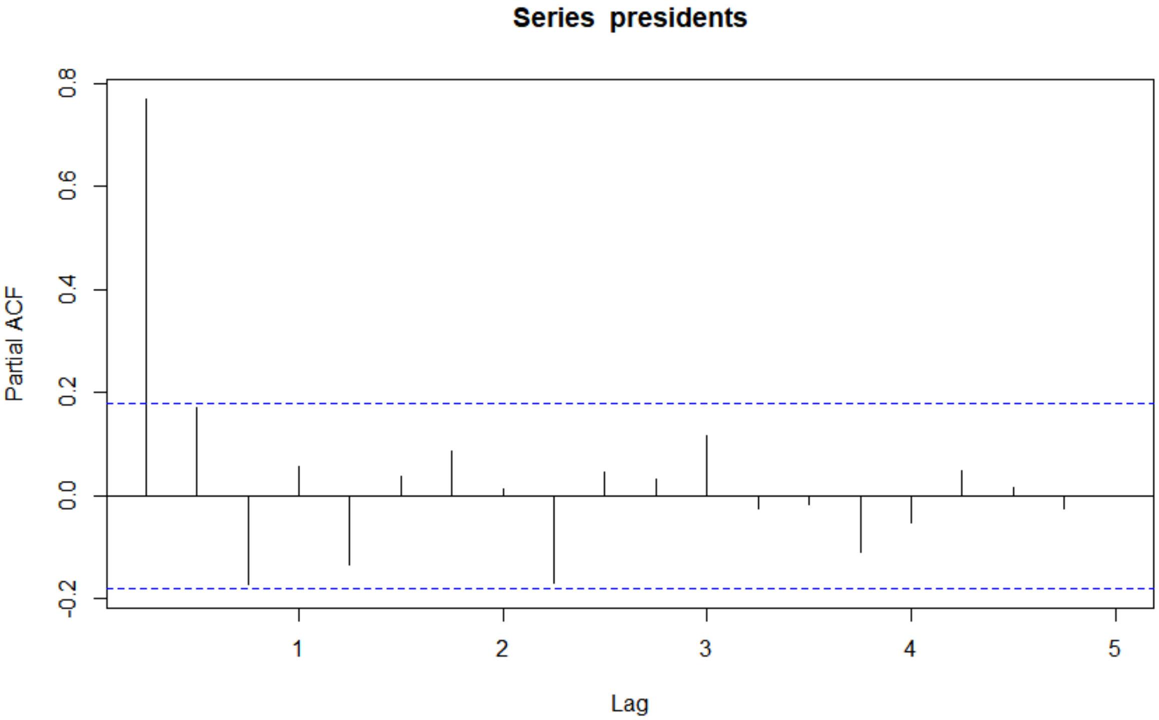


Figure 9.5. ACF plot of LakeHuron dataset

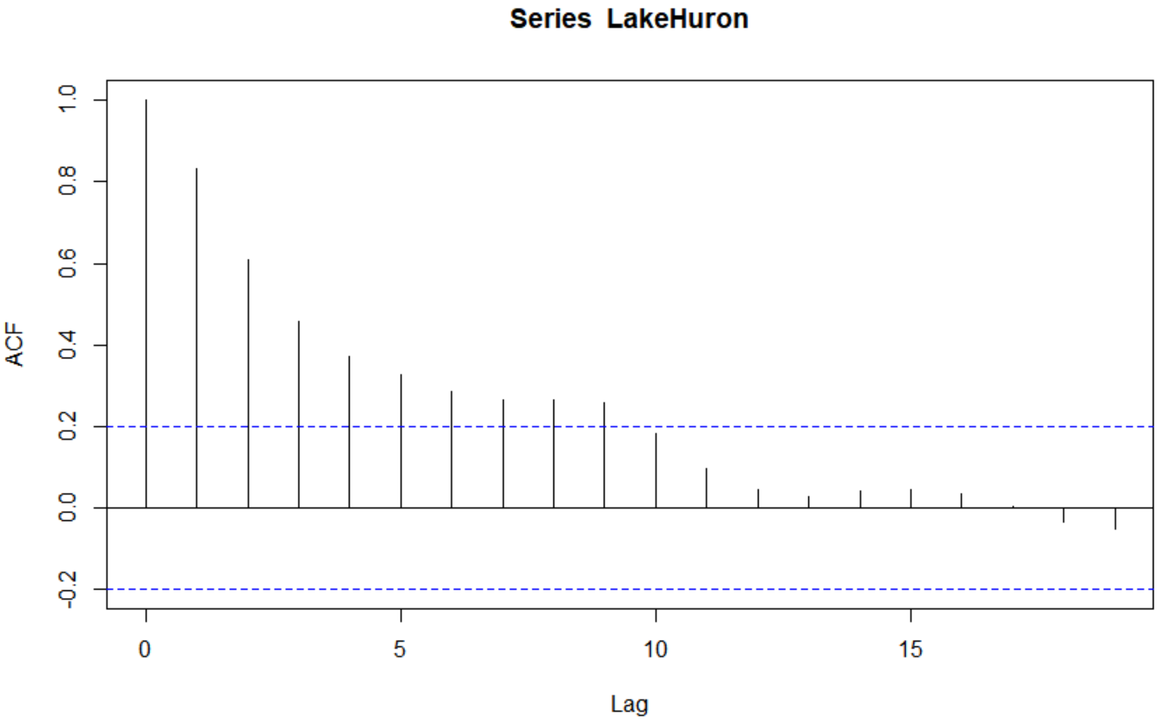
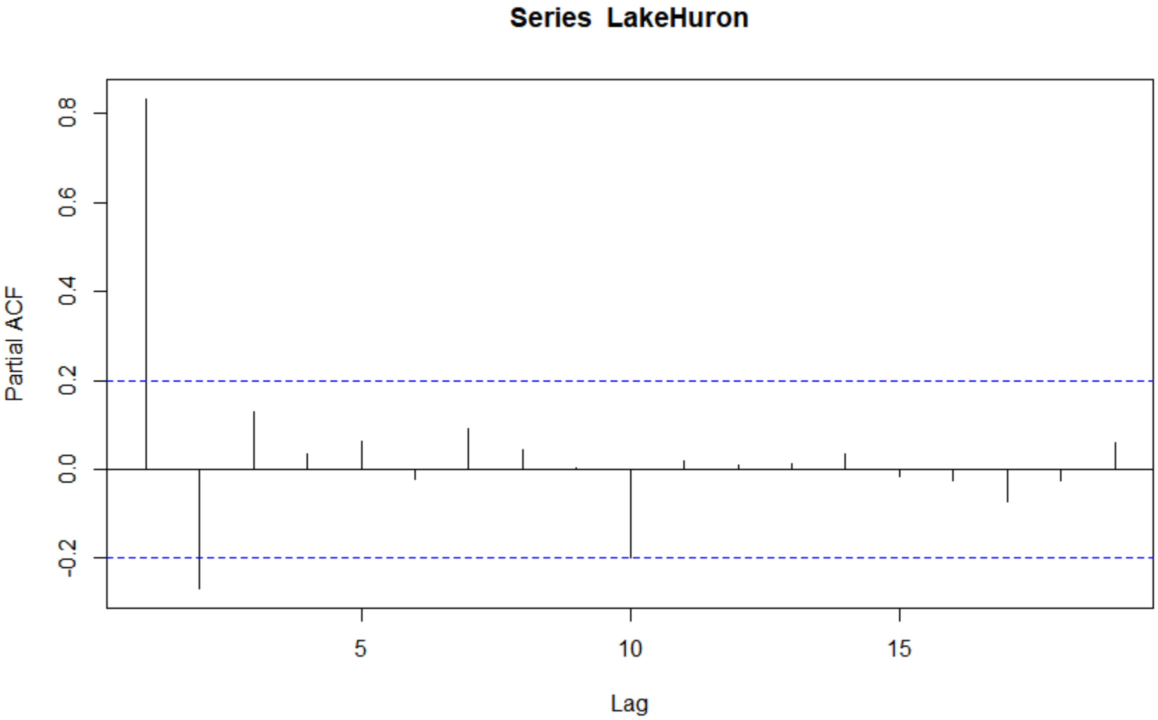


Figure 9.6. PACF plot of LakeHuron dataset



The function predict in R gives $\hat{Y}$ and se where the Chebychev Prediction Interval is given by $[\hat{Y} - 1.96se, \hat{Y} + 1.96se]$ for the 1 step ahead. Below are R outputs obtained using predict function for the two datasets.

For LakeHuron dataset

```
> predict(arima(dat, order = c(2,0,0)), n.ahead = 1)

#dat is the training data for the LakeHuron dataset

$pred

Time Series:

Start = 97

End = 97

Frequency = 1

[1] 579.1357


$se

Time Series:

Start = 97

End = 97

Frequency = 1

[1] 0.6948871


For presidents dataset

> predict(arima(Tdata, order = c(1,0,0)), n.ahead = 1)

$pred

Time Series:

Start = 120

End = 120

Frequency = 1
```

```
[1] 29.92367


$se
Time Series:
Start = 120
End = 120
Frequency = 1
[1] 9.272505
```

The following two plots show the prediction intervals produced using locpi and locpi2 for both datasets. The function Points in R is used to plot the prediction interval for the one that considers the time series structure and the function abline is used to plot the parallel lines.

Figure 9.7. PIs for LakeHuron dataset

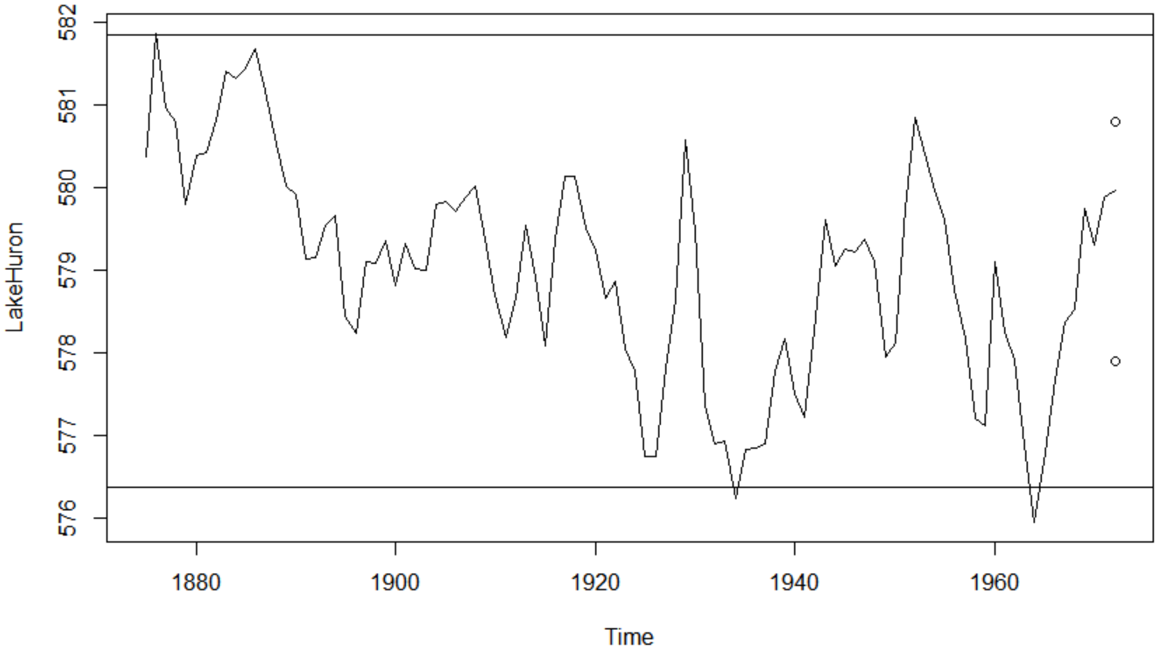
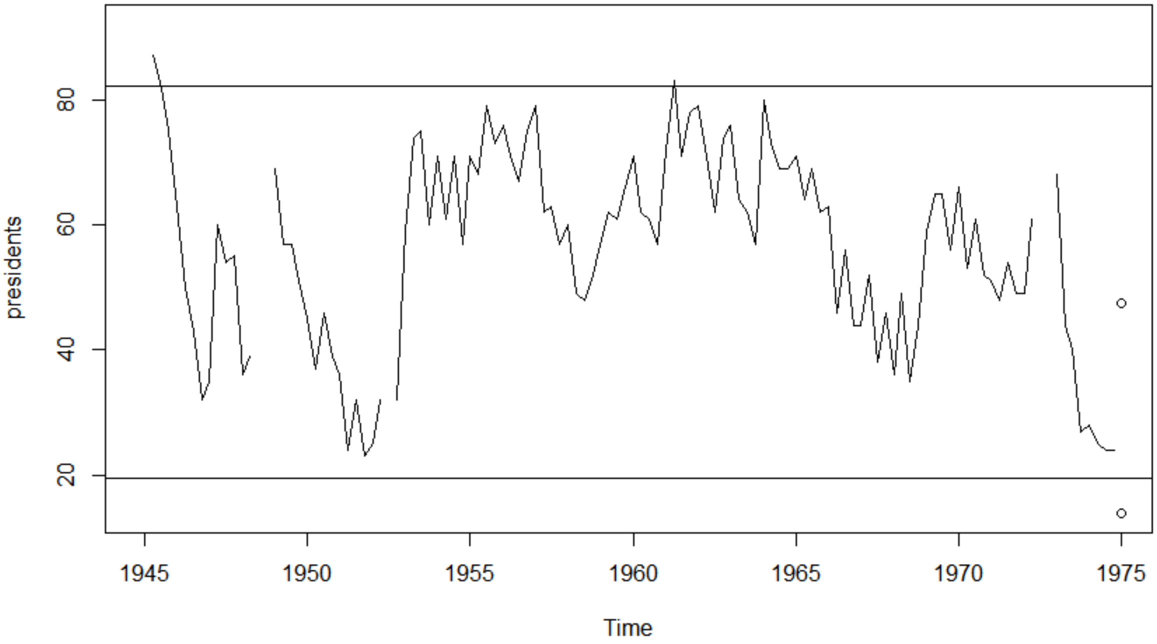


Figure 9.8. PIs for presidents dataset



R output for getting both prediction intervals for the LakeHuron dataset is given below, the codes for the presidents dataset are similar to this one.

```
> dat=LakeHuron[1:96]

> outAR2t=arima(dat,c(2,0,0))

> outAR2t


Call:

arima(x = dat, order = c(2, 0, 0))


Coefficients:

ar1      ar2  intercept

1.0477  -0.2570   579.0051

s.e.  0.0992   0.1017     0.3310
```

```
$sigma^2 estimated as 0.4829:  log likelihood = -101.94,  aic = 211.87


> locpi2(outAR2t$resid,k=2)

$LPI

[1] -1.227476

$UPI

[1] 1.65287


> tauhat=579.0051*(1-1.0477+0.2570)

> tauhat

[1] 121.1858

> yhat=121.1858+(1.0477*LakeHuron[96])-(0.2570*LakeHuron[95])

> yhat

[1] 579.1357

> locpi(dat)

$LPI

[1] 576.3737


$UPI

[1] 581.8533


> lnL=yhat-1.227476

> lnL

[1] 577.9082

> unL=yhat+1.65287

> unL
```

```
[1] 580.7886

> LakeHuron[97]

[1] 579.89


> plot.ts(LakeHuron)

> abline(576.3737,0)

> abline( 581.8533,0)

> points(1972,577.9082)

> points(1972,580.7886)
```

It should be noted that the $\hat{Y}$ obtained using the predict function in R is more or less the same as the one obtained after fitting our data and substituting the estimates for the parameters in the general model as shown above.

CHAPTER 10

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 (7.1) 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).

The correction factors for the prediction intervals of this paper compensate for estimation of the model parameters and model selection for moderate n. Hyndman and Athanasopoulos (2018, last paragraph of $\oint$8.8) note that ARIMA-based prediction intervals tend to be too narrow, so actual coverage is less than the nominal coverage. See Bhansali (1981) for the effects of estimating the order of the time series model.

There is a large literature on time series PIs, especially for AR(p) models, and the bootstrap is often used. See Alonso, Peńa, and Romo (2002, 2003), Brockwell and Davis (2016), Clements and Kim (2007), deLuna (2000), Hyndman and Athanasopoulos (2018), Kabaila and He (2007), Lu and Wang (2020), Pan and Politis (2016a), Pascual, Romo, and Ruiz (2001), Thombs and Schucany (1990), Vidoni (2009), and Wolf and Wunderli (2015) for references. Some papers on the shorth include Chen and Shao (1999), Grübel (1988), and Einmahl and Mason (1992). See Hong, Kuffner, and Martin (2018) for why classical PIs after AIC variable selection do not work.

Some prediction intervals for stochastic processes include Pan and Politis (2016b), Vidoni (2004), and Vit (1973). Mykland (2003) described how to convert prediction regions into investment strategies. Pankratz (1983, p. 106) notes that the random walk model has been found to be a good model for many stock price time series.

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 `msarsim` simulates AR model selection using the Yule Walker equations with AIC and the *R* function `ar.yw`. The *tspack* function `msmasim` simulates MA model selection using the GMLE with AIC_C using the *R* function `auto.arima` from the Hyndman and Khandakar (2008) *forecast package*. Also see Hyndman and Athanasopoulos (2018).

The aicmatrix is also somewhat useful for GARCH models. The aicmatrix is made in one of the *R* time series help files. Using I_I and submodels helps to quickly find a small number of good models to examine. The function `aicmat` makes the aicmatrix for ARIMA(p, d, q) models with d fixed while the function `saics` makes the aicmatrix for ARIMA$(p, d, q) \times (P, D, Q)_s$ models with d, P, D, Q and s fixed. The function `pimasim` was used to simulate the prediction intervals. The *tspack* function `pitsvssim` simulates PI (3.6) after model selection using the GMLE with AIC_C using the *R* function `auto.arima` from the Hyndman and Khandakar (2008) *forecast package*. Also see Hyndman and Athanasopoulos (2018). The *tspack* function `rwpisim` was used for the random walk simulation.

REFERENCES

[1] 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.

[2] Alonso, A.M., Peńa, D., and Romo, J. (2002), “Forecasting Time Series With Sieve Bootstrap,” *Journal of Statistical Planning and Inference*, 100, 1-11.

[3] Alonso, A.M., Peńa, D., and Romo, J. (2003), “On Sieve Bootstrap Prediction Intervals,” *Statistical & Probability Letters*, 65, 13-20.

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

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

[6] 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.

[7] 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.

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

[9] Breiman, L. (1996), “Bagging Predictors,” *Machine Learning*, 24, 123-140.

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

[11] Brockwell, P.J., and Davis, R.A. (2016), *Introduction to Time Series and Forecasting*, 3rd ed., Springer, New York, NY.

[12] Bühlmann, P. (1994), “Blockwise Bootstrapped Empirical Process for Stationary Sequence,” *The Annals of Statistics*, 22, 995-1012.

[13] Bühlmann, P. (1997), “Sieve Bootstrap for Time Series,” *Bernoulli*, 3, 5123-148.

[14] Bühlmann, P. (2002), “Bootstraps for Time Series,” *Statistical Science*, 17, 52-72.

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

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

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

[18] Chen, M.H., and Shao, Q.M. (1999), “Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistics* 8, 69-92.

[19] Chen, S.X. (2016), “Peter Hall’s Contributions to the Bootstrap,” *The Annals of Statistics*, 44, 1821-1836.

[20] Chew, V. (1966), “Confidence, Prediction and Tolerance Regions for the Multivariate Normal Distribution,” *Journal of the American Statistical Association,* 61, 605-617.

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

[22] Clements, M.P., and Kim, N. (2007), “Bootstrapping Prediction Intervals for Autoregressive Time Series,” *Computational Statistics & Data Analysis*, 51, 3580-3594.

[23] deLuna, X. (2000), “Prediction Intervals Based on Autoregressive Forecasts,” *Journal of the Royal Statistical Society, D,* 49, 87-93.

[24] 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.

[25] Durbin, J. (1959), “Efficient Estimation of Parameters in Moving-Average Models,” *Biometrika*, 46, 306-316.

[26] Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM, Philadelphia, PA.

[27] Efron, B. (2014), “Estimation and Accuracy After Model Selection,” (with discussion), *Journal of the American Statistical Association,* 109, 991-1007.

[28] Einmahl, J.H.J., and Mason, D.M. (1992), “Generalized Quantile Processes,” *The Annals of Statistics*, 20, 1062-1078.

[29] Freedman, D.A. (1981), “Bootstrapping Regression Models,” *The Annals of Statistics*, 9, 1218-1228.

[30] Frey, J. (2013), “Data-Driven Nonparametric Prediction Intervals,” *Journal of Statistical Planning and Inference*, 143, 1039-1048.

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

[32] Gray, H.L., and Odell P.L., (1966), “On Sums and Products of Rectangular Variates,” *Biometrika*, 53, 615-617.

[33] Grübel, R. (1988), “The Length of the Shorth,” *The Annals of Statistics,* 16, 619-628.

[34] Hall, P. (1988), “Theoretical Comparisons of Bootstrap Confidence Intervals,” (with discussion), *The Annals of Statistics*, 16, 927-985.

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

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

[37] Hannan, E.J. (1980), “The Estimation of the Order of an ARMA Process,” *The Annals of Statistics*, 8, 1071-1081.

[38] Hannan, E.J., and Quinn, B.G. (1979), “The Determination of the Order of an Autoregression,” *Jouurnal of the Royal Statistical Society, B*, 41, 190-195.

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

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

[41] Hong, L., Kuffner, T.A., and Martin, R. (2018), “On Overfitting and Post-Selection Uncertainty Assessments,” *Biometrika*, 105, 221-224.

[42] Hurvich, C., and Tsai, C.L. (1989), “Regression and Time Series Model Selection in Small

Samples," *Biometrika*, 76, 297-307.

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

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

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

[46] Kabaila, P., and He, Z. (2007), "Improved Prediction Limits for AR(p) and ARC(p) Processes," *Journal of Time Series Analysis*, 29, 213-223.

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

[47] 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.

[48] 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.

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

[50] Lee, Y.S., and Scholtes, S. (2014), "Empirical Prediction Intervals Revisited," *International Journal of Forecasting*, 30, 217-234.

[51] 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.

[52] 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.

[53] Lu, X., and Wang, L. (2020), "Bootstrap Prediction Interval for ARMA Models with Unknown Orders," *Revstat-Statistical Journal*, 18, 375-396.

[54] Machado, J.A.F., and Parente, P. (2005), "Bootstrap Estimation of Covariance Matrices Via the Percentile Method," *Econometrics Journal,* 8, 70–78.

[55] Mallows, C. (1973), “Some Comments on C_p,” *Technometrics,* 15, 661-676.

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

[57] Marengo, J.E., Farnsworth, D.L., and Stefanic, L. (2017), “A Geometric Derivation of the Irwin-Hall Distribution,” *International Journal of Mathematics and Mathematical Sciences*, 2017, online.

[58] Masters, T. (1995), *Neural, Novel, & Hybrid Algorithms for Time Series Prediction*, Wiley, New York, NY.

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

[60] Mykland, P.A. (2003), “Financial Options and Statistical Prediction Intervals,” *The Annals of Statistics*, 31, 1413-1438.

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

[62] Olive, D.J. (2007), “Prediction Intervals for Regression Models,” *Computational Statistics & Data Analysis,* 51, 3115-3122.

[63] Olive, D.J. (2013), “Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data,” *International Journal of Statistics and Probability*, 2, 90-100.

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

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

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

[67] Olive, D.J. (2018), “Applications of Hyperellipsoidal Prediction Regions,” *Statistical Papers*, 59, 913-931.

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

[69] Olive, D. J., and Hawkins, D. M. (2005), “Variable Selection for 1D Regression Models,” *Technometrics*, 47, 43-50.

[70] Olive, D.J., Rathnayake, R.C., and Haile, M.G. (2021), "Prediction Intervals for GLMs, GAMs, and Some Survival Regression Models," *Communications in Statistics: Theory and Methods*, to appear.

[71] Pan, L., and Politis, D.N. (2016a), "Bootstrap Prediction Intervals for Markov Processes," *Computational Statistics & Data Analysis*, 100, 467-494.

[72] Pan, L., and Politis, D.N. (2016b), "Bootstrap Prediction Intervals for Linear, Nonlinear, and Nonparametric Autoregressions," *Journal of Statistical Planning and Inference*, 177, 1-27.

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

[74] Pascual, L., Romo, J., and Ruiz, E., (2004),"Bootstrap Predictive Inference for ARIMA Processes," *Journal of Time Series Analysis*, 25, 449-465.

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

[76] 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.

[77] Politis, D.N. (2003), "The Impact of Bootstrap Methods on Time Series Analysis," *Statistical Science*, 18, 219-230.

[78] 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.

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

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

[81] Rathnayake, R.C., and Olive, D.J. (2021), "Bootstrapping Some GLMs and Survival Regression Models After Variable Selection, *"Communications in Statistics: Theory and Methods*, to appear.

[82] R Core Team (2016), “R: a Language and Environment for Statistical Computing,” R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

[83] 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.

[84] Roach, S.A. (1963), “The Frequency Distribution of the Sample Mean Where Each Member of the Sample is Drawn from a Different Rectangular Distribution,” *Biometrika*, 50, 508-513.

[85] Ross, S., M. (2014), *Introduction to Probability Models*, 11th ed., Academic Press, San Diego, CA.

[86] Schwarz, G. (1978), “Estimating the Dimension of a Model,” *The Annals of Statistics,* 6, 461-464.

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

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

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

[90] Shibata, R. (1976), “Selection of the Order of an Autoregressive Model by Akaike’s Information Criterion,” *Biometrika*, 63, 117-126.

[91] Su, Z., and Cook, R.D. (2012), “Inner Envelopes: Efficient Estimation in Multivariate Linear Regression,” *Biometrika*, 99, 687-702.

[92] Thombs, L.A. and Schucany, W.R. (1990), “Bootstrap Prediction Intervals for Autoregression,” *Journal of the American Statistical Association*, 85, 486-492.

[93] Vidoni, P. (2004), “Improved Prediction Intervals for Stochastic Process Models,” *Journal of Time Series Analysis*, 25, 137-154.

[94] Vidoni, P. (2009), “A Simple Procedure for Computing Improved Prediction Intervals for Autoregressive Models,” *Journal of Time Series Analysis*, 30, 577-590.

[95] Vit, P. (1973), “Interval Prediction for a Poisson Process,” *Biometrika*, 60, 667-668.

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

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

[98] Wolf, M., and Wunderli, D. (2015), "Bootstrap Joint Prediction Regions," *Journal of Time Series Analysis*, 36, 352-376.

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

[100] 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.

VITA

Graduate School
Southern Illinois University

Mulubrhan G. Haile

gmulubrhan@gmail.com

University of Asmara
Bachelor of Science, July 2009

Southern Illinois University at Carbondale
Master of Science, Mathematics, May 2017

Special Honors and Awards:
Dissertation Research Assistantship Award (Fall 2021)

Dissertation Paper Title:
Inference for Time Series after Variable Selection

Major Professor: Dr. David Olive

Publications:
Olive, D.J., Rathnayake, R.C., and Haile, M.G. (2021), “Prediction Intervals for GLMs, GAMs, and Some Survival Regression Models,” *Communications in Statistics: Theory and Methods*, to appear. https://doi.org/10.1080/03610926.2021.1887238