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.*

$X_i - \bar{X} \longrightarrow$ mean adjust

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

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

Aim: 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*

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

this prediction might be non-linear

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

$$\min_{g(X_1 \ldots X_n)} \underbrace{E\left(X_{n+1} - \tilde{X}_{n+1}\right)^2}_{MSE} \Rightarrow g(X_1, \ldots X_n) = E(X_{n+1} | X_1, \ldots, X_n)$$

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

$\pm E(X_{n+1}|X_1^n)$

$$
\begin{aligned}
E\left[X_{n+1}-\underbrace{g(X_1^n)}_{?!}\right]^2 &= E\left[X_{n+1}-E(X_{n+1}|X_1^n)+E(X_{n+1}|X_1^n)-g(X_1^n)\right]^2 \\
&= E([X_{n+1}-E(X_{n+1}|X_1^n)]^2)+E([E(X_{n+1}|X_1^n)-g(X_1^n)]^2) \\
&\quad + \underbrace{2E([X_{n+1}-E(X_{n+1}|X_1^n)][E(X_{n+1}|X_1^n)-g(X_1^n)])}_{0}
\end{aligned}
$$

→ $(a+b)^2 = a^2+b^2+2ab$

*

Note that

$G(Y) = E\,E(Y|X)$ *

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

→ given information

* $E(X_{n+1}|X_1^n) - E\,\underline{E(X_{n+1}|X_1^n)}$

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

$= 0$

so that

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

is not a function of g

and it is non-negative ✓

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. ■

$$E\underbrace{\Big(E(X_{n+1} \mid X_1^n) - g(X_1^n)\Big)^2}_{\geq 0}$$

$$\Rightarrow \min E\Big(E(X_{n+1} \mid X_1^n) - g(X_1^n)\Big) = 0$$

it happens if $g(X_1^n) = E(X_{n+1} \mid X_1^n)$ with prob 1

$$\boxed{\min_a (X - a)^2 \Rightarrow X = a}$$

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

$$\text{prediction} \leftarrow \hat{X}_{n+1} = \sum_{j=1}^{n} \underbrace{a_{nj}}_{\rightarrow \text{ needs to be estimated}} X_{n+1-j} \longrightarrow \text{linear process}$$

where the constants a_{nj} are found by minimising

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

$\underset{a_n}{\text{minimize}}\ S(a_n) \Rightarrow$ find the values of a_{nj}

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

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

$$\frac{\partial S(a_n)}{\partial a_{nk}} = \frac{\partial E\left(X_{n+1} - \sum_{j=1}^{n} a_{nj} \underline{X_{n+1-j}}\right)^2}{\partial a_{nk}}.$$

$$= 2E\left(X_{n+1} - \sum a_{nj} X_{n+1-j}\right) X_{n+1-k} = 0$$

$$\Rightarrow E(X_{n+1} X_{n+1-k}) - \sum a_{nj} E(X_{n+1-j} X_{n+1-k}) = 0$$

Since the mean was zero

$$\Rightarrow \operatorname{cov}(X_{n+1}, X_{n+1-k}) = \sum_{j=1}^{n} a_{nj} \operatorname{cov}(X_{n+1-j}, X_{n+1-k})$$

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

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

This can be written in matrix form $\Gamma_n \mathbf{a}_n = \gamma_n$ by putting

stationarity

If invertible $\hookrightarrow \Gamma_n^{-1}$

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

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, $\mathbf{a}_n^T X_1^n$ can be found by the solution

$$\mathbf{a}_n = \Gamma_n^{-1}\gamma_n. \qquad (5.3)$$

Note 5.2 *The matrix Γ_n is nonnegative definite.* → no unique solution for an but $a_n^T X_1^n = \hat{X}_{n+1}$ is unique

- *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*

$$\mathbf{a}_n = \Gamma_n^{-1}\gamma_n, \qquad (5.4)$$

⇒ we have a unique prediction for X_{n+1}

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. → unique prediction

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

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

exercise 5.1

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 X_{n+1} and $X_1, \dots, X_n$ (past observation)

$$cov(X_1^{n+1}) = cov(\begin{bmatrix} X_{n+1} \\ X_1^n \end{bmatrix}) = \left[\begin{array}{c|c} \overset{\Sigma_{11}}{\gamma(n+1, n+1)} & \overset{\Sigma_{12}}{\gamma_n'} \\ \hline \underset{\Sigma_{21}}{\gamma_n} & \underset{\Sigma_{22}}{\Gamma_n} \end{array}\right] \quad \text{Block matrix}$$

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

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

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.

From chapter 1

$$\begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

$$X^{(1)} \mid X^{(2)} \sim N\left(\mu^{(1)} + \Sigma_{12} \Sigma_{22}^{-1} (X^{(2)} - \mu^{(2)}) , \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right)$$

conditional expectation

$$\Rightarrow E(X^{n+1} \mid X_1^n) = 0 + \gamma_n^T \Gamma_n^{-1} (X_1^n - 0)$$

$$= \underbrace{\gamma_n^T \Gamma_n^{-1}}_{a_n^T} X_1^n$$

$1-\Phi_1 z - \Phi_2 z^2 = 0$ → roots must be outside the unit circle (causality)

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*

$a_n = \Gamma_n^{-1} \gamma_n$

$$\Gamma_n = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & & & \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{pmatrix}$$

$n=1 \quad \Gamma_1 = \gamma(0)$

$\gamma_1 = \gamma(1) \quad \Longrightarrow a_{11} = \frac{\gamma(1)}{\gamma(0)}$

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

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)

$\Gamma_2 = \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix}$ $\quad \Gamma_n a_n = \gamma_n$ $\qquad \gamma_n = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(n) \end{pmatrix}$

$\gamma_2 = \begin{pmatrix} \gamma(1) \\ \gamma(2) \end{pmatrix} \quad a_n = \begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} \Longrightarrow$

$$a_{21}\gamma(0) + a_{22}\gamma(1) = \gamma(1) \checkmark$$
$$a_{21}\gamma(1) + a_{22}\gamma(0) = \gamma(2) \checkmark$$

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

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

Solution!

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),

in AR $Cov(Z_t, X_s) = 0$ $t > s$

past observation

$$E\{\overbrace{[X_3 - (\phi_1 X_2 + \phi_2 X_1)]}^{Z_3} X_1\} = E(Z_3 X_1) = 0,$$
$$E\{\underbrace{[X_3 - (\phi_1 X_2 + \phi_2 X_1)]}_{Z_3} X_2\} = E(Z_3 X_2) = 0,$$

past observation

we know that we have a unique solution for the coefficients in ARMA

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* $j = 3, 4, \ldots, n$.

AR(2)

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

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

one - step ahead prediction

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

Equation(5.5) →

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

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 $n \geq 1$,

error

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

where for $n \geq 2$,

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

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.

correct

Example 5.2 (Using the Durbin–Levinson Algorithm) *To use the algorithm, start with* $a_{00} = 0$ *and* $\nu_0 = \gamma(0)$. *Then, for* $n = 1$, PACF in AR $\leftarrow a_{11} = \frac{\rho(1)}{1} = \rho(1)$

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

$v_1 = v_0(1 - a_{11}^2)$

For $n = 2$,

$$a_{22} = \frac{\rho(2) - a_{11}\rho(1)}{1 - a_{11}\rho(1)}, \checkmark \quad a_{21} = a_{11} - a_{22}a_{11} \checkmark$$

$$\nu_2 = \nu_1[1 - a_{22}^2] = \gamma(0)[1 - a_{11}^2][1 - a_{22}^2]. \checkmark$$

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

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

$$a_{22} = \frac{\rho(2) - a_{11}\rho(1)}{1 - a_{11}\rho(1)}$$

Yule-walker equations

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,*

PACF lag 1

$$a_{11} = \rho(1) = \frac{\phi_1}{1-\phi_2} \quad ✓$$

PACF lag 2

$$a_{22} = \frac{\rho(2) - \rho^2(1)}{1-\rho^2(1)} = \frac{\left[\phi_1\left(\frac{\phi_1}{1-\phi_2}\right) + \phi_2\right] - \left(\frac{\phi_1}{1-\phi_2}\right)^2}{1 - \left(\frac{\phi_1}{1-\phi_2}\right)^2} = \phi_2$$

$$a_{21} = \rho(1)[1-\phi_2] = \phi_1$$

PACF lag 3

$$a_{33} = \frac{\rho(3) - \phi_1\rho(2) - \phi_2\rho(1)}{1-\phi_1\rho(1) - \phi_2\rho(2)} = 0.$$

$a_{31} = \phi_1$ $a_{32} = \phi_2$

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

$n=4$

$a_{44} = 0$ $a_{43} = 0$ $a_{42} = \phi_2$ $a_{41} = \phi_1$

$X_5 = \phi_1 X_4 + \phi_2 X_3$

$[X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cancel{\overset{0}{Z_t}}]$

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

error term

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

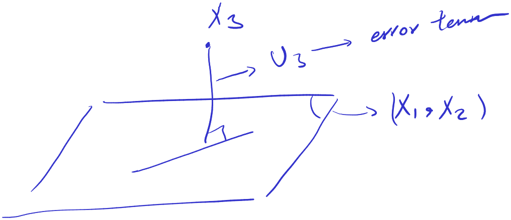
- For $t = 0$, the summation is empty and therefore the best linear predictor of X_1 is $\underline{\hat{X}_1 = 0}$, the unconditional mean of the time series.
- $U_{n+1} = X_{n+1} - \sum_{j=1}^{n} a_{nj} X_{n+1-j}$ is uncorrelated with $H_n = (X_1, \ldots, X_n)$ and hence with any linear combination 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.

$$U_1 = X_1 - \hat{X}_1 = X_1 - 0$$

$$U_2 = X_2 - \hat{X}_2 = X_2 - (a_{11} X_1)$$

$$U_3 = X_3 - \hat{X}_3 = X_3 - (a_{21} X_2 + a_{22} X_1)$$



$$\text{corr}\,(U_3, \underbrace{g(X_1, X_2)}_{\text{linear function}}) = 0$$

Let

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

and define

Lower triangular

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

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.

↳ product of elements on the main diagonal. ✓

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

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

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

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

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*

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

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

looks like a linear process

$Z_i \longrightarrow \tilde{Z}_i$

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:*

Correct

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

ARMA(2,3)

$\hat{z}_i$

$$\hat{X}_{n+1} = \begin{cases} \sum \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) & 1 \leq n < 3 \\ & \end{cases}$$

$n \geq 3$

$$\hat{X}_{n+1} = \phi_1 X_n + \phi_2 X_{n-1} + \cancel{Z_n} + \theta_1 \hat{Z}_{n-1} + \theta_2 \hat{Z}_{n-2} + \theta_3 \hat{Z}_{n-3}$$

$E(Z_t) = 0$

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

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

if $1 \leq n < p$ $\quad \hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj} \left(X_{n+1-j} - \hat{X}_{n+1-j} \right)$

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

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

and

$$\hat{X}_{n+1} = \phi_1 X_n + \theta_{n1}\underbrace{(X_n - \hat{X}_n)}_{\text{innovation term}}, \quad n \geq 1.$$

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.

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

  based on $X_1 \cdots X_n$

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

  observation $n+m$

  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

  infinite past

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

  observation $n+m$

  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_{n+m} *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 (4.20) we can write

causality

$$X_{n+m} = \sum_{j=0}^{\infty} \psi_j Z_{n+m-j} \qquad \psi_0 = 1 \tag{5.15}$$

Using the AR(∞) representation in (4.21) we get:

invertability

$$Z_{n+m} = \sum_{j=0}^{\infty} \pi_j X_{n+m-j} \qquad \pi_0 = 1 \qquad * \tag{5.16}$$

Running conditional expectations through both sides of (5.15) gives

$$\tilde{X}_{n+m} = \sum_{j=0}^{\infty} \psi_j \overbrace{\tilde{Z}_{n+m-j}}^{\text{innovation}} = \sum_{\boxed{j=m}}^{\infty} \psi_j Z_{n+m-j}. \qquad ✓ \tag{5.17}$$

- 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 (4.21), Z_t is completely determined by the conditioning set and hence $E(Z_t|X_n, \ldots, X_0, X_{-1}, \ldots) = Z_t$.

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

conditional expectation

$$\tilde{X}_{n+m} = E(X_{n+m} \mid X_1^n)$$

$$= \sum_{j\geq 0}^{\infty} \psi_j \underbrace{E(Z_{n+m-j} \mid X_1^n)} = \sum_{j\geq m}^{\infty} \psi_j Z_{n+m-j}$$

future

$$E(Z_{n+m-j} \mid X_1^n) = \begin{cases} 0 = E(Z_{n+m-j}) & \text{if } \underbrace{n+m-j > n} \equiv \underline{\underline{j<m}} \\ Z_{n+m-j} & \text{if } \underline{n+m-j} \leq \underline{n} \equiv j \geq m \end{cases}$$

past

Recursive form of prediction: In principle, the Z_t needed to create forecasts from (5.17) could be calculated using the infinite autoregressive representation (4.21). 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.

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

$$\xrightarrow[\text{expectation}]{\text{conditional}} \qquad E(Z_{n+m} \mid X_1^n) = \sum_{j=0}^{\infty} \pi_j E(X_{n+m-j} \mid X_1^n)$$

future noise

$$E(X_{n+m-j} \mid X_1^n) = \begin{cases} X_{n+m-j} & \underbrace{n+m-j \le n}_{\text{past}} \quad j \ge m \\ \tilde{X}_{n+m-j} & \underbrace{n+m-j > n}_{\text{future}} \quad j < m \end{cases}$$

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

Mean Squared Error for m- step ahead prediction: Subtracting (5.17) from (5.15) 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

$$\tilde{X}_{n+m} = \mu + \underbrace{\sum_{j=m}^{\infty} \psi_j Z_{n+m-j}}_{\rightarrow 0} \tag{5.22}$$

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

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

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

innovation

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

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

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

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

↑ based on ARMA(p,q) model

$$\begin{cases} \tilde{X}_t^n = 0 & t \leq 0 \\ \tilde{Z}_t^n = 0 & t \leq 0 \text{ or } \underbrace{t > n}_{\text{future}} \\ \tilde{X}_t^n = X_t & 1 \leq t \leq n \longrightarrow \text{observation} \end{cases}$$

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

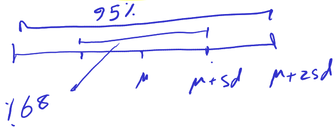
$$\tilde{X}^n_{n+m} \pm c_{\alpha/2}\sqrt{E(X_{n+m} - \tilde{X}_{n+m})^2} \tag{5.25}$$

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.


95%
.68
μ
μ+sd
μ+2sd