STAT 626: Outline of Lecture 12
Forecasting Based on the Infinite Past (§3.5)

1. **Forecasting a One-sided MA(∞) Time Series; Infinite Past**

2. Example 3.22: Long-Range Forecasts,

3. **Prediction Error Variance (P_{n+1}^n) and Forecast Interval:**

$$\tilde{x}_{n+1}^n \pm 1.96\sqrt{P_{n+1}^n}$$

4. Example 3.23: Forecasting an ARMA(1,1) Series

5. **Example 3.24 Forecasting the Recruitment Series**

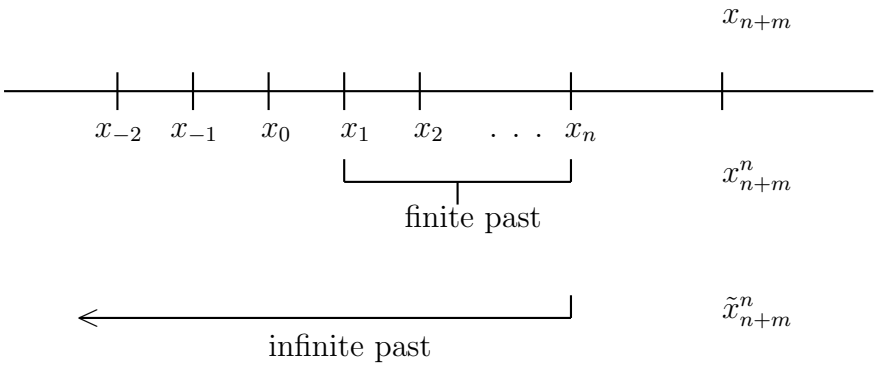
6. Example 3.25: Backcasting an ARMA(1,1) Model

7. **Reading Assignment: Examples 3.36, Random Walk, and 3.37, IMA (1,1) and EWMA: Holt-Winter's Method.**
   These simple models are extremely important in applied forecasting.

Forecasting a Stationary Series Based on the Infinite Past

Assumption: The ACF or the parameters, and the past are known.

A pictorial setup for forecasting the future values $x_{n+m}, m = 1, 2, \ldots$:



(a) What are the forecast, forecast error, and forecast error variance?

$$x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{n+m-j},$$

Forecast: $\tilde{x}^n_{n+m} = \sum_{j=m}^{\infty} \psi_j w_{n+m-j}$.

Forecast error: $x_{n+m} - \tilde{x}^n_{n+m} = \sum_{j=0}^{m-1} \psi_j w_{n+m-j}$

Error variance: $P^n_{n+m} = \mathrm{Var}(x_{n+m} - \tilde{x}^n_{n+m}) = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2$

(b) Their 95% forecast intervals?

$$\tilde{x}_{n+m}^n \pm 1.96\sqrt{P_{n+m}^n}.$$

EXAMPLES:

1. Forecasting a Causal AR(1) Model

$$x_t = \phi x_{t-1} + w_t.$$

Its predictor?

Its prediction error variance?

Its 95% forecast interval?

2. Forecasting Causal ARMA Models:

3. One-Sided MA(∞) Processes:

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

with absolutely summable coefficients.

The forecast? The prediction error? Its variance?

The General Principle of Statistical Forecasting

(a) Forecating: Begins when a good model is identified for the time series,

(b) Given the time series data $x_1, \ldots, x_n$: **What are the principles for model-based forecasting ?**

$$x_t = f(\beta, \text{Past of the Series}) + w_t.$$

Example:

$$x_t = \phi x_{t-1} + w_t.$$

Principle: Replace the unknowns by their best ESTIMATES.
Example:

$$x_t = w_t + \theta w_{t-1}.$$

(c) **Forecasting ARMA Models**

Recall that causal ARMA models can be written as One-Sided MA(∞) of a white noise:

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

and **invertible** ARMA models can be written as One-Sided AR(∞);

$$x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} + w_t.$$

Example 3.21 Prediction for an MA(1)

The innovations algorithm lends itself well to prediction for moving average processes. Consider an MA(1) model, $x_t = w_t + \theta w_{t-1}$. Recall that $\gamma(0) = (1+\theta^2)\sigma_w^2$, $\gamma(1) = \theta\sigma_w^2$, and $\gamma(h) = 0$ for $h > 1$. Then, using Property 3.6, we have

$$\begin{aligned} \theta_{n1} &= \theta\sigma_w^2/P_n^{n-1} \\ \theta_{nj} &= 0, \quad j = 2, \ldots, n \\ P_1^0 &= (1+\theta^2)\sigma_w^2 \\ P_{n+1}^n &= (1+\theta^2 - \theta\theta_{n1})\sigma_w^2. \end{aligned}$$

Finally, from (3.77), the one-step-ahead predictor is

$$x_{n+1}^n = \theta\left(x_n - x_n^{n-1}\right)\sigma_w^2/P_n^{n-1}.$$

Forecasting ARMA Processes

The general prediction equations (3.60) provide little insight into forecasting for ARMA models in general. There are a number of different ways to express these forecasts, and each aids in understanding the special structure of ARMA prediction. Throughout, we assume x_t is a causal and invertible ARMA(p,q) process, $\phi(B)x_t = \theta(B)w_t$, where $w_t \sim$ iid N$(0,\sigma_w^2)$. In the non-zero mean case, $E(x_t) = \mu_x$, simply replace x_t with $x_t - \mu_x$ in the model. First, we consider two types of forecasts. We write x_{n+m}^n to mean the minimum mean square error predictor of x_{n+m} based on the data $\{x_n, \ldots, x_1\}$, that is,

$$x_{n+m}^n = E(x_{n+m} \mid x_n, \ldots, x_1).$$

For ARMA models, it is easier to calculate the predictor of x_{n+m}, assuming we have the complete history of the process $\{x_n, x_{n-1}, \ldots, x_1, x_0, x_{-1}, \ldots\}$. We will denote the predictor of x_{n+m} *based on the infinite past* as

$$\widetilde{x}_{n+m} = E(x_{n+m} \mid x_n, x_{n-1}, \ldots, x_1, x_0, x_{-1}, \ldots).$$

In general, x_{n+m}^n and $\widetilde{x}_{n+m}$ are not the same, but the idea here is that, for large samples, $\widetilde{x}_{n+m}$ will provide a good approximation to x_{n+m}^n.

Now, write x_{n+m} in its causal and invertible forms:

$$x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{n+m-j}, \quad \psi_0 = 1 \tag{3.82}$$

$$w_{n+m} = \sum_{j=0}^{\infty} \pi_j x_{n+m-j}, \quad \pi_0 = 1. \tag{3.83}$$

Then, taking conditional expectations in (3.82), we have

$$\widetilde{x}_{n+m} = \sum_{j=0}^{\infty} \psi_j \widetilde{w}_{n+m-j} = \sum_{j=m}^{\infty} \psi_j w_{n+m-j}, \tag{3.84}$$

because, by causality and invertibility,

$$\widetilde{w}_t = E(w_t \mid x_n, x_{n-1}, \ldots, x_0, x_{-1}, \ldots) = \begin{cases} 0 & t > n \\ w_t & t \leq n. \end{cases}$$

Similarly, taking conditional expectations in (3.83), we have

$$0 = \widetilde{x}_{n+m} + \sum_{j=1}^{\infty} \pi_j \widetilde{x}_{n+m-j},$$

or

$$\widetilde{x}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \widetilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j}, \tag{3.85}$$

using the fact $E(x_t \mid x_n, x_{n-1}, \ldots, x_0, x_{-1}, \ldots) = x_t$, for $t \leq n$. Prediction is accomplished recursively using (3.85), starting with the one-step-ahead predictor, $m = 1$, and then continuing for $m = 2, 3, \ldots$. Using (3.84), we can write

$$x_{n+m} - \widetilde{x}_{n+m} = \sum_{j=0}^{m-1} \psi_j w_{n+m-j},$$

so the mean-square prediction error can be written as

$$P_{n+m}^n = E(x_{n+m} - \widetilde{x}_{n+m})^2 = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2. \tag{3.86}$$

Also, we note, for a fixed sample size, n, the prediction errors are correlated. That is, for $k \geq 1$,

$$E\{(x_{n+m} - \widetilde{x}_{n+m})(x_{n+m+k} - \widetilde{x}_{n+m+k})\} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j \psi_{j+k}. \tag{3.87}$$

Example 3.22 Long-Range Forecasts

Consider forecasting an ARMA process with mean μ_x. Replacing x_{n+m} with $x_{n+m} - \mu_x$ in (3.82), and taking conditional expectation as is in (3.84), we deduce that the m-step-ahead forecast can be written as

$$\widetilde{x}_{n+m} = \mu_x + \sum_{j=m}^{\infty} \psi_j w_{n+m-j}. \tag{3.88}$$

Noting that the ψ-weights dampen to zero exponentially fast, it is clear that

$$\widetilde{x}_{n+m} \to \mu_x \tag{3.89}$$

exponentially fast (in the mean square sense) as $m \to \infty$. Moreover, by (3.86), the mean square prediction error

$$P^n_{n+m} \to \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2 = \gamma_x(0) = \sigma_x^2, \tag{3.90}$$

exponentially fast as $m \to \infty$; recall (3.45).

It should be clear from (3.89) and (3.90) that ARMA forecasts quickly settle to the mean with a constant prediction error as the forecast horizon, m, grows. This effect can be seen in Figure 3.6 on page 119 where the Recruitment series is forecast for 24 months; see Example 3.24.

When n is small, the general prediction equations (3.60) can be used easily. When n is large, we would use (3.85) by truncating, because we do not observe $x_0, x_{-1}, x_{-2}, \ldots$, and only the data $x_1, x_2, \ldots, x_n$ are available. In this case, we can truncate (3.85) by setting $\sum_{j=n+m}^{\infty} \pi_j x_{n+m-j} = 0$. The truncated predictor is then written as

$$\widetilde{x}^n_{n+m} = -\sum_{j=1}^{m-1} \pi_j \widetilde{x}^n_{n+m-j} - \sum_{j=m}^{n+m-1} \pi_j x_{n+m-j}, \tag{3.91}$$

which is also calculated recursively, $m = 1, 2, \ldots$. The mean square prediction error, in this case, is approximated using (3.86).

For AR(p) models, and when $n > p$, equation (3.67) yields the exact predictor, x^n_{n+m}, of x_{n+m}, and there is no need for approximations. That is, for $n > p$, $\widetilde{x}^n_{n+m} = \widetilde{x}_{n+m} = x^n_{n+m}$. Also, in this case, the one-step-ahead prediction error is $E(x_{n+1} - x^n_{n+1})^2 = \sigma_w^2$. For pure MA($q$) or ARMA($p, q$) models, truncated prediction has a fairly nice form.

Property 3.7 Truncated Prediction for ARMA

For ARMA(p, q) models, the truncated predictors for $m = 1, 2, \ldots,$ are

$$\widetilde{x}^n_{n+m} = \phi_1 \widetilde{x}^n_{n+m-1} + \cdots + \phi_p \widetilde{x}^n_{n+m-p} + \theta_1 \widetilde{w}^n_{n+m-1} + \cdots + \theta_q \widetilde{w}^n_{n+m-q}, \tag{3.92}$$

where $\widetilde{x}^n_t = x_t$ for $1 \le t \le n$ and $\widetilde{x}^n_t = 0$ for $t \le 0$. The truncated prediction errors are given by: $\widetilde{w}^n_t = 0$ for $t \le 0$ or $t > n$, and

$$\widetilde{w}^n_t = \phi(B)\widetilde{x}^n_t - \theta_1 \widetilde{w}^n_{t-1} - \cdots - \theta_q \widetilde{w}^n_{t-q}$$

for $1 \le t \le n$.

Example 3.23 Forecasting an ARMA$(1,1)$ Series
Given data $x_1, \ldots, x_n$, for forecasting purposes, write the model as

$$x_{n+1} = \phi x_n + w_{n+1} + \theta w_n.$$

Then, based on (3.92), the one-step-ahead truncated forecast is

$$\widetilde{x}^n_{n+1} = \phi x_n + 0 + \theta \widetilde{w}^n_n.$$

For $m \geq 2$, we have

$$\widetilde{x}^n_{n+m} = \phi \widetilde{x}^n_{n+m-1},$$

which can be calculated recursively, $m = 2, 3, \ldots$.

To calculate $\widetilde{w}^n_n$, which is needed to initialize the successive forecasts, the model can be written as $w_t = x_t - \phi x_{t-1} - \theta w_{t-1}$ for $t = 1, \ldots, n$. For truncated forecasting using (3.92), put $\widetilde{w}^n_0 = 0$, $x_0 = 0$, and then iterate the errors forward in time

$$\widetilde{w}^n_t = x_t - \phi x_{t-1} - \theta \widetilde{w}^n_{t-1}, \quad t = 1, \ldots, n.$$

The approximate forecast variance is computed from (3.86) using the ψ-weights determined as in Example 3.11. In particular, the ψ-weights satisfy $\psi_j = (\phi + \theta)\phi^{j-1}$, for $j \geq 1$. This result gives

$$P^n_{n+m} = \sigma_w^2 \left[1 + (\phi + \theta)^2 \sum_{j=1}^{m-1} \phi^{2(j-1)} \right] = \sigma_w^2 \left[1 + \frac{(\phi + \theta)^2 (1 - \phi^{2(m-1)})}{(1 - \phi^2)} \right].$$

To assess the precision of the forecasts, prediction intervals are typically calculated along with the forecasts. In general, $(1-\alpha)$ prediction intervals are of the form

$$x^n_{n+m} \pm c_{\frac{\alpha}{2}} \sqrt{P^n_{n+m}}, \tag{3.93}$$

where $c_{\alpha/2}$ is chosen to get the desired degree of confidence. For example, if the process is Gaussian, then choosing $c_{\alpha/2} = 2$ will yield an approximate 95% prediction interval for x_{n+m}. If we are interested in establishing prediction intervals over more than one time period, then $c_{\alpha/2}$ should be adjusted appropriately, for example, by using Bonferroni's inequality [see (4.55) in Chapter 4 or Johnson and Wichern, 1992, Chapter 5].

Example 3.24 Forecasting the Recruitment Series
Using the parameter estimates as the actual parameter values, Figure 3.6 shows the result of forecasting the Recruitment series given in Example 3.17 over a 24-month horizon, $m = 1, 2, \ldots, 24$. The actual forecasts are calculated as

$$x^n_{n+m} = 6.74 + 1.35 x^n_{n+m-1} - .46 x^n_{n+m-2}$$

for $n = 453$ and $m = 1, 2, \ldots, 12$. Recall that $x^s_t = x_t$ when $t \leq s$. The forecasts errors P^n_{n+m} are calculated using (3.86). Recall that $\widehat{\sigma}^2_w = 89.72$,

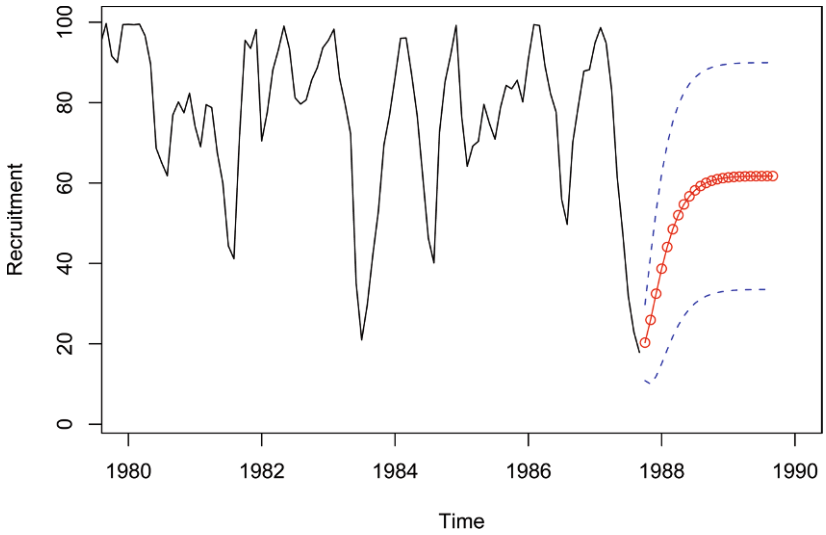


Fig. 3.6. Twenty-four month forecasts for the Recruitment series. The actual data shown are from about January 1980 to September 1987, and then the forecasts plus and minus one standard error are displayed.

and using (3.40) from Example 3.11, we have $\psi_j = 1.35\psi_{j-1} - .46\psi_{j-2}$ for $j \geq 2$, where $\psi_0 = 1$ and $\psi_1 = 1.35$. Thus, for $n = 453$,

$$\begin{aligned} P^n_{n+1} &= 89.72, \\ P^n_{n+2} &= 89.72(1 + 1.35^2), \\ P^n_{n+3} &= 89.72(1 + 1.35^2 + [1.35^2 - .46]^2), \end{aligned}$$

and so on.

Note how the forecast levels off quickly and the prediction intervals are wide, even though in this case the forecast limits are only based on one standard error; that is, $x^n_{n+m} \pm \sqrt{P^n_{n+m}}$.

To reproduce the analysis and Figure 3.6, use the following commands:

```
regr = ar.ols(rec, order=2, demean=FALSE, intercept=TRUE)
fore = predict(regr, n.ahead=24)
ts.plot(rec, fore$pred, col=1:2, xlim=c(1980,1990),
    ylab="Recruitment")
lines(fore$pred, type="p", col=2)
lines(fore$pred+fore$se, lty="dashed", col=4)
lines(fore$pred-fore$se, lty="dashed", col=4)
```

We complete this section with a brief discussion of backcasting. In backcasting, we want to predict x_{1-m}, for $m = 1, 2, \ldots$, based on the data $\{x_1, \ldots, x_n\}$. Write the backcast as

$$x_{1-m}^n = \sum_{j=1}^{n} \alpha_j x_j. \tag{3.94}$$

Analogous to (3.74), the prediction equations (assuming $\mu_x = 0$) are

$$\sum_{j=1}^{n} \alpha_j E(x_j x_k) = E(x_{1-m} x_k), \quad k = 1, \ldots, n, \tag{3.95}$$

or

$$\sum_{j=1}^{n} \alpha_j \gamma(k-j) = \gamma(m+k-1), \quad k = 1, \ldots, n. \tag{3.96}$$

These equations are precisely the prediction equations for forward prediction. That is, $\alpha_j \equiv \phi_{nj}^{(m)}$, for $j = 1, \ldots, n$, where the $\phi_{nj}^{(m)}$ are given by (3.75). Finally, the backcasts are given by

$$x_{1-m}^n = \phi_{n1}^{(m)} x_1 + \cdots + \phi_{nn}^{(m)} x_n, \quad m = 1, 2, \ldots. \tag{3.97}$$

Example 3.25 Backcasting an ARMA$(1,1)$

Consider an ARMA$(1,1)$ process, $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$; we will call this the *forward model*. We have just seen that best linear prediction backward in time is the same as best linear prediction forward in time for stationary models. Because we are assuming ARMA models are Gaussian, we also have that minimum mean square error prediction backward in time is the same as forward in time for ARMA models.[4] Thus, the process can equivalently be generated by the *backward model*,

$$x_t = \phi x_{t+1} + \theta v_{t+1} + v_t,$$

where $\{v_t\}$ is a Gaussian white noise process with variance σ_w^2. We may write $x_t = \sum_{j=0}^{\infty} \psi_j v_{t+j}$, where $\psi_0 = 1$; this means that x_t is uncorrelated with $\{v_{t-1}, v_{t-2}, \ldots\}$, in analogy to the forward model.

Given data $\{x_1, \ldots, x_n\}$, truncate $v_n^n = E(v_n \,|\, x_1, \ldots, x_n)$ to zero and then iterate backward. That is, put $\widetilde{v}_n^n = 0$, as an initial approximation, and then generate the errors backward

$$\widetilde{v}_t^n = x_t - \phi x_{t+1} - \theta \widetilde{v}_{t+1}^n, \quad t = (n-1), (n-2), \ldots, 1.$$

Then,

$$\widetilde{x}_0^n = \phi x_1 + \theta \widetilde{v}_1^n + \widetilde{v}_0^n = \phi x_1 + \theta \widetilde{v}_1^n,$$

because $\widetilde{v}_t^n = 0$ for $t \leq 0$. Continuing, the general truncated backcasts are given by

$$\widetilde{x}_{1-m}^n = \phi \widetilde{x}_{2-m}^n, \quad m = 2, 3, \ldots \ .$$

[4] In the stationary Gaussian case, (a) the distribution of $\{x_{n+1}, x_n, \ldots, x_1\}$ is the same as (b) the distribution of $\{x_0, x_1 \ldots, x_n\}$. In forecasting we use (a) to obtain $E(x_{n+1} \,|\, x_n, \ldots, x_1)$; in backcasting we use (b) to obtain $E(x_0 \,|\, x_1, \ldots, x_n)$. Because (a) and (b) are the same, the two problems are equivalent.

3.6 Estimation

Throughout this section, we assume we have n observations, $x_1, \ldots, x_n$, from a causal and invertible Gaussian ARMA(p, q) process in which, initially, the order parameters, p and q, are known. Our goal is to estimate the parameters, $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$, and σ_w^2. We will discuss the problem of determining p and q later in this section.

We begin with method of moments estimators. The idea behind these estimators is that of equating population moments to sample moments and then solving for the parameters in terms of the sample moments. We immediately see that, if $E(x_t) = \mu$, then the method of moments estimator of μ is the sample average, $\bar{x}$. Thus, while discussing method of moments, we will assume $\mu = 0$. Although the method of moments can produce good estimators, they can sometimes lead to suboptimal estimators. We first consider the case in which the method leads to optimal (efficient) estimators, that is, AR(p) models.

When the process is AR(p),

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t,$$

the first $p + 1$ equations of (3.47) and (3.48) lead to the following:

Definition 3.10 *The* **Yule–Walker equations** *are given by*

$$\gamma(h) = \phi_1\gamma(h-1) + \cdots + \phi_p\gamma(h-p), \quad h = 1, 2, \ldots, p, \tag{3.98}$$

$$\sigma_w^2 = \gamma(0) - \phi_1\gamma(1) - \cdots - \phi_p\gamma(p). \tag{3.99}$$

In matrix notation, the Yule–Walker equations are

$$\Gamma_p\boldsymbol{\phi} = \boldsymbol{\gamma}_p, \quad \sigma_w^2 = \gamma(0) - \boldsymbol{\phi}'\boldsymbol{\gamma}_p, \tag{3.100}$$

where $\Gamma_p = \{\gamma(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ is a $p \times 1$ vector, and $\boldsymbol{\gamma}_p = (\gamma(1), \ldots, \gamma(p))'$ is a $p \times 1$ vector. Using the method of moments, we replace $\gamma(h)$ in (3.100) by $\widehat{\gamma}(h)$ [see equation (1.34)] and solve

$$\widehat{\boldsymbol{\phi}} = \widehat{\Gamma}_p^{-1}\widehat{\boldsymbol{\gamma}}_p, \quad \widehat{\sigma}_w^2 = \widehat{\gamma}(0) - \widehat{\boldsymbol{\gamma}}_p'\widehat{\Gamma}_p^{-1}\widehat{\boldsymbol{\gamma}}_p. \tag{3.101}$$

These estimators are typically called the Yule–Walker estimators. For calculation purposes, it is sometimes more convenient to work with the sample ACF. By factoring $\widehat{\gamma}(0)$ in (3.101), we can write the Yule–Walker estimates as

$$\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{R}}_p^{-1}\widehat{\boldsymbol{\rho}}_p, \quad \widehat{\sigma}_w^2 = \widehat{\gamma}(0)\left[1 - \widehat{\boldsymbol{\rho}}_p'\widehat{\boldsymbol{R}}_p^{-1}\widehat{\boldsymbol{\rho}}_p\right], \tag{3.102}$$

where $\widehat{R}_p = \{\widehat{\rho}(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix and $\widehat{\boldsymbol{\rho}}_p = (\widehat{\rho}(1), \ldots, \widehat{\rho}(p))'$ is a $p \times 1$ vector.

For AR(p) models, if the sample size is large, the Yule–Walker estimators are approximately normally distributed, and $\widehat{\sigma}_w^2$ is close to the true value of σ_w^2. We state these results in Property 3.8; for details, see Appendix B, §B.3.