Time Series (MATH5845)

Dr. Atefeh Zamani
Based on the notes by Prof. William T.M. Dunsmuir

T2 2025

Chapter 5

Prediction of Weakly Stationary Time Series

Contents

5.1 Generalities

Note 5.1
- *For development of this material we do not always require that the series be stationary but we do require it to have finite variance.*
- *Since we can always mean adjust our series prior to using the following methods and algorithms it is no loss of generality to assume that the series has zero mean function.*

Let $\{X_t\}$ be a time series with zero mean function and covariance function

$$\gamma(i,j) = E(X_i X_j).$$

The aim is to forecast a future value of a time series given previous observed values. We will develop this topic only for the situation where we forecast X_{n+1} given observations on $X_1^n = (X_n, \ldots, X_1)'$. For longer lead times (X_{n+l}, $l > 1$) similar ideas can be developed using the understanding gained from the case $l = 1$.

Theorem 5.1 ***The optimal forecast****, in the sense of* **mimimising mean squared prediction error**, *is the conditional expectation*

$$\tilde{X}_{n+1} = E(X_{n+1}|X_1^n).$$

Note that for non-Gaussian processes this is difficult to compute.

Proof. Let $g(X_1^n)$ be any function of the past of the time series which is going to be used for prediction:

$$\begin{aligned} E\left[X_{n+1} - g(X_1^n)\right]^2 &= E\left[X_{n+1} - E(X_{n+1}|X_1^n) + E(X_{n+1}|X_1^n) - g(X_1^n)\right]^2 \\ &= E([X_{n+1} - E(X_{n+1}|X_1^n)]^2) + E([E(X_{n+1}|X_1^n) - g(X_1^n)]^2) \\ &\quad + 2E([X_{n+1} - E(X_{n+1}|X_1^n)]\,[E(X_{n+1}|X_1^n) - g(X_1^n)]) \end{aligned}$$

Note that

$$\begin{aligned} &E([X_{n+1} - E(X_{n+1}|X_1^n)]\,[E(X_{n+1}|X_1^n) - g(X_1^n)]) \\ &= E\{E([X_{n+1} - E(X_{n+1}|X_1^n)]\,[E(X_{n+1}|X_1^n) - g(X_1^n)]\,|X_1^n)\} \\ &= E\{[E(X_{n+1}|X_1^n) - g(X_1^n)]\,E([X_{n+1} - E(X_{n+1}|X_1^n)]\,|X_1^n)\} \\ &= E\{[E(X_{n+1}|X_1^n) - g(X_1^n)] \times 0\} = 0 \end{aligned}$$

so that

$$E\left[X_{n+1} - g(X_1^n)\right]^2 = E([X_{n+1} - E(X_{n+1}|X_1^n)]^2) + E([E(X_{n+1}|X_1^n) - g(X_1^n)]^2)$$

in which the first term does not change with choice of function $g(\cdot)$ and the second term is minimised when the r.v in the expectation is 0 almost surely, that is when

$$E(X_{n+1}|X_1^n) = g(X_1^n)$$

proving the result. ■

5.1.1 Linear Prediction

One possible prediction is the minimum mean squared error forecast ("best") based on *linear* functions of the past. These linear predictions depend only on the second-order moments of the process, which are easy to estimate from the data. In general, the best linear predictor, BLP in abbreviation, of X_{n+1} given X_1^n is in the form

$$\hat{X}_{n+1} = \sum_{j=1}^{n} a_{nj} X_{n+1-j}$$

where the constants a_{nj} are found by minimising

$$S(\mathbf{a}_n) = E(X_{n+1} - \sum_{j=1}^{n} a_{nj} X_{n+1-j})^2.$$

Since $S(\mathbf{a}_n)$ is a quadratic function in $\mathbf{a}_n = (a_{n1}, \ldots, a_{nn})^T$ bounded below by zero, there is at least one value of $\mathbf{a}_n$ which minimises it and the minimum satisfies

$$\frac{\partial S(\mathbf{a}_n)}{\partial a_{nk}} = -2E[(X_{n+1} - \sum_{j=1}^{n} a_{nj} X_{n+1-j}) X_{n+1-k}] = 0, \quad k = 1, \ldots, n. \tag{5.1}$$

That is, the a_{nj} satisfy the system of equations for $k = 1, \ldots, n$

$$\begin{aligned} 0 &= E(X_{n+1} X_{n+1-k}) - \sum_{j=1}^{n} a_{nj} E(X_{n+1-j} X_{n+1-k}) \\ &= \gamma(n+1-k, n+1) - \sum_{j=1}^{n} a_{nj} \gamma(n+1-k, n+1-j). \end{aligned} \tag{5.2}$$

This can be written in matrix form

$$\Gamma_n \mathbf{a}_n = \gamma_n$$

by putting

$$\Gamma_n = \begin{bmatrix} \gamma(n,n) & \cdots & \gamma(n,1) \\ \vdots & \ddots & \vdots \\ \gamma(1,n) & \cdots & \gamma(1,1) \end{bmatrix}, \ \gamma_n = \begin{bmatrix} \gamma(n,n+1) \\ \vdots \\ \gamma(1,n+1) \end{bmatrix}.$$

Thus is has been shown that, for each n, the coefficients $\mathbf{a}_n = (a_{n1}, \ldots, a_{nn})^T$ that give the best linear predictor, $a^T X_1^n$ can be found by the solution

$$\mathbf{a}_n = \Gamma_n^{-1} \gamma_n \tag{5.3}$$

where Γ_n and γ_n are defined above in terms of the general covariance function $\gamma(i,j)$.

Note 5.2 *The matrix Γ_n is nonnegative definite.*

- *If Γ_n is singular, there are many solutions to* (5.4)*, but $\hat{X}_{n+1}$ is unique.*
- *If Γ_n is nonsingular, the elements of $\mathbf{a}_n$ are unique, and are given by*

$$\mathbf{a}_n = \Gamma_n^{-1} \gamma_n, \tag{5.4}$$

For ARMA models, the fact that $\sigma_Z^2 > 0$ *and* $\gamma(h) \to 0$ *as* $h \to \infty$ *is enough to ensure that* Γ_n *is positive definite.*

The **mean square one-step-ahead *linear* prediction error** is

$$E(X_{n+1} - \hat{X}_{n+1})^2 = \gamma(n+1, n+1) - \gamma_n' \Gamma_n \gamma_n \tag{5.5}$$

Note 5.3 *If a process is* ***Gaussian****,* ***minimum mean square error predictors and best linear predictors (BLP) are the same****. Specifically, let*

$$\Gamma_n = cov(X_1^n) = [\gamma(n+1-i, n+1-j)]_{i,j=1}^n, \quad \gamma_n = [\gamma(n+1-i, n+1)]_{i=1}^n$$

and note that

$$cov(X_1^{n+1}) = cov(\begin{bmatrix} X_{n+1} \\ X_1^n \end{bmatrix}) = \begin{bmatrix} \gamma(n+1, n+1) & \gamma_n' \\ \gamma_n & \Gamma_n \end{bmatrix}$$

and that for a zero mean Gaussian process (not necessarily stationary) we have the conditional distribution of X_{n+1} *given observations on* X_1^n *is*

$$X_{n+1} | X_1^n = x_1^n \sim N\left(\gamma_n' \Gamma_n^{-1} x_1^n, \gamma(n+1, n+1) - \gamma_n' \Gamma_n^{-1} \gamma_n\right).$$

Note that

$$\hat{X}_{n+1} = E(X_{n+1} | X_1^n = x_1^n) = \gamma_n' \Gamma_n^{-1} x_1^n$$

which is a linear function of past observations. Note also that the unconditional variance, $\gamma(n+1, n+1)$*, of* X_{n+1} *is reduced in the prediction conditional on the past.*

Example 5.1 (Prediction for an AR(2)) *Suppose we have a causal AR(2) process* $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$*, and one observation* $X_1 = x_1$*. Then, using equation* (5.4)*, the one-step-ahead prediction of* X_2 *based on* X_1 *is*

$$\hat{X}_2 = a_{11} x_1 = \frac{\gamma(1)}{\gamma(0)} x_1 = \rho(1) x_1.$$

Now, suppose we want the one-step-ahead prediction of X_3 *based on two observations* X_1 *and* X_2*; i.e.,* $\hat{X}_3 = a_{21} x_2 + a_{22} x_1$*. We could use* (5.2)

$$\begin{aligned} a_{21}\gamma(0) + a_{22}\gamma(1) &= \gamma(1) \\ a_{21}\gamma(1) + a_{22}\gamma(0) &= \gamma(2) \end{aligned}$$

to solve for a_{21} *and* a_{22}*, or use the matrix form in* (5.4) *and solve*

$$\begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} = \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix}^{-1} \begin{pmatrix} \gamma(1) \\ \gamma(2) \end{pmatrix},$$

On the other hand, it should be apparent from the model that $\hat{X}_3 = \phi_1 X_2 + \phi_2 X_1$*. Because* $\phi_1 X_2 + \phi_2 X_1$ *satisfies the prediction equations* (5.2)*,*

$$\begin{aligned} E\{[X_3 - (\phi_1 X_2 + \phi_2 X_1)] X_1\} &= E(Z_3 X_1) = 0, \\ E\{[X_3 - (\phi_1 X_2 + \phi_2 X_1)] X_2\} &= E(Z_3 X_2) = 0, \end{aligned}$$

by the uniqueness of the coefficients in this case, we conclude that that $a_{21} = \phi_1$ *and* $a_{22} = \phi_2$*. Continuing in this way, it is easy to verify that, for* $n \geq 2$*,*

$$\hat{X}_{n+1} = \phi_1 X_n + \phi_2 X_{n-1}.$$

That is, $a_{n1} = \phi_1$*,* $a_{n2} = \phi_2$*, and* $a_{nj} = 0$*, for* $j = 3, 4, \ldots, n$*.*

Note 5.4 *It can be shown that if the time series is a causal AR(p) process, then, for* $n \geq p$*,*

$$\hat{X}_{n+1} = \phi_1 X_n + \phi_2 X_{n-1} + \ldots + \phi_p X_{n-p+1}, \tag{5.6}$$

5.1.2 The Durbin-Levinson Algorithm and Partial Autocorrelations

This algorithm was invented independently by Levinson (1947) and Durbin (1960) and is used to determine the linear prediction coefficients in a stationary times series (using true autocorrelations) as well as to get the Yule-Walker estimates (using estimated autocorrelations) recursively for a stationary time series using a series of increasing order AR(p) processes.

When the time series $\{X_t\}$ is weakly stationary so that $\gamma(i,j) = \gamma(i-j)$ we have

$$\Gamma_n = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \ddots & \vdots \\ & & \ddots & \gamma(1) \\ \gamma(n-1) & \cdots & \gamma(1) & \gamma(0) \end{bmatrix}, \quad \gamma_n = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(n-1) \end{bmatrix}.$$

and in this case we denote the solution as

$$\mathbf{a}_n = \Gamma_n^{-1}\gamma_n \tag{5.7}$$

which is the solution to the Yule-Walker equations introduced previously. This solution can also be written in terms of autocorrations $\rho(h)$ by dividing the autocovariances, $\gamma(h)$, by $\gamma(0)$. Note that in this case

$$E(X_{n+1} - \hat{X}_{n+1})^2 = \gamma(0) - \gamma_n'\Gamma_n\gamma_n \tag{5.8}$$

The system of Yule-Walker equations can be solved recursively, in n, by the Durbin-Levinson recursion.

Algorithm 1 (The Durbin-Levinson Algorithm) *Equations* (5.7) *and* (5.8) *can be solved iteratively as follows:*

$$a_{00} = 0, \qquad \nu_0 = \gamma(0). \tag{5.9}$$

For $n \geq 1$,

$$a_{nn} = \frac{\rho(n) - \sum_{k=1}^{n-1} a_{n-1,k}\rho(n-k)}{1 - \sum_{k=1}^{n-1} a_{n-1,k}\rho(k)}, \quad \nu_n = \nu_{n-1}(1 - a_{nn}^2,) \tag{5.10}$$

where for $n \geq 2$,

$$a_{nk} = a_{n-1,k} - a_{nn}a_{n-1,n-k}, \qquad k = 1, 2, \ldots, n-1. \tag{5.11}$$

A proof of this result is given in Brockwell and Davis [1991], P. 170. Note that the mean squared error in prediction reduces from stage $n-1$ to stage n by an amount $[1 - \phi_{nn}^2]$ defined in terms of the partial autocorrelation ϕ_{nn}^2. Hence for an autoregressive process of order p the mean squared error of prediction will not be reduced by using historical data further in the past than p time points.

Example 5.2 (Using the Durbin–Levinson Algorithm) *To use the algorithm, start with* $a_{00} = 0$ *and* $\nu_0 = \gamma(0)$. *Then, for* $n = 1$,

$$a_{11} = \rho(1), \quad \nu_1 = \gamma(0)[1 - a_{11}^2].$$

For $n = 2$,

$$a_{22} = \frac{\rho(2) - a_{11}\rho(1)}{1 - a_{11}\rho(1)}, \quad a_{21} = a_{11} - a_{22}a_{11}$$
$$\nu_2 = \nu_1[1 - a_{22}^2] = \gamma(0)[1 - a_{11}^2][1 - a_{22}^2].$$

For $n = 3$,

$$a_{33} = \frac{\rho(3) - a_{21}\rho(2) - a_{22}\rho(1)}{1 - a_{21}\rho(1)} - a_{22}\rho(2),$$
$$a_{32} = a_{22} - a_{33}a_{21}, \quad a_{31} = a_{21} - a_{33}a_{22}$$
$$\nu_3 = \nu_2[1 - a_{33}^2] = \gamma(0)[1 - a_{11}^2][1 - a_{22}^2][1 - a_{33}^2].$$

and so on. Note that, in general, the standard error of the one-step-ahead forecast is the square root of

$$\nu_n = \gamma(0)\prod_{j=1}^{n}[1 - a_{jj}^2] \tag{5.12}$$

Example 5.3 (The PACF of an AR(2)) *Recall that for AR(2),* $\rho(h) - \phi_1\rho(h-1) - \phi_2\rho(h-2) = 0$ *for* $h \geq 1$. *When* $h = 1, 2, 3$, *we have* $\rho(1) = \phi_1/(1 - \phi_2)$, $\rho(2) = \phi_1\rho(1) + \phi_2$, $\rho(3) - \phi_1\rho(2) - \phi_2\rho(1) = 0$. *Thus,*

$$a_{11} = \rho(1) = \frac{\phi_1}{1 - \phi_2}$$
$$a_{22} = \frac{\rho(2) - \rho^2(1)}{1 - \rho^2(1)} = \frac{\left[\phi_1\left(\frac{\phi_1}{1-\phi_2}\right) + \phi_2\right] - \left(\frac{\phi_1}{1-\phi_2}\right)^2}{1 - \left(\frac{\phi_1}{1-\phi_2}\right)^2} = \phi_2$$
$$a_{21} = \rho(1)[1 - \phi_2] = \phi_1$$
$$a_{33} = \frac{\rho(3) - \phi_1\rho(2) - \phi_2\rho(1)}{1 - \phi_1\rho(1) - \phi_2\rho(2)} = 0.$$

Notice that $a_{22} = \phi_2$ *for an AR(2) model.*

5.2 The innovations representation.

For each n, let a_{nj} continue to denote the values that solve these equations and therefore give the best linear predictors. We write the prediction error for each observation ("innovation") as

$$\begin{aligned} U_{t+1} &= X_{t+1} - \hat{X}_{t+1} \\ &= X_{t+1} - \sum_{j=1}^{t} a_{tj} X_{t+1-j}, \quad t = 0, \ldots, n. \end{aligned}$$

with the convention that, for $t = 0$, the summation is empty and therefore the best linear predictor of X_1 is $\hat{X}_1 = 0$, the unconditional mean of the time series. Because of equation (5.1) it follows that $U_{n+1} = X_{n+1} - \sum_{j=1}^{n} a_{nj} X_{n+1-j}$ is uncorrelated with $H_n = (X_1, \ldots, X_n)$ and hence with any linear combination of elements in H_n. In particular previous prediction errors, U_{t+1} for $t = 0, \ldots, n-1$ are linear combinations of elements in H_n and hence U_{n+1} is uncorrelated with the previous linear prediction errors $U_1, \ldots, U_n$. Since this is true for each $n \geq 1$ it follows that the series of linear prediction errors are mutually uncorrelated.

Let

$$\gamma(i, j) = E(X_i X_j)$$

and define

$$A_n = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ -a_{11} & 1 & 0 & & & \vdots \\ -a_{22} & -a_{21} & 1 & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ & & & & 1 & 0 \\ -a_{n-1,n-1} & & & & -a_{n-1,1} & 1 \end{bmatrix}$$

where the $a_{n,j}$ are the solutions generated from equation (5.1). Then

$$\mathbf{U}_n = A_n \mathbf{X}_n$$

where $\mathbf{X}_n = (X_1, \ldots, X_n)'$. Note that $\det(A_n) = 1$ so that $C_n = A_n^{-1}$ exists and is also lower triangular. Now, since $\mathbf{U}_n = \mathbf{X}_n - \hat{\mathbf{X}}_n$, we have

$$\begin{aligned} \hat{\mathbf{X}}_n &= \mathbf{X}_n - \mathbf{U}_n \\ &= A_n^{-1} \mathbf{U}_n - \mathbf{U}_n \\ &= (A_n^{-1} - I) \mathbf{U}_n \\ &= \Theta_n \mathbf{U}_n \end{aligned}$$

where

$$\Theta_n = (C_n - I) = \begin{bmatrix} 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \theta_{11} & 0 & 0 & & & 0 \\ \theta_{22} & \theta_{21} & 0 & 0 & & 0 \\ \vdots & & & \ddots & & \\ & & & & 0 & 0 \\ \theta_{n-1,n-1} & & & & \theta_{n-1,1} & 0 \end{bmatrix}.$$

Hence

$$\hat{\mathbf{X}}_n = \Theta_n \mathbf{U}_n = \Theta_n (\mathbf{X}_n - \hat{\mathbf{X}}_n)$$

and, because Θ_n is lower triangular, recursive calculation of 1-step ahead prediction can be obtained as follows,

$$\hat{X}_{n+1} = \begin{cases} 0, & n = 0 \\ \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) & n = 1, 2, \ldots \end{cases} \tag{5.13}$$

in terms of previous 1-step ahead prediction errors. The prediction mean squared error (MSE) is denoted

$$v_n = E(X_{n+1} - \hat{X}_{n+1})^2. \tag{5.14}$$

The innovations algorithm of Brockwell and Davis [2002], Page 62, provides a recursive (and fast) method of generating the coefficients $\theta_{n,j}$ and the one-step ahead linear prediction mean squared errors (variances) v_n.

Algorithm 2 (The Innovations Algorithm) *The best linear predictor of X_{n+1} is obtained by equation (5.13) where*

- $v_0 = \gamma(1,1)$,
- $\theta_{n,n-k} = v_k^{-1}(\gamma(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j, \quad 0 \le k < n$
- $v_n = \gamma(n+1,n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j$ *is the prediction MSE*

Remark 1 *1. Stationarity is not required in the innovations algorithm.*

2. The innovations algorithm can be used to express X_{n+1} as

$$\begin{aligned} X_{n+1} &= (X_{n+1} - \hat{X}_{n+1}) + \hat{X}_{n+1} \\ &= \sum_{j=0}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) \end{aligned}$$

where $\theta_{n0} = 1$. This expresses X_{n+1} as a linear combination of 1-step ahead prediction errors (innovations).

3. If $\{X_t\}$ is an ARMA(p,q) process the innovations algorithm simplifies to the following, Brockwell and Davis [2002], Page 88:

$$\hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & 1 \le n < m = \max(p,q) \\ \phi_1 X_n + \cdots \phi_p X_{n+1-p} + \sum_{j=1}^{q} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), & n \ge m \end{cases}$$

Example 5.4 *AR(p) time series: put $q = 1$ and $\theta_1 = 0$ to get*

$$\hat{X}_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n+1-p}, \quad n \ge p.$$

Example 5.5 *ARMA$(1,1)$ time series. To simplify notation let $r_n = v_n/\sigma^2$. Then*

$$\begin{aligned} r_0 &= (1 + 2\theta\phi + \theta^2)/(1 - \phi^2) = \frac{\mathrm{var}(X_1)}{\sigma^2} \\ \theta_{n,1} &= \theta/r_{n-1} \to \theta, \quad n \to \infty \\ r_n &= 1 + \theta^2 - \theta^2/r_{n-1} \to 1, \quad n \to \infty \end{aligned}$$

and

$$\hat{X}_{n+1} = \phi_1 X_n + \theta_{n1}(X_n - \hat{X}_n), \quad n \ge 1.$$

Example 5.6 *MA(1) time series.*

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}), \quad 1 \leq n$$

where

$$\begin{aligned}
\theta_{n,j} &= 0, \quad 2 \leq j \leq n \\
\theta_{n1} &= \theta\sigma^2/v_n = \theta/r_n \to \theta, \ n \to \infty \\
v_0 &= (1+\theta^2)\sigma^2 \\
v_n &= [1+\theta^2 - \theta^2\sigma^2/v_{n-1}]\sigma^2 \to \sigma^2, \ n \to \infty
\end{aligned}$$

5.2.1 General theory of forecasting m steps ahead for ARMA(p,q) processes

The following notes are based primarily on Shumway et al. [2000].

Let $\{X_t\}$ be a stationary, causal and invertible ARMA(p,q) process $\phi(B)X_t = \theta(B)Z_t$ where $Z_t \overset{\text{i.i.d}}{\sim} (0,\sigma^2)$ process. Note that the conditional expectation is always the minimum mean square predictor and under the assumptions on the noise process $\{Z_t\}$ this is a linear function of $(X_n,\ldots,X_1)$.

- **Prediction based on finite history:** Let

  $$X_{n+m}^n = E(X_{n+m}|X_n,\ldots,X_1)$$

  be the minimum mean squared predictor m steps ahead from forecast horizon n using values of the process from times $n, n-1,\ldots,1$.

- **Prediction based on infinite history:** Let

  $$\tilde{X}_{n+m} = E(X_{n+m}|X_n,\ldots,X_0,X_{-1}\ldots,)$$

  be the minimum mean squared predictor m steps ahead from forecast horizon n using the infinite past of the process.

Note 5.5 *$\tilde{X}_{n+m}$ is an idealization that cannot be used in practice. But for large sample sizes n, X_{n+m}^n will be a good approximation to $\tilde{X}_{n+m}$.*

Lets start with writing X_{n+m} in its causal and invertible form. Using the MA(∞) representation in (4.23) we can write

$$X_{n+m} = \sum_{j=0}^{\infty} \psi_j Z_{n+m-j} \tag{5.15}$$

Using the AR(∞) representation in (4.24) we get:

$$Z_{n+m} = \sum_{j=0}^{\infty} \pi_j X_{n+m-j} \tag{5.16}$$

Running conditional expectations through both sides of (5.15) gives

$$\tilde{X}_{n+m} = \sum_{j=0}^{\infty} \psi_j \tilde{Z}_{n+m-j} = \sum_{j=m}^{\infty} \psi_j Z_{n+m-j}. \tag{5.17}$$

The last equality in (5.17) follows from two facts:

- If $t > n$ then Z_t is in the future relative to the conditioning set and by the i.i.d properties $E(Z_t|X_n, \ldots, X_0, X_{-1}, \ldots) = E(Z_t) = 0$.
- If $t \leq n$ then, using (4.24), Z_t is completely determined by the conditioning set and hence $E(Z_t|X_n, \ldots, X_0, X_{-1}, \ldots) = Z_t$.

Recursive form of prediction: In principle, the Z_t needed to create forecasts from (5.17) could be calculated using the infinite autoregressive representation (4.24). There is an alternative recursive method derived by taking expectations, conditional on the complete past, on both sides of (5.16) to get

$$0 = \tilde{X}_{n+m} + \sum_{j=1}^{\infty} \pi_j \tilde{X}_{n+m-j}.$$

Re-arranging this and using $E(X_t|X_n, \ldots, X_0, X_{-1}, \ldots) = X_t$ when $t \leq n$ gives

$$\tilde{X}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{X}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j X_{n+m-j} \tag{5.18}$$

Starting with the 1-step ahead predictor, $m = 1$, (5.18) can be used recursively for $m > 1$ step ahead predictors.

Mean Squared Error for m- step ahead prediction: Subtracting (5.17) from (4.24) gives

$$X_{n+m} - \tilde{X}_{n+m} = \sum_{j=0}^{m-1} \psi_j Z_{n+m-j} \tag{5.19}$$

and hence the mean squared error of prediction is

$$E(X_{n+m} - \tilde{X}_{n+m})^2 = \sigma^2 \sum_{j=0}^{m-1} \psi_j^2. \tag{5.20}$$

and for $k \geq 1$ covariances between forecasts at different lead times are

$$E(X_{n+m} - \tilde{X}_{n+m})(X_{n+m+k} - \tilde{X}_{n+m+k}) = \sigma^2 \sum_{j=0}^{m-1} \psi_j \psi_{j+k}. \tag{5.21}$$

.

These results are intuitively reasonable:

- It is obvious from (5.20) forecast variability increases with lead time away from the time at which forecasts are calculated. In the limit, as $m \to \infty$, the mean squared error of forecasts in (5.20) tends to the process variance:

$$\lim_{m\to\infty} E(X_{n+m} - \tilde{X}_{n+m})^2 = \lim_{m\to\infty} \sigma^2 \sum_{j=0}^{m-1} \psi_j^2 = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 = \gamma_X(0)$$

  This convergence is assured because $\psi_j \searrow 0$ exponentially fast as $j \to \infty$.

- Likewise from (5.21) the prediction errors at different lead times are correlated.

Including a non-zero mean term: All the above holds if X_t is replaced by $X_t - \mu$ where $\mu = E(X_t)$. Then the equivalent of (5.17) is

$$\tilde{X}_{n+m} = \mu + \sum_{j=m}^{\infty} \psi_j Z_{n+m-j} \tag{5.22}$$

and, since $\psi_j \searrow 0$ exponentially fast as $j \to \infty$, $\sum_{j=m}^{\infty} \psi_j Z_{n+m-j} \to 0$ almost surely and hence $\tilde{X}_{n+m} \to \mu$ for $m \to \infty$. Hence the long range forecast of an ARMA process is the stationary mean μ.

Prediction using finite information: In practice only $X_n, X_{n-1}, \ldots, X_1$ is observable for creating predictors.

- When n is small or moderate the general methods of finding the linear prediction coefficients of Section 5.1 can be used.
- For any n the innovations algorithm (or Kalman filter, not discussed in this course) can be used to produce forecasts.
- Alternatively, for reasonably large n compared to the decay rate of the AR(∞) π_j weights and MA(∞) ψ_j weights, an approximate recursion is useful.

For the last instance, the second sum in (5.18) is truncated as

$$\tilde{X}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{X}_{n+m-j} - \sum_{j=m}^{n+m-1} \pi_j X_{n+m-j} \tag{5.23}$$

assuming that the π_j weights are sufficiently small so as to make $\sum_{j=n+m}^{\infty} \pi_j X_{n+m-j}$ small.

We continue to use (5.20) to assess the variability of this forecast.

Truncated Prediction in ARMA models: These are obtained by intializing the "pre-period" ($t \leq 0$) values of the series and prediction errors to zero. Let $\tilde{X}_t^n = 0$ for $t \leq 0$ and $\tilde{Z}_t^n = 0$ for $t \leq 0$ or $t > n$. Also denote (obviously) $\tilde{X}_t^n = X_t$ for $1 \leq t \leq n$. Define truncated prediction errors

$$\tilde{Z}_t^n = \phi(B)\tilde{X}_t^n - \theta_1 \tilde{Z}_{t-1}^n - \cdots - \theta_q \tilde{Z}_{t-q}^n, \quad 1 \leq t \leq n.$$

Then, using the ARMA defining equation, form the truncated predictors as

$$\tilde{X}_{n+m}^n = \phi_1 \tilde{X}_{n+m-1}^n + \cdots + \phi_p \tilde{X}_{n+m-p}^n + \theta_1 \tilde{Z}_{n+m-1}^n + \cdots + \theta_q \tilde{Z}_{n+m-q}^n \tag{5.24}$$

for $m = 1, 2, \ldots$.

Prediction intervals: $(1-\alpha)100\%$ prediction intervals are

$$\tilde{X}_{n+m}^n \pm c_{\alpha/2} \sqrt{E(X_{n+m} - \tilde{X}_{n+m})^2} \tag{5.25}$$

where $c_{\alpha/2}$ is selected to give the required coverage in the assumed distribution for the Z_t. For Gaussian processes $c_{\alpha/2} = 1.96$ for 95% coverage.

Note that these are individual lead time (m) prediction intervals and, when asembled for several lead times, do not give nominal coverage $(1-\alpha)100\%$ for the future sample path over the lead time. Bonferroni adjustments could be used but are typically not in practice. Bootstrap ideas can also be applied here.

- For non-Gaussian data, it is tricky to get correct coverage and bootstrap ideas are used.
- Typically nominal 95% intervals (around plus or minus two standard deviations of prediction error) are "frighteningly too large" for many consumers of forecasts and sometimes people present 68% intervals using one standard deviation around the prediction.

5.3 Exercises

Exercise 5.1 *Prove* (5.5).

Exercise 5.2 *For the ARMA(1, 1) model show, using* (5.24) *that the truncated recursive predictors are obtained as follows:*

$$\tilde{X}^n_{n+1} = \phi X_n + \theta \tilde{Z}^n_n, \quad m = 1, \quad \textit{and} \quad \tilde{X}^n_{n+m} = \phi \tilde{X}^n_{n+m-1}, \quad m > 1$$

and truncated forecast errors, initiated with $\tilde{Z}^n_0 = 0$, $X_0 = 0$, *are*

$$\tilde{Z}^n_t = X_t - \phi X_{t-1} - \theta \tilde{Z}^n_{t-1}. \quad t = 1, \ldots, n.$$

Exercise 5.3 *Simplify the general formulae in Exercise 5.2 for the special cases* $p = 1, q = 0$ *and* $p = 0, q = 1$.

- *Discuss whether the forecasts and forecast mean squared error seem intuitively reasonable.*
- *Sketch the forecasts and prediction intervals as lead time* m *increases in these cases.*

Exercise 5.4 *a) Let* $\{X_t\}$ *be a Gaussian AR(p) process with zero mean. Using the fact that the best linear predictor* $\hat{X}_{n+1|n,\ldots,1} = \sum_{j=1}^n \phi_{nj} X_{n+1-j}$ *of* X_{n+1} *given* $X_n, \ldots, X_1$ *is equal to the conditional expectation* $E(X_{n+1}|X_n, \ldots, X_1)$ *for Gaussian processes, show that when* $n \geq p$,

$$\phi_{nj} = \begin{cases} \phi_j, \ j = 1, \ldots, p \\ 0, \ j > p \end{cases}$$

and hence, by definition of partial autocorrelations $\alpha(h)$ *that* $\phi_{hh} = 0$ *if* $p < h$.

b) Interpretation of partial autocorrelations. Let $\hat{X}_{n+1|n,\ldots,2}$ *be the best linear predictor of* X_{n+1} *given* $X_n, \ldots, X_2$ *and* $\hat{X}_{1|2,\ldots,n}$ *be the best linear predictor of* X_1 *given* $X_2, \ldots, X_n$. *By following these suggested steps (or otherwise) show that* $\phi_{nn} = corr(X_1 - \hat{X}_{1|2,\ldots,n}, X_{n+1} - \hat{X}_{n+1|n,\ldots,2})$*:*

- *Refer to the Durbin Levinson recursion as stated in the notes. Left multiply both sides by* $(X_n, \ldots, X_2)$ *and add* $\phi_{nn} X_1$ *to both sides to get*

$$\hat{X}_{n+1|n,\ldots,1} = \hat{X}_{n+1|n,\ldots,2} + \phi_{nn}(X_1 - \hat{X}_{1|2,\ldots n}).$$

- *Subtract both sides from* X_{n+1} *to get*

$$X_{n+1} - \hat{X}_{n+1|n,\ldots,1} = (X_{n+1} - \hat{X}_{n+1|n,\ldots,2}) - \phi_{nn}(X_1 - \hat{X}_{1|2,\ldots n}).$$

- *Multiply both sides by* $X_1 - \hat{X}_{1|2,\ldots n}$ *and take expectations throughout to prove the result. Recall that* $X_1 - \hat{X}_{1|2,\ldots n}$ *is in the space spanned by* $X_1, \ldots, X_n$ *and hence* $X_{n+1} - \hat{X}_{n+1|n,\ldots,1}$ *is orthogonal to it.*

5.4 Tutorial: Week 4

The following two practical examples are based on Shumway and Stoffer (2017).

Global Warming

Consider the global temperature series, provided `gtemp_land` data set from the `astsa` package. The data are the global mean land temperature index from 1880 to 2023. In particular, the data are deviations, measured in degrees centigrade, from the 1951–1980 average. We note an apparent upward trend in the series during the latter part of the twentieth century that has been used as an argument for the global warming hypothesis. Note also the leveling off at about 1935 and then another rather sharp upward trend at about 1970.

(i) Plot the data along with its ACF and PACF and comment.

(ii) There is an upward trend in the data. Use differencing to detrend the data. Plot the deta along with ACF and PACF and comment.

(iii) Fit an ARMA(p,q) model to `gtemp_land` after differencing and check the residuals.

(iv) After deciding on an appropriate model, forecast (with limits) the next 10 years. Comment.

Sulfur dioxide levels from the LA pollution study

`so2` data set in the `astsa` package provides sulfur dioxide levels which is used in a study on the possible effects of temperature and pollution on weekly mortality in Los Angeles County. The data is collected from 1970 to 1980.

(i) Plot the data along with its ACF and PACF.

(ii) There seems to be a trend in the data. Use differencing to detrend the data. Plot the differenced series along with its ACF and PACF and comment.

(iii) Fit an ARMA(p, q) model to the series and perform all of the necessary diagnostics.

(iv) After deciding on an appropriate model, forecast the data into the future four time periods ahead (about one month) and calculate 95prediction intervals for each of the four forecasts. Comment.

(v) Take `log` from the original series and fit the model again. Compare the results.