Chapter 5

Prediction

Prerequisites

• The best linear predictor.

• Some idea of what a basis of a vector space is.

Objectives

• Understand that prediction using a long past can be difficult because a large matrix has to be inverted, thus alternative, recursive method are often used to avoid direct inversion.

• Understand the derivation of the Levinson-Durbin algorithm, and why the coefficient, $\phi_{t,t}$, corresponds to the partial correlation between $X_1$ and $X_{t+1}$.

• Understand how these predictive schemes can be used write space of $\overline{sp}(X_t, X_{t-1}, \ldots, X_1)$ in terms of an orthogonal basis $\overline{sp}(X_t - P_{X_{t-1},X_{t-2},\ldots,X_1}(X_t), \ldots, X_1)$.

• Understand how the above leads to the Wold decomposition of a second order stationary time series.

• To understand how to approximate the prediction for an ARMA time series into a scheme which explicitly uses the ARMA structure. And this approximation improves geometrically, when the past is large.

One motivation behind fitting models to a time series is to forecast future unobserved observations - which would not be possible without a model. In this chapter we consider forecasting, based on the assumption that the model and/or autocovariance structure is known.
5.1 Forecasting given the present and infinite past

In this section we will assume that the linear time series \( \{X_t\} \) is both causal and invertible, that is

\[
X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} = \sum_{i=1}^{\infty} b_i X_{t-i} + \varepsilon_t, \tag{5.1}
\]

where \( \{\varepsilon_t\} \) are iid random variables (recall Definition 2.2.2). Both these representations play an important role in prediction. Furthermore, in order to predict \( X_{t+k} \) given \( X_t, X_{t-1}, \ldots \) we will assume that the infinite past is observed. In later sections we consider the more realistic situation that only the finite past is observed. We note that since \( X_t, X_{t-1}, X_{t-2}, \ldots \) is observed that we can obtain \( \varepsilon_\tau \) (for \( \tau \leq t \)) by using the invertibility condition

\[
\varepsilon_\tau = X_\tau - \sum_{i=1}^{\infty} b_i X_{\tau-i}.
\]

Now we consider the prediction of \( X_{t+k} \) given \( \{X_\tau; \tau \leq t\} \). Using the MA(\( \infty \)) presentation (since the time series is causal) of \( X_{t+k} \) we have

\[
X_{t+k} = \sum_{j=0}^{\infty} a_{j+k} \varepsilon_{t-j} + \sum_{j=0}^{k-1} a_j \varepsilon_{t+k-j},
\]

since \( E[\sum_{j=0}^{k-1} a_j \varepsilon_{t+k-j} | X_t, X_{t-1}, \ldots] = E[\sum_{j=0}^{k-1} a_j \varepsilon_{t+k-j}] = 0 \). Therefore, the best linear predictor of \( X_{t+k} \) given \( X_t, X_{t-1}, \ldots \), which we denote as \( X_t(k) \) is

\[
X_t(k) = \sum_{j=0}^{\infty} a_{j+k} \varepsilon_{t-j} = \sum_{j=0}^{\infty} a_{j+k}(X_{t-j} - \sum_{i=1}^{\infty} b_i X_{t-i-j}). \tag{5.2}
\]

\( X_t(k) \) is called the \( k \)-step ahead predictor and it is straightforward to see that it’s mean squared error is

\[
E[X_{t+k} - X_t(k)]^2 = E\left[\sum_{j=0}^{k-1} a_j \varepsilon_{t+k-j}\right]^2 = \text{var}[\varepsilon_t] \sum_{j=0}^{k} a_j^2, \tag{5.3}
\]

where the last line is due to the uncorrelatedness and zero mean of the innovations.

Often we would like to obtain the \( k \)-step ahead predictor for \( k = 1, \ldots, n \) where \( n \) is some
time in the future. We now explain how $X_t(k)$ can be evaluated recursively using the invertibility assumption.

Step 1 Use invertibility in (5.1) to give

$$X_t(1) = \sum_{i=1}^{\infty} b_i X_{t+1-i},$$

and $E[X_{t+1} - X_t(1)]^2 = \text{var}[\varepsilon_t]$

Step 2 To obtain the 2-step ahead predictor we note that

$$X_{t+2} = \sum_{i=2}^{\infty} b_i X_{t+2-i} + b_1 X_{t+1} + \varepsilon_{t+2}$$

$$= \sum_{i=2}^{\infty} b_i X_{t+2-i} + b_1 [X_t(1) + \varepsilon_{t+1}] + \varepsilon_{t+2},$$

thus it is clear that

$$X_t(2) = \sum_{i=2}^{\infty} b_i X_{t+2-i} + b_1 X_t(1)$$

and $E[X_{t+2} - X_t(2)]^2 = \text{var}[\varepsilon_t] (b_1^2 + 1) = \text{var}[\varepsilon_t] (a_2^2 + a_1^2)$.

Step 3 To obtain the 3-step ahead predictor we note that

$$X_{t+3} = \sum_{i=3}^{\infty} b_i X_{t+3-i} + b_2 X_{t+1} + b_1 X_{t+2} + \varepsilon_{t+3}$$

$$= \sum_{i=3}^{\infty} b_i X_{t+3-i} + b_2 (X_t(1) + \varepsilon_{t+1}) + b_1 (X_t(2) + b_1 \varepsilon_{t+1} + \varepsilon_{t+2}) + \varepsilon_{t+3}.$$

Thus

$$X_t(3) = \sum_{i=3}^{\infty} b_i X_{t+2-i} + b_2 X_t(1) + b_1 X_t(2)$$

and $E[X_{t+3} - X_t(3)]^2 = \text{var}[\varepsilon_t] [(b_2 + b_1^2)^2 + b_1^2 + 1] = \text{var}[\varepsilon_t] (a_3^2 + a_2^2 + a_1^2)$. 

135
Step $k$. Using the arguments above it is easily seen that

$$X_t(k) = \sum_{i=k}^{\infty} b_i X_{t+k-i} + \sum_{i=1}^{k-1} b_i X_t(k-i).$$

Thus the $k$-step ahead predictor can be recursively estimated.

We note that the predictor given above is based on the assumption that the infinite past is observed. In practice this is not a realistic assumption. However, in the special case that time series is an autoregressive process of order $p$ (with AR parameters $\{\phi_j\}_{j=1}^{p}$) and $X_t, \ldots, X_{t-m}$ is observed where $m \geq p-1$, then the above scheme can be used for forecasting. More precisely,

$$X_t(1) = \sum_{j=1}^{p} \phi_j X_{t+1-j}$$

$$X_t(k) = \sum_{j=k}^{p} \phi_j X_{t+k-j} + \sum_{j=1}^{k-1} \phi_j X_t(k-j) \text{ for } 2 \leq k \leq p$$

$$X_t(k) = \sum_{j=1}^{p} \phi_j X_t(k-j) \text{ for } k > p. \quad (5.4)$$

However, in the general case more sophisticated algorithms are required when only the finite past is known.

**Example: Forecasting yearly temperatures**

We now fit an autoregressive model to the yearly temperatures from 1880-2008 and use this model to forecast the temperatures from 2009-2013. In Figure 5.1 we give a plot of the temperature time series together with it’s ACF. It is clear there is some trend in the temperature data, therefore we have taken second differences, a plot of the second difference and its ACF is given in Figure 5.2.

We now use the command `ar.yule(res1, order.max=10)` (we will discuss in Chapter 7 how this function estimates the AR parameters) to estimate the the AR parameters. The function `ar.yule` uses the AIC to select the order of the AR model. When fitting the second differences from (from 1880-2008 - a data set of length of 127) the AIC chooses the AR(7) model

$$X_t = -1.1472X_{t-1} - 1.1565X_{t-2} - 1.0784X_{t-3} - 0.7745X_{t-4} - 0.6132X_{t-5} - 0.3515X_{t-6} - 0.1575X_{t-7} + \varepsilon_t,$$
Figure 5.1: Yearly temperature from 1880-2013 and the ACF.

Figure 5.2: Second differences of yearly temperature from 1880-2013 and its ACF.

with \( \text{var}[\varepsilon_t] = \sigma^2 = 0.02294 \). An ACF plot after fitting this model and then estimating the residuals \( \{\varepsilon_t\} \) is given in Figure 5.3. We observe that the ACF of the residuals ‘appears’ to be uncorrelated, which suggests that the AR(7) model fitted the data well. Later we cover the Ljung-Box test, which is a method for checking this claim. However since the residuals are estimated residuals and not the true residual, the results of this test need to be taken with a large pinch of salt. We will show that when the residuals are estimated from the data the error bars given in the ACF plot are not correct and the Ljung-Box test is not pivotal (as is assumed when deriving the limiting distribution
under the null the model is correct). By using the sequence of equations

- 

Figure 5.3: An ACF plot of the estimated residuals \( \{ \tilde{\varepsilon}_t \} \).

\[
\hat{X}_{127}(1) = -1.1472X_{127} - 1.1565X_{126} - 1.0784X_{125} - 0.7745X_{124} - 0.6132X_{123} \\
-0.3515X_{122} - 0.1575X_{121} \\
\hat{X}_{127}(2) = -1.1472\hat{X}_{127}(1) - 1.1565X_{127} - 1.0784X_{126} - 0.7745X_{125} - 0.6132X_{124} \\
-0.3515X_{123} - 0.1575X_{122} \\
\hat{X}_{127}(3) = -1.1472\hat{X}_{127}(2) - 1.1565\hat{X}_{127}(1) - 1.0784X_{127} - 0.7745X_{126} - 0.6132X_{125} \\
-0.3515X_{124} - 0.1575X_{123} \\
\hat{X}_{127}(4) = -1.1472\hat{X}_{127}(3) - 1.1565\hat{X}_{127}(2) - 1.0784\hat{X}_{127}(1) - 0.7745X_{127} - 0.6132X_{126} \\
-0.3515X_{125} - 0.1575X_{124} \\
\hat{X}_{127}(5) = -1.1472\hat{X}_{127}(4) - 1.1565\hat{X}_{127}(3) - 1.0784\hat{X}_{127}(2) - 0.7745\hat{X}_{127}(1) - 0.6132X_{127} \\
-0.3515X_{126} - 0.1575X_{125}.
\]

We can use \( \hat{X}_{127}(1), \ldots, \hat{X}_{127}(5) \) as forecasts of \( X_{128}, \ldots, X_{132} \) (we recall are the second differences), which we then use to construct forecasts of the temperatures. A plot of the second difference forecasts together with the true values are given in Figure 5.4. From the forecasts of the second differences we can obtain forecasts of the original data. Let \( Y_t \) denote the temperature at time \( t \)
and $X_t$ its second difference. Then $Y_t = -Y_{t-2} + 2Y_{t-1} + X_t$. Using this we have

\[ \hat{Y}_{127}(1) = -Y_{126} + 2Y_{127} + X_{127}(1) \]
\[ \hat{Y}_{127}(2) = -Y_{127} + 2Y_{127}(1) + X_{127}(2) \]
\[ \hat{Y}_{127}(3) = -Y_{127}(1) + 2Y_{127}(2) + X_{127}(3) \]

and so forth.

We note that (5.3) can be used to give the mse error. For example

\[ E[X_{128} - \hat{X}_{127}(1)]^2 = \sigma_t^2 \]
\[ E[X_{128} - \hat{X}_{127}(1)]^2 = (1 + \phi_t^2)\sigma_t^2 \]

If we believe the residuals are Gaussian we can use the mean squared error to construct confidence intervals for the predictions. Assuming for now that the parameter estimates are the true parameters (this is not the case), and $X_t = \sum_{j=0}^{\infty} \psi_j(\hat{\phi})\varepsilon_{t-j}$ is the MA($\infty$) representation of the AR(7) model, the mean square error for the $k$th ahead predictor is

\[ \sigma^2 \sum_{j=0}^{k-1} \psi_j(\hat{\phi})^2 \text{ (using (5.3))} \]

thus the 95% CI for the prediction is

\[ X_t(k) \pm 1.96\sigma^2 \sum_{j=0}^{k-1} \psi_j(\hat{\phi})^2, \]

however this confidence interval for not take into account $X_t(k)$ uses only parameter estimators and not the true values. In reality we need to take into account the approximation error here too.

If the residuals are not Gaussian, the above interval is not a 95% confidence interval for the prediction. One way to account for the non-Gaussianity is to use bootstrap. Specifically, we rewrite the AR(7) process as an MA($\infty$) process

\[ X_t = \sum_{j=0}^{\infty} \psi_j(\hat{\phi})\varepsilon_{t-j}. \]
Hence the best linear predictor can be rewritten as

\[ X_t(k) = \sum_{j=k}^{\infty} \psi_j(\hat{\phi}) \varepsilon_{t+k-j} \]

thus giving the prediction error

\[ X_{t+k} - X_t(k) = \sum_{j=0}^{k-1} \psi_j(\hat{\phi}) \varepsilon_{t+k-j}. \]

We have the prediction estimates, therefore all we need is to obtain the distribution of \( \sum_{j=0}^{k-1} \psi_j(\hat{\phi}) \varepsilon_{t+k-j} \).

This can be done by estimating the residuals and then using bootstrap\(^1\) to estimate the distribution of \( \sum_{j=0}^{k-1} \psi_j(\hat{\phi}) \varepsilon_{t+k-j} \), using the empirical distribution of \( \sum_{j=0}^{k-1} \psi_j(\hat{\phi}) \varepsilon_{t+k-j}^* \). From this we can construct the 95% CI for the forecasts.

---

\(^1\)Residual bootstrap is based on sampling from the empirical distribution of the residuals i.e. construct the “bootstrap” sequence \( \{ \varepsilon_{t+k-j}^* \}_j \) by sampling from the empirical distribution \( \hat{F}(x) = \frac{1}{n} \sum_{t=p+1}^{n} I(\hat{\varepsilon}_t \leq x) \) (where \( \hat{\varepsilon}_t = X_t - \sum_{j=1}^{p} \hat{\phi}_j X_{t-j} \)). This sequence is used to construct the bootstrap estimator \( \sum_{j=0}^{k-1} \psi_j(\hat{\phi}) \varepsilon_{t+k-j}^* \). By doing this several thousand times we can evaluate the empirical distribution of \( \sum_{j=0}^{k-1} \psi_j(\hat{\phi}) \varepsilon_{t+k-j}^* \) using these bootstrap samples. This is an estimator of the distribution function of \( \sum_{j=0}^{k-1} \psi_j(\hat{\phi}) \varepsilon_{t+k-j} \).
series of length of 127. It may be more appropriate to fit an ARMA model to this time series.

**Exercise 5.1** In this exercise we analyze the Sunspot data found on the course website. In the data analysis below only use the data from 1700 - 2003 (the remaining data we will use for prediction). In this section you will need to use the function `ar.yw` in R.

(i) Fit the following models to the data and study the residuals (using the ACF). Using this decide which model

\[
X_t = \mu + A \cos(\omega t) + B \sin(\omega t) + \varepsilon_t \text{ or } X_t = \mu + \varepsilon_t
\]

is more appropriate (take into account the number of parameters estimated overall).

(ii) Use these models to forecast the sunspot numbers from 2004-2013.

```r
diff1 = global.mean[c(2:134)] - global.mean[c(1:133)]
diff2 = diff1[c(2:133)] - diff1[c(1:132)]
res1 = diff2[c(1:127)]
residualsar7 <- ar.yw(res1, order.max=10)$resid
residuals <- residualsar7[-c(1:7)]
Forecast using the above model
res = c(res1,rep(0,5))
```

141
5.2 Review of vector spaces

In next few sections we will consider prediction/forecasting for stationary time series. In particular to find the best linear predictor of \( X_{t+1} \) given the finite past \( X_t, \ldots, X_1 \). Setting up notation our aim is to find

\[
X_{t+1|t} = P_{X_t, \ldots, X_1}(X_{t+1}) = X_{t+1|t, \ldots, 1} = \sum_{j=1}^{t} \phi_{t,j} X_{t+1-j},
\]

where \( \{\phi_{t,j}\} \) are chosen to minimise the mean squared error \( \min_{\phi} \mathbb{E}(X_{t+1} - \sum_{j=1}^{t} \phi_{t,j} X_{t+1-j})^2 \).

Basic results from multiple regression show that

\[
\begin{pmatrix}
\phi_{t,1} \\
\vdots \\
\phi_{t,t}
\end{pmatrix} = \Sigma_t^{-1} \tau_t,
\]

where \( (\Sigma_t)_{i,j} = \mathbb{E}(X_i X_j) \) and \( (\tau_t)_i = \mathbb{E}(X_{t-i} X_{t+1}) \). Given the covariances this can easily be done. However, if \( t \) is large a brute force method would require \( O(t^3) \) computing operations to calculate (5.7). Our aim is to exploit stationarity to reduce the number of operations. To do this, we will briefly discuss the notion of projections on a space, which help in our derivation of computationally more efficient methods.

Before we continue we first discuss briefly the idea of a a vector space, inner product spaces, Hilbert spaces, spans and basis. A more complete review is given in Brockwell and Davis (1998), Chapter 2.

First a brief definition of a vector space. \( \mathcal{X} \) is called an vector space if for every \( x, y \in \mathcal{X} \) and \( a, b \in \mathbb{R} \) (this can be generalised to \( \mathbb{C} \)), then \( ax + by \in \mathcal{X} \). An inner product space is a vector space which comes with an inner product, in other words for every element \( x, y \in \mathcal{X} \) we can defined an innerproduct \( \langle x, y \rangle \), where \( \langle \cdot, \cdot \rangle \) satisfies all the conditions of an inner product. Thus for every element \( x \in \mathcal{X} \) we can define its norm as \( \|x\| = \langle x, x \rangle \). If the inner product space is complete (meaning the limit of every sequence in the space is also in the space) then the innerproduct space is a Hilbert space (see wiki).

**Example 5.2.1**  
(i) The classical example of a Hilbert space is the Euclidean space \( \mathbb{R}^n \) where the innerproduct between two elements is simply the scalar product, \( \langle x, y \rangle = \sum_{i=1}^{n} x_i y_i \).
(ii) The subset of the probability space \((\Omega, F, P)\), where all the random variables defined on \(\Omega\) have a finite second moment, i.e. \(E(X^2) = \int_{\Omega} X(\omega)^2 dP(\omega) < \infty\). This space is denoted as \(L^2(\Omega, F, P)\). In this case, the inner product is \(\langle X, Y \rangle = E(XY)\).

(iii) The function space \(L^2[\mathbb{R}, \mu]\), where \(f \in L^2[\mathbb{R}, \mu]\) if \(f\) is \(\mu\)-measureable and

\[
\int_{\mathbb{R}} |f(x)|^2 d\mu(x) < \infty,
\]

is a Hilbert space. For this space, the inner product is defined as

\[
\langle f, g \rangle = \int_{\mathbb{R}} f(x)g(x) d\mu(x).
\]

In this chapter we will not use this function space, but it will be used in Chapter ?? (when we prove the Spectral representation theorem).

It is straightforward to generalize the above to complex random variables and functions defined on \(\mathbb{C}\). We simply need to remember to take conjugates when defining the innerproduct, i.e. \(\langle X, Y \rangle = E(X\overline{Y})\) and \(\langle f, g \rangle = \int_{\mathbb{C}} f(z)\overline{g(z)} d\mu(z)\).

In this chapter our focus will be on certain spaces of random variables which have a finite variance.

**Basis**

The random variables \(\{X_t, X_{t-1}, \ldots, X_1\}\) span the space \(\mathcal{X}_t^1\) (denoted as \(\text{sp}(X_t, X_{t-1}, \ldots, X_1)\)), if for every \(Y \in \mathcal{X}_t^1\), there exists coefficients \(\{a_j \in \mathbb{R}\}\) such that

\[
Y = \sum_{j=1}^{t} a_j X_{t+1-j}.
\]  

(5.5)

Moreover, \(\text{sp}(X_t, X_{t-1}, \ldots, X_1) = \mathcal{X}_t^1\) if for every \(\{a_j \in \mathbb{R}\}, \sum_{j=1}^{t} a_j X_{t+1-j} \in \mathcal{X}_t^1\). We now define the basis of a vector space, which is closely related to the span. The random variables \(\{X_t, \ldots, X_1\}\) form a basis of the space \(\mathcal{X}_t^1\), if for every \(Y \in \mathcal{X}_t^1\) we have a representation (5.5) and this representation is unique. More precisely, there does not exist another set of coefficients \(\{b_j\}\) such that \(Y = \sum_{j=1}^{t} b_j X_{t+1-j}\). For this reason, one can consider a basis as the minimal span, that is the smallest set of elements which can span a space.

**Definition 5.2.1 (Projections)** The projection of the random variable \(Y\) onto the space spanned
by \( \mathbb{P}(X_t, X_{t-1}, \ldots, X_1) \) (often denoted as \( P_{X_t, X_{t-1}, \ldots, X_1}(Y) \)) is defined as \( P_{X_t, X_{t-1}, \ldots, X_1}(Y) = \sum_{j=1}^t c_j X_{t+1-j} \), where \( \{c_j\} \) are chosen such that the difference \( Y - P_{X_t, X_{t-1}, \ldots, X_1}(Y) \) is uncorrelated (orthogonal/parallel) to any element in \( \mathbb{P}(X_t, X_{t-1}, \ldots, X_1) \). In other words, \( P_{X_t, X_{t-1}, \ldots, X_1}(Y) \) is the best linear predictor of \( Y \) given \( X_t, \ldots, X_1 \).

**Orthogonal basis**

An orthogonal basis is a basis, where every element in the basis is orthogonal to every other element in the basis. It is straightforward to orthogonalize any given basis using the method of projections.

To simplify notation let \( X_{t|t-1} = P_{X_{t-1}, \ldots, X_1}(X_t) \). By definition, \( X_t - X_{t|t-1} \) is orthogonal to the space \( \mathbb{P}(X_{t-1}, X_{t-2}, \ldots, X_1) \). In other words \( X_t - X_{t|t-1} \) and \( X_s \) (\( 1 \leq s \leq t \)) are orthogonal \( \text{cov}(X_s, (X_t - X_{t|t-1})) \), and by a similar argument \( X_t - X_{t|t-1} \) and \( X_s - X_{s|s-1} \) are orthogonal.

Thus by using projections we have created an orthogonal basis \( X_1, (X_2 - X_{2|1}), \ldots, (X_t - X_{t|t-1}) \) of the space \( \mathbb{P}(X_1, (X_2 - X_{2|1}), \ldots, (X_t - X_{t|t-1})) \). By construction it clear that \( \mathbb{P}(X_1, (X_2 - X_{2|1}), \ldots, (X_t - X_{t|t-1})) \) is a subspace of \( \mathbb{P}(X_t, \ldots, X_1) \). We now show that \( \mathbb{P}(X_1, (X_2 - X_{2|1}), \ldots, (X_t - X_{t|t-1})) = \mathbb{P}(X_t, \ldots, X_1) \).

To do this we define the sum of spaces. If \( U \) and \( V \) are two orthogonal vector spaces (which share the same innerproduct), then \( y \in U \oplus V \), if there exists a \( u \in U \) and \( v \in V \) such that \( y = u + v \). By the definition of \( \mathcal{X}_t^1 \), it is clear that \( (X_t - X_{t|t-1}) \notin \mathcal{X}_t^1 \), but \( (X_t - X_{t|t-1}) \notin \mathcal{X}_t^1 \).

Hence \( \mathcal{X}_t^1 = \mathbb{P}(X_t - X_{t|t-1}) \oplus \mathcal{X}_{t-1}^1 \). Continuing this argument we see that \( \mathcal{X}_t^1 = \mathbb{P}(X_t - X_{t|t-1}) \oplus \mathbb{P}(X_{t-1} - X_{t-1|t-2}) \oplus \ldots \oplus \mathbb{P}(X_1) \). Hence \( \mathbb{P}(X_t, \ldots, X_1) = \mathbb{P}(X_t - X_{t|t-1}, \ldots, X_2 - X_{2|1}, X_1) \).

Therefore for every \( P_{X_t, \ldots, X_1}(Y) = \sum_{j=1}^t a_j X_{t+1-j} \), there exists coefficients \( \{b_j\} \) such that

\[
P_{X_t, \ldots, X_1}(Y) = P_{X_t, X_{t-1}, \ldots, X_2, X_2 - X_{2|1}, X_1}(Y) = \sum_{j=1}^t P_{X_{t+1-j}, X_{t+1-j} - X_{t+1-j|t-j}, X_1}(Y) = \sum_{j=1}^{t-1} b_j (X_{t+1-j} - X_{t+1-j|t-j}) + b_t X_1,
\]

where \( b_j = E(Y(X_j - X_{j|j-1}))/E(X_j - X_{j|j-1})^2 \). A useful application of orthogonal basis is the ease of obtaining the coefficients \( b_j \), which avoids the inversion of a matrix. This is the underlying idea behind the innovations algorithm proposed in Brockwell and Davis (1998), Chapter 5.

### 5.2.1 Spaces spanned by infinite number of elements

The notions above can be generalised to spaces which have an infinite number of elements in their basis (and are useful to prove Wold’s decomposition theorem). Let now construct the space spanned
by infinite number random variables \( \{X_t, X_{t-1}, \ldots \} \). As with anything that involves \( \infty \) we need to define precisely what we mean by an infinite basis. To do this we construct a sequence of subspaces, each defined with a finite number of elements in the basis. We increase the number of elements in the subspace and consider the limit of this space. Let \( \mathcal{X}_t^{-n} = \mathcal{sp}(X_t, \ldots, X_{-n}) \), clearly if \( m > n \), then \( \mathcal{X}_t^{-n} \subset \mathcal{X}_t^{-m} \). We define \( \mathcal{X}_t^{-\infty} \) as \( \mathcal{X}_t^{-\infty} = \bigcup_{n=1}^{\infty} \mathcal{X}_t^{-n} \), in other words if \( Y \in \mathcal{X}_t^{-\infty} \), then there exists an \( n \) such that \( Y \in \mathcal{X}_t^{-n} \). However, we also need to ensure that the limits of all the sequences lie in this infinite dimensional space, therefore we close the space by defining defining a new space which includes the old space and also includes all the limits. To make this precise suppose the sequence of random variables is such that \( Y_s \in \mathcal{X}_t^{-s} \), and \( E(Y_{s_1} - Y_{s_2})^2 \to 0 \) as \( s_1, s_2 \to \infty \). Since the sequence \( \{Y_s\} \) is a Cauchy sequence there exists a limit. More precisely, there exists a random variable \( Y \), such that \( E(Y_s - Y)^2 \to 0 \) as \( s \to \infty \). Since the closure of the space, \( \overline{\mathcal{X}_t^{-n}} \), contains the set \( \mathcal{X}_t^{-n} \) and all the limits of the Cauchy sequences in this set, then \( Y \in \overline{\mathcal{X}_t^{-\infty}} \). We let

\[
\overline{\mathcal{X}_t^{-\infty}} = \mathcal{sp}(X_t, X_{t-1}, \ldots),
\]

**The orthogonal basis of \( \mathcal{sp}(X_t, X_{t-1}, \ldots) \)**

An orthogonal basis of \( \mathcal{sp}(X_t, X_{t-1}, \ldots) \) can be constructed using the same method used to orthogonalize \( \mathcal{sp}(X_t, X_{t-1}, \ldots, X_1) \). The main difference is how to deal with the initial value, which in the case of \( \mathcal{sp}(X_t, X_{t-1}, \ldots, X_1) \) is \( X_1 \). The analogous version of the initial value in infinite dimension space \( \mathcal{sp}(X_t, X_{t-1}, \ldots) \) is \( X_{-\infty} \), but this it not a well defined quantity (again we have to be careful with these pesky infinities).

Let \( X_{t-1}(1) \) denote the best linear predictor of \( X_t \) given \( X_{t-1}, X_{t-2}, \ldots \). As in Section 5.2 it is clear that \( (X_t - X_{t-1}(1)) \) and \( X_s \) for \( s \leq t-1 \) are uncorrelated and \( \overline{\mathcal{X}_t^{-\infty}} = \mathcal{sp}(X_t - X_{t-1}(1)) \oplus \overline{\mathcal{X}_{t-1}^{-\infty}} \), where \( \overline{\mathcal{X}_t^{-\infty}} = \mathcal{sp}(X_t, X_{t-1}, \ldots) \). Thus we can construct the orthogonal basis \( (X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots \) and the corresponding space \( \mathcal{sp}(X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots) \). It is clear that \( \mathcal{sp}(X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots \subset \mathcal{sp}(X_t, X_{t-1}, \ldots) \). However, unlike the finite dimensional case it is not clear that they are equal, roughly speaking this is because \( \mathcal{sp}(X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots \) lacks the initial value \( X_{-\infty} \). Of course the time \( -\infty \) in the past is not really a well defined quantity. Instead, the way we overcome this issue is that we define the initial starting random variable as the intersection of the subspaces, more precisely let \( \mathcal{X}_{-\infty} = \cap_{n=-\infty}^{\infty} \mathcal{X}_t^{-\infty} \).

Furthermore, we note that since \( X_n - X_{n-1}(1) \) and \( X_s \) (for any \( s \leq n \)) are orthogonal, then \( \mathcal{sp}(X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots) \) and \( \mathcal{X}_{-\infty} \) are orthogonal spaces. Using \( \mathcal{X}_{-\infty} \), we have
\[ \Theta_{j=0}^{t} \mathbb{P}(X_{t-j} - X_{t-j-1}(1)) \oplus \mathcal{A}_{\infty} = \mathbb{P}(X_t, X_{t-1}, \ldots). \]

We will use this result when we prove the Wold decomposition theorem (in Section 5.7).

### 5.3 Levinson-Durbin algorithm

We recall that in prediction the aim is to predict \( X_{t+1} \) given \( X_t, X_{t-1}, \ldots, X_1 \). The best linear predictor is

\[
X_{t+1|t} = P_{X_1, \ldots, X_t} (X_{t+1}) = X_{t+1|t, \ldots, 1} = \sum_{j=1}^{t} \phi_{t,j} X_{t+1-j}, \tag{5.7}
\]

where \( \{\phi_{t,j}\} \) are chosen to minimise the mean squared error, and are the solution of the equation

\[
\begin{pmatrix}
\phi_{t,1} \\
\vdots \\
\phi_{t,t}
\end{pmatrix} = \Sigma_{t}^{-1} \Sigma_t,
\tag{5.8}
\]

where \( (\Sigma_t)_{i,j} = \text{E}(X_i X_j) \) and \( (\Sigma_t)_i = \text{E}(X_{t-i} X_{t+1}) \). Using standard methods, such as Gauss-Jordan elimination, to solve this system of equations requires \( O(t^3) \) operations. However, we recall that \( \{X_t\} \) is a stationary time series, thus \( \Sigma_t \) is a Toeplitz matrix, by using this information in the 1940s Norman Levinson proposed an algorithm which reduced the number of operations to \( O(t^2) \). In the 1960s, Jim Durbin adapted the algorithm to time series and improved it.

We first outline the algorithm. We recall that the best linear predictor of \( X_{t+1} \) given \( X_t, \ldots, X_1 \) is

\[
X_{t+1|t} = \sum_{j=1}^{t} \phi_{t,j} X_{t+1-j}. \tag{5.9}
\]

The mean squared error is \( r(t+1) = \text{E}[X_{t+1} - X_{t+1|t}]^2 \). Given that the second order stationary covariance structure, the idea of the Levinson-Durbin algorithm is to recursively estimate \( \{\phi_{t,j}; j = 1, \ldots, t\} \) given \( \{\phi_{t-1,j}; j = 1, \ldots, t-1\} \) (which are the coefficients of the best linear predictor of \( X_t \) given \( X_{t-1}, \ldots, X_1 \)). Let us suppose that the autocovariance function \( c(k) = \text{cov}[X_0, X_k] \) is known.

The Levinson-Durbin algorithm is calculated using the following recursion.

Step 1  \( \phi_{1,1} = c(1)/c(0) \) and \( r(2) = \text{E}[X_2 - X_{2|1}]^2 = \text{E}[X_2 - \phi_{1,1} X_1]^2 = 2c(0) - 2\phi_{1,1}c(1) \).
Step 2 For $j = t$

$$\phi_{t,t} = \frac{c(t) - \sum_{j=1}^{t-1} \phi_{t-1,j} c(t-j)}{r(t)}$$

$$\phi_{t,j} = \phi_{t-1,j} - \phi_{t,t} \phi_{t-1,t-j} \quad 1 \leq j \leq t - 1,$$

and $r(t+1) = r(t)(1 - \phi_{t,t}^2)$.

We give two proofs of the above recursion.

**Exercise 5.2**

(i) Suppose $X_t = \phi X_{t-1} + \varepsilon_t$ (where $|\phi| < 1$). Use the Levinson-Durbin algorithm, to deduce an expression for $\phi_{t,j}$ for $(1 \leq j \leq t)$.

(ii) Suppose $X_t = \phi \varepsilon_{t-1} + \varepsilon_t$ (where $|\phi| < 1$). Use the Levinson-Durbin algorithm (and possibly Maple/Matlab), deduce an expression for $\phi_{t,j}$ for $(1 \leq j \leq t)$. (recall from Exercise 3.4 that you already have an analytic expression for $\phi_{t,t}$).

### 5.3.1 A proof based on projections

Let us suppose $\{X_t\}$ is a zero mean stationary time series and $c(k) = \text{E}(X_k X_0)$. Let $P_{X_t,...,X_2}(X_1)$ denote the best linear predictor of $X_1$ given $X_t, \ldots, X_2$ and $P_{X_t,...,X_2}(X_{t+1})$ denote the best linear predictor of $X_{t+1}$ given $X_t, \ldots, X_2$. Stationarity means that the following predictors share the same coefficients

$$X_{t|t-1} = \sum_{j=1}^{t-1} \phi_{t-1,j} X_{t-j} \quad P_{X_t,...,X_2}(X_{t+1}) = \sum_{j=1}^{t-1} \phi_{t-1,j} X_{t+1-j} \quad (5.10)$$

$$P_{X_t,...,X_2}(X_1) = \sum_{j=1}^{t-1} \phi_{t-1,j} X_{j+1}.$$

The last line is because stationarity means that flipping a time series round has the same correlation structure. These three relations are an important component of the proof.

Recall our objective is to derive the coefficients of the best linear predictor of $P_{X_t,...,X_1}(X_{t+1})$ based on the coefficients of the best linear predictor $P_{X_{t-1},...,X_1}(X_t)$. To do this we partition the space $\mathcal{sp}(X_t, \ldots, X_2, X_1)$ into two orthogonal spaces $\mathcal{sp}(X_t, \ldots, X_2, X_1) = \mathcal{sp}(X_t, \ldots, X_2, X_1) \oplus \mathcal{sp}(X_t, \ldots, X_2, X_1)$.
Therefore by uncorrelatedness we have the partition

\[
X_{t+1|t} = P_{X_t,...,X_2(X_{t+1})} + P_{X_t-P_{X_t,...,X_2}(X_t)}(X_{t+1})
\]

\[
= \sum_{j=1}^{t-1} \phi_{t-1,j} X_{t+1-j} + \phi_{t,t} (X_1 - P_{X_t,...,X_2}(X_t))
\]

by (5.10)

\[
= \sum_{j=1}^{t-1} \phi_{t-1,j} X_{t+1-j} + \phi_{t,t} \left( X_1 - \sum_{j=1}^{t-1} \phi_{t-1,j} X_{j+1} \right).
\]

(5.11)

We start by evaluating an expression for \(\phi_{t,t}\) (which in turn will give the expression for the other coefficients). It is straightforward to see that

\[
\phi_{t,t} = \frac{E(X_{t+1}(X_1 - P_{X_t,...,X_2}(X_1)))}{E(X_1 - P_{X_t,...,X_2}(X_1))^2}
\]

\[
= \frac{E[(X_{t+1} - P_{X_t,...,X_2}(X_{t+1}) + P_{X_t,...,X_2}(X_{t+1}))(X_1 - P_{X_t,...,X_2}(X_1))]}{E(X_1 - P_{X_t,...,X_2}(X_1))^2}
\]

\[
= \frac{E[(X_{t+1} - P_{X_t,...,X_2}(X_{t+1}))(X_1 - P_{X_t,...,X_2}(X_1))]}{E(X_1 - P_{X_t,...,X_2}(X_1))^2}
\]

(5.12)

Therefore we see that the numerator of \(\phi_{t,t}\) is the partial covariance between \(X_{t+1}\) and \(X_1\) (see Section 3.2.2), furthermore the denominator of \(\phi_{t,t}\) is the mean squared prediction error, since by stationarity

\[
E(X_1 - P_{X_t,...,X_2}(X_1))^2 = E(X_t - P_{X_{t-1},...,X_1}(X_t))^2 = r(t)
\]

(5.13)

Returning to (5.12), expanding out the expectation in the numerator and using (5.13) we have

\[
\phi_{t,t} = \frac{E(X_{t+1}(X_1 - P_{X_t,...,X_2}(X_1)))}{r(t)} = \frac{c(0) - E[X_{t+1}P_{X_t,...,X_2}(X_1)]}{r(t)} = \frac{c(0) - \sum_{j=1}^{t-1} \phi_{t-1,j} c(t-j)}{r(t)},
\]

(5.14)

which immediately gives us the first equation in Step 2 of the Levinson-Durbin algorithm. To
obtain the recursion for \( \phi_{t,j} \) we use (5.11) to give

\[
X_{t+1 | t} = \sum_{j=1}^{t} \phi_{t,j} X_{t+1-j} \\
= \sum_{j=1}^{t-1} \phi_{t-1,j} X_{t+1-j} + \phi_{t,t} \left( X_1 - \sum_{j=1}^{t-1} \phi_{t-1,j} X_{j+1} \right).
\]

To obtain the recursion we simply compare coefficients to give

\[
\phi_{t,j} = \phi_{t-1,j} - \phi_{t,t} \phi_{t-1,t-j} \quad 1 \leq j \leq t - 1.
\]

This gives the middle equation in Step 2. To obtain the recursion for the mean squared prediction error we note that by orthogonality of \( \{X_t, \ldots, X_2\} \) and \( X_1 - P_{X_t,\ldots,X_2}(X_1) \) we use (5.11) to give

\[
r(t+1) = E(X_{t+1} - X_{t+1 | t})^2 = E[X_{t+1} - P_{X_t,\ldots,X_2}(X_{t+1}) - \phi_{t,t} (X_1 - P_{X_t,\ldots,X_2}(X_1))^2
\]
\[
= E[X_{t+1} - P_{X_t,\ldots,X_2}(X_{t+1})]^2 + \phi_{t,t}^2 E[X_1 - P_{X_t,\ldots,X_2}(X_1)]^2
\]
\[
- 2\phi_{t,t} E[(X_{t+1} - P_{X_t,\ldots,X_2}(X_{t+1}))(X_1 - P_{X_t,\ldots,X_2}(X_1))] \\
= r(t) + \phi_{t,t}^2 r(t) - 2\phi_{t,t} E[X_{t+1}(X_1 - P_{X_t,\ldots,X_2}(X_1))]
\]
\[
= r(t)[1 - \phi_{t,t}^2].
\]

This gives the final part of the equation in Step 2 of the Levinson-Durbin algorithm.


### 5.3.2 A proof based on symmetric Toeplitz matrices

We now give an alternative proof which is based on properties of the (symmetric) Toeplitz matrix. We use (5.8), which is a matrix equation where

\[
\Sigma_t \begin{pmatrix}
\phi_{t,1} \\
\vdots \\
\phi_{t,t}
\end{pmatrix} = r_t,
\]  

(5.15)
with
\[\Sigma_t = \begin{pmatrix}
c(0) & c(1) & c(2) & \ldots & c(t-1) \\
c(1) & c(0) & c(1) & \ldots & c(t-2) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
c(t-1) & c(t-2) & \vdots & \vdots & c(0)
\end{pmatrix}\]
and
\[\tau_t = \begin{pmatrix}
c(1) \\
c(2) \\
\vdots \\
c(t)
\end{pmatrix}.
\]

The proof is based on embedding \(\tau_{t-1}\) and \(\Sigma_{t-1}\) into \(\Sigma_t\) and using that \(\Sigma_{t-1}\phi_{t-1} = \tau_{t-1}\).

To do this, we define the \((t-1) \times (t-1)\) matrix \(E_{t-1}\) which basically swops round all the elements in a vector
\[E_{t-1} = \begin{pmatrix}
0 & 0 & 0 & \ldots & 0 & 1 \\
0 & 0 & 0 & \ldots & 1 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
1 & 0 & \vdots & \vdots & 0 & 0 & 0
\end{pmatrix},
\]
(recall we came across this swopping matrix in Section 3.2.2). Using the above notation, we have
the interesting block matrix structure
\[\Sigma_t = \begin{pmatrix}
\Sigma_{t-1} & E_{t-1}\tau_{t-1} \\
r'_{t-1}E_{t-1} & c(0)
\end{pmatrix}
\]
and
\[\tau_t = (r'_{t-1}, c(t))'.
\]

Returning to the matrix equations in (5.15) and substituting the above into (5.15) we have
\[\Sigma_t \phi_{t-1} = \tau_t, \quad \Rightarrow \quad \begin{pmatrix}
\Sigma_{t-1} & E_{t-1}\tau_{t-1} \\
r'_{t-1}E_{t-1} & c(0)
\end{pmatrix} \begin{pmatrix}
\phi_{t-1,t} \\
\phi_{t,t}
\end{pmatrix} = \begin{pmatrix}
\tau_{t-1} \\
c(t)
\end{pmatrix},
\]
where \(\phi'_{t-1,t} = (\phi_{1,t}, \ldots, \phi_{t-1,t})\). This leads to the two equations
\[\Sigma_{t-1}\phi_{t-1,t} + E_{t-1}\tau_{t-1}\phi_{t,t} = \tau_{t-1} \quad (5.16)
\]
\[r'_{t-1}E_{t-1}\phi_{t-1,t} + c(0)\phi_{t,t} = c(t). \quad (5.17)
\]

We first show that equation (5.16) corresponds to the second equation in the Levinson-Durbin
algorithm. Multiplying (5.16) by $\Sigma_{t-1}^{-1}$, and rearranging the equation we have

$$
\Phi_{t-1,t} = \frac{\Sigma_{t-1}^{-1} Y_{t-1} - \Sigma_{t-1}^{-1} E_{t-1} E_{t-1}' \Phi_{t,t}}{-E_{t-1} E_{t-1}' \Phi_{t,t}}.
$$

Thus we have

$$
\Phi_{t-1,t} = \Phi_{t-1} - \phi_{t,t} E_{t-1} E_{t-1}' \Phi_{t,t}.
$$

This proves the second equation in Step 2 of the Levinson-Durbin algorithm.

We now use (5.17) to obtain an expression for $\phi_{t,t}$, which is the first equation in Step 1. Substituting (5.18) into $\Phi_{t-1,t}$ of (5.17) gives

$$
r_{t-1}' E_{t-1} \left( \Phi_{t-1} - \phi_{t,t} E_{t-1} \Phi_{t-1} \right) + c(0) \phi_{t,t} = c(t).
$$

Thus solving for $\phi_{t,t}$ we have

$$
\phi_{t,t} = \frac{c(t) - r_{t-1}' E_{t-1} \Phi_{t-1}}{c(0) - r_{t-1}' E_{t-1} \Phi_{t-1}}.
$$

Noting that $r(t) = c(0) - r_{t-1}' E_{t-1} \Phi_{t-1}$. (5.20) is the first equation of Step 2 in the Levinson-Durbin equation.

Note from this proof we do not need that the (symmetric) Toeplitz matrix is positive semi-definite. See Pourahmadi (2001), Chapter 7.

5.3.3 Using the Durbin-Levinson to obtain the Cholesky decomposition of the precision matrix

We recall from Section 3.2.1 that by sequentially projecting the elements of random vector on the past elements in the vector gives rise to Cholesky decomposition of the inverse of the variance/covariance (precision) matrix. This is exactly what was done in when we make the Durbin-Levinson
algorithm. In other words,

\[
\begin{pmatrix}
\frac{X_1}{\sqrt{r(1)}} \\
\frac{X_1 - \phi_{1,1}X_2}{\sqrt{r(2)}} \\
\vdots \\
\frac{X_n - \sum_{j=1}^{n-1} \phi_{n-1,j}X_{n-j}}{\sqrt{r(n)}}
\end{pmatrix} = I_n
\]

Therefore, if \( \Sigma_n = \text{var}[X_n] \), where \( X_n = (X_1, \ldots, X_n) \), then \( \Sigma_n^{-1} = L_nD_nL_n' \), where

\[
L_n = \begin{pmatrix}
1 & 0 & \ldots & \ldots & \ldots & 0 \\
-\phi_{1,1} & 1 & 0 & \ldots & \ldots & 0 \\
-\phi_{2,2} & -\phi_{2,1} & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
-\phi_{n-1,n-1} & -\phi_{n-1,n-2} & -\phi_{n-1,n-3} & \ldots & \ldots & 1
\end{pmatrix}
\]

(5.21)

and \( D_n = \text{diag}(r_1^{-1}, r_2^{-1}, \ldots, r_n^{-1}) \).

### 5.4 Forecasting for ARMA processes

Given the autocovariance of any stationary process the Levinson-Durbin algorithm allows us to systematically obtain one-step predictors of second order stationary time series without directly inverting a matrix.

In this section we consider forecasting for a special case of stationary processes, the ARMA process. We will assume throughout this section that the parameters of the model are known.

We showed in Section 5.1 that if \( \{X_t\} \) has an AR\((p)\) representation and \( t > p \), then the best linear predictor can easily be obtained using (5.4). Therefore, when \( t > p \), there is no real gain in using the Levinson-Durbin for prediction of AR\((p)\) processes. However, we do use it in Section 7.1.1 for recursively obtaining estimators of autoregressive parameters at increasingly higher orders.

Similarly if \( \{X_t\} \) satisfies an ARMA\((p,q)\) representation, then the prediction scheme can be simplified. Unlike the AR\((p)\) process, which is \( p \)-Markovian, \( P_{X_t, X_{t-1}, \ldots, X_1}(X_{t+1}) \) does involve all regressors \( X_t, \ldots, X_1 \). However, some simplifications are still possible. To explain how, let us
suppose that $X_t$ satisfies the ARMA($p, q$) representation

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i},$$

where $\{\varepsilon_t\}$ are iid zero mean random variables and the roots of $\phi(z)$ and $\theta(z)$ lie outside the unit circle. For the analysis below, we define the variables $\{W_t\}$, where $W_t = X_t$ for $1 \leq t \leq p$ and for $t > \max(p, q)$ let $W_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$ (which is the MA($q$) part of the process). Since $X_{p+1} = \sum_{j=1}^{p} \phi_j X_{t+1-j} + W_{p+1}$ and so forth it is clear that $\mathbb{E}(X_1, \ldots, X_t) = \mathbb{E}(W_1, \ldots, W_t)$ (i.e., they are linear combinations of each other). We will show for $t > \max(p, q)$ that

$$X_{t+1|t} = P_{X_t, \ldots, X_t}(X_{t+1}) = \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i (X_{t+1-i} - X_{t+1-i}|t-i), \quad (5.22)$$

for some $\theta_{t,i}$ which can be evaluated from the autocovariance structure. To prove the result we use the following steps:

$$P_{X_t, \ldots, X_t}(X_{t+1}) = \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i X_{t+1-i}$$

$$= \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i (X_{t+1-i} - \mathbb{E}(X_{t+1-i} | X_t, \ldots, X_t))$$

$$= \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i P_{W_{t+1-i}, \ldots, W_{t+1-i}}(X_{t+1-i} - \mathbb{E}(X_{t+1-i} | X_t, \ldots, X_t))$$

$$= \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i P_{W_{t+1-i}, \ldots, W_{t+1-i}}(X_{t+1-i} - \mathbb{E}(X_{t+1-i} | X_t, \ldots, X_t))$$

$$= \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i P_{W_{t+1-i}, \ldots, W_{t+1-i}}(X_{t+1-i} - \mathbb{E}(X_{t+1-i} | X_t, \ldots, X_t))$$

$$= \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i (W_{t+1-i} - \mathbb{E}(W_{t+1-i} | X_t, \ldots, X_t))$$

$$= X_{t+1-i} - \mathbb{E}(X_{t+1-i} | X_t, \ldots, X_t), \quad (5.23)$$

this gives the desired result. Thus given the parameters $\{\theta_{t,i}\}$ is straightforward to construct the predictor $X_{t+1|t}$. It can be shown that $\theta_{t,i} \rightarrow \theta_i$ as $t \rightarrow \infty$ (see Brockwell and Davis (1998)).
Example 5.4.1 (MA(q)) In this case, the above result reduces to

$$\hat{X}_{t+1|t} = \sum_{i=1}^{q} \theta_{t,i} \left( X_{t+1-i} - \hat{X}_{t+1-i|t-i} \right).$$

We now state a few results which will be useful later.

Lemma 5.4.1 Suppose \( \{X_t\} \) is a stationary time series with spectral density \( f(\omega) \). Let \( X_t = (X_1, \ldots, X_t) \) and \( \Sigma_t = \text{var}(X_t) \).

(i) If the spectral density function is bounded away from zero (there is some \( \gamma > 0 \) such that \( \inf_\omega f(\omega) > 0 \)), then for all \( t \),

$$\min(\Sigma_t) \geq \gamma (\text{where } \lambda_{\min} \text{ and } \lambda_{\max} \text{ denote the smallest and largest absolute eigenvalues of the matrix}).$$

(ii) Further, \( \lambda_{\max}(\Sigma_t^{-1}) \leq \gamma^{-1} \).

(Since for symmetric matrices the spectral norm and the largest eigenvalue are the same, then \( \|\Sigma_t^{-1}\|_{\text{spec}} \leq \gamma^{-1} \)).

(iii) Analogously, \( \sup_\omega f(\omega) \leq M < \infty \), then \( \lambda_{\max}(\Sigma_t) \leq M \) (hence \( \|\Sigma_t\|_{\text{spec}} \leq M \)).

PROOF. See Chapter 8.

Remark 5.4.1 Suppose \( \{X_t\} \) is an ARMA process, where the roots \( \phi(z) \) and and \( \theta(z) \) have absolute value greater than \( 1 + \delta_1 \) and less than \( \delta_2 \), then the spectral density \( f(\omega) \) is bounded by

$$\text{var}(\xi_t) \leq f(\omega) \leq \text{var}(\xi_t) \frac{(1-\frac{1}{\delta_2})^{2p}}{(1-\frac{1}{\delta_1})^{2p}}.$$  
Therefore, from Lemma 5.4.1 we have that \( \lambda_{\max}(\Sigma_t) \) and \( \lambda_{\max}(\Sigma_t^{-1}) \) is bounded uniformly over \( t \).

The prediction can be simplified if we make a simple approximation (which works well if \( t \) is relatively large). For \( 1 \leq t \leq \max(p,q) \), set \( \hat{X}_{t+1|t} = X_t \) and for \( t > \max(p,q) \) we define the recursion

$$\hat{X}_{t+1|t} = \sum_{j=1}^{p} \phi_j X_{t+1-j} + \sum_{i=1}^{q} \theta_i (X_{t+1-i} - \hat{X}_{t+1-i|t-i}). \quad (5.24)$$

This approximation seems plausible, since in the exact predictor (5.23), \( \theta_{t,i} \rightarrow \theta_i \). Note that this approximation is often used the case of prediction of other models too. We now derive a bound
for this approximation. In the following proposition we show that the best linear predictor of \(X_{t+1}\) given \(X_1, \ldots, X_t, X_{t+1|t}\), the approximating predictor \(\hat{X}_{t+1|t}\) and the best linear predictor given the infinite past, \(X_t(1)\) are asymptotically equivalent. To do this we obtain expressions for \(X_t(1)\) and \(\hat{X}_{t+1|t}\)

\[
X_t(1) = \sum_{j=1}^{\infty} b_j X_{t+1-j} \quad \text{(since \(X_{t+1} = \sum_{j=1}^{\infty} b_j X_{t+1-j} + \epsilon_{t+1}\))}
\]

Furthermore, by iterating (5.24) backwards we can show that

\[
\hat{X}_{t+1|t} = \sum_{j=1}^{\max(p,q)} b_j X_{t+1-j} + \sum_{j=1}^{\max(p,q)} \gamma_j X_j
\]

where \(|\gamma_j| \leq C \rho^j\), with \(1/(1+\delta) < \rho < 1\) and the roots of \(\theta(z)\) are outside \((1+\delta)\). We give a proof in the remark below.

**Remark 5.4.2** We prove (5.25) for the MA(1) model \(X_t = \theta X_{t-1} + \epsilon_t\). We recall that \(X_{t-1}(1) = \sum_{j=0}^{t-1} (-\theta)^j X_{t-j-1}\) and

\[
\begin{align*}
\hat{X}_{t|t-1} &= \theta \left( X_{t-1} - \hat{X}_{t-1|t-2} \right) \\
X_t - \hat{X}_{t|t-1} &= -\theta \left( X_{t-1} - \hat{X}_{t-1|t-2} \right) + X_t \\
&= \sum_{j=0}^{t-1} (-\theta)^j X_{t-j-1} + (-\theta)^t \left( X_1 - \hat{X}_{1|0} \right).
\end{align*}
\]

Thus we see that the first \((t-1)\) coefficients of \(X_{t-1}(1)\) and \(\hat{X}_{t|t-1}\) match.

Next, we prove (5.25) for the ARMA(1,2). We first note that \(\mathbb{sp}(X_1, X_t, \ldots, X_t) = \mathbb{sp}(W_1, W_2, \ldots, W_t)\), where \(W_1 = X_1\) and for \(t \geq 2\) \(W_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t\). The corresponding approximating predictor is defined as \(\hat{W}_{2|1} = W_1, \hat{W}_{3|2} = W_2\) and for \(t > 3\)

\[
\hat{W}_{t|t-1} = \theta_1 [W_{t-1} - \hat{W}_{t-1|t-2}] + \theta_2 [W_{t-2} - \hat{W}_{t-2|t-3}].
\]
Note that by using (5.24), the above is equivalent to
\[
\begin{align*}
\hat{X}_{t+1|t} - \phi_1 X_t &= \theta_1 \left[ X_t - \hat{X}_{t|t-1} \right] + \theta_2 \left[ X_{t-1} - \hat{X}_{t-1|t-2} \right] \\
&= (W_t - \hat{W}_{t|t-1}) + \left( W_{t-1} - \hat{W}_{t-1|t-2} \right)
\end{align*}
\]

By subtracting the above from \( W_{t+1} \) we have
\[
W_{t+1} - \hat{W}_{t+1|t} = -\theta_1 (W_t - \hat{W}_{t|t-1}) - \theta_2 (W_{t-1} - \hat{W}_{t-1|t-2}) + W_{t+1}. \tag{5.26}
\]

It is straightforward to rewrite \( W_{t+1} - \hat{W}_{t+1|t} \) as the matrix difference equation
\[
\begin{pmatrix}
W_{t+1} - \hat{W}_{t+1|t} \\
W_t - \hat{W}_{t|t-1}
\end{pmatrix}
= \begin{pmatrix}
\theta_1 & \theta_2 \\
-1 & 0
\end{pmatrix}
\begin{pmatrix}
W_t - \hat{W}_{t|t-1} \\
W_{t-1} - \hat{W}_{t-1|t-2}
\end{pmatrix}
+ \begin{pmatrix}
W_{t+1} \\
0
\end{pmatrix}
\]

We now show that \( \varepsilon_{t+1} \) and \( W_{t+1} - \hat{W}_{t+1|t} \) lead to the same difference equation except for some initial conditions, it is this that will give us the result. To do this we write \( \varepsilon_t \) as function of \( \{W_t\} \) (the irreducible condition). We first note that \( \varepsilon_t \) can be written as the matrix difference equation
\[
\begin{pmatrix}
\varepsilon_{t+1} \\
\varepsilon_t
\end{pmatrix}
= \begin{pmatrix}
\theta_1 & \theta_2 \\
-1 & 0
\end{pmatrix}
\begin{pmatrix}
\varepsilon_t \\
\varepsilon_{t-1}
\end{pmatrix}
+ \begin{pmatrix}
W_{t+1} \\
0
\end{pmatrix}. \tag{5.27}
\]

Thus iterating backwards we can write
\[
\varepsilon_{t+1} = \sum_{j=0}^{\infty} (-1)^j [Q^j]_{(1,1)} W_{t+1-j} = \sum_{j=0}^{\infty} \tilde{b}_j W_{t+1-j},
\]

where \( \tilde{b}_j = (-1)^j [Q^j]_{(1,1)} \) (noting that \( \tilde{b}_0 = 1 \)) denotes the \((1,1)\)th element of the matrix \( Q^j \) (note we did something similar in Section 2.4.1). Furthermore the same iteration shows that
\[
\varepsilon_{t+1} = \sum_{j=0}^{t-3} (-1)^j [Q^j]_{(1,1)} W_{t+1-j} + (-1)^{t-2} [Q^{t-2}]_{(1,1)} \varepsilon_3 \\
= \sum_{j=0}^{t-3} \tilde{b}_j W_{t+1-j} + (-1)^{t-2} [Q^{t-2}]_{(1,1)} \varepsilon_3. \tag{5.28}
\]
Therefore, by comparison we see that

\[ \varepsilon_{t+1} - \sum_{j=0}^{t-3} \tilde{b}_j W_{t+1-j} = (-1)^{t-2} [Q^{t-2} \xi_3]_1 = \sum_{j=t-2}^{\infty} \tilde{b}_j W_{t+1-j}. \]

We now return to the approximation prediction in (5.26). Comparing (5.27) and (5.27) we see that they are almost the same difference equations. The only difference is the point at which the algorithm starts. \( \varepsilon_t \) goes all the way back to the start of time. Whereas we have set initial values for \( \tilde{W}_{2|1} = W_1, \tilde{W}_{3|2} = W_2 \), thus \( \tilde{\xi}_3 = (W_3 - W_2, W_2 - W_1) \). Therefore, by iterating both (5.27) and (5.27) backwards, focusing on the first element of the vector and using (5.28) we have

\[ \varepsilon_{t+1} - \tilde{\varepsilon}_{t+1} = (-1)^{t-2} [Q^{t-2} \xi_3]_1 + (-1)^{t-2} [Q^{t-2} \xi_3]_1 \]

Replacing \( W_t \) with \( X_t - \phi_1 X_{t-1} \) gives (5.25), where the \( b_j \) can be easily deduced from \( \tilde{b}_j \) and \( \phi_1 \).

**Proposition 5.4.1** Suppose \( \{X_t\} \) is an ARMA process where the roots of \( \phi(z) \) and \( \theta(z) \) have roots which are greater in absolute value than \( 1+\delta \). Let \( X_{t+1|t}, \tilde{X}_{t+1|t} \) and \( X_t(1) \) be defined as in (5.23), (5.24) and (5.2) respectively. Then

\[ \text{E}[X_{t+1|t} - \tilde{X}_{t+1|t}]^2 \leq K \rho^t, \]

(5.29)

\[ \text{E}[\tilde{X}_{t+1|t} - X_t(1)]^2 \leq K \rho^t \]

(5.30)

\[ |\text{E}[X_{t+1} - X_{t+1|t}]^2 - \sigma^2| \leq K \rho^t \]

(5.31)

for any \( \frac{1}{1+\delta} < \rho < 1 \) and \( \text{var}(\varepsilon_t) = \sigma^2 \).

**PROOF.** The proof of (5.29) becomes clear when we use the expansion \( X_{t+1} = \sum_{j=1}^{\infty} b_j X_{t+1-j} + \)
Evaluating the best linear predictor of $X_{t+1}$ given $X_t, \ldots, X_1$, using the autoregressive expansion gives

$$X_{t+1|t} = \sum_{j=1}^{\infty} b_j P_{X_t, \ldots, X_1}(X_{t+1-j}) + \frac{P_{X_t, \ldots, X_1}(\varepsilon_{t+1})}{C \implies j} = 0.$$

Therefore by using (5.25) we see that the difference between the best linear predictor and $\hat{X}_{t+1|t}$ is

$$X_{t+1|t} - \hat{X}_{t+1|t} = \sum_{j=-\max(p,q)}^{\infty} b_{t+j} P_{X_t, \ldots, X_1}(X_{j+1}) + \sum_{j=1}^{\max(p,q)} \gamma_j X_j = I + II.$$

By using (5.25), it is straightforward to show that the second term $E[I^2] = E[\sum_{j=1}^{\max(p,q)} \gamma_j X_{t-j}]^2 \leq C \rho^t$, therefore what remains is to show that $E[I^2]$ attains a similar bound. As Zijuan pointed out, by definitions of projections, $E[P_{X_t, \ldots, X_1}(X_{j+1})^2] \leq E[X_{j+1}^2]$, which immediately gives the bound, instead we use a more convoluted proof. To obtain a bound, we first obtain a bound for $E[P_{X_t, \ldots, X_1}(X_{j+1})^2]$. Basic results in linear regression shows that

$$P_{X_t, \ldots, X_1}(X_{j+1}) = \beta'_{j,t} X_t,$$

where $\beta_{j,t} = \Sigma_t^{-1} r_{t,j}$, with $\beta'_{j,t} = (\beta_{1,j,t}, \ldots, \beta_{j,t,t})$, $X'_t = (X_1, \ldots, X_t)$, $\Sigma_t = E(X_t X'_t)$ and $r_{t,j} = E(X_t X_j)$. Substituting (5.32) into $I$ gives

$$\sum_{j=-\max(p,q)}^{\infty} b_{t+j} P_{X_t, \ldots, X_1}(X_{j+1}) = \sum_{j=-\max(p,q)}^{\infty} b_{t+j} \beta'_{j,t} X_t = \left( \sum_{j=1}^{\infty} b_j r'_{t,j} \right) \Sigma_t^{-1} X_t.$$

Therefore the mean squared error of $I$ is

$$E[I^2] = \left( \sum_{j=-\max(p,q)}^{\infty} b_{t+j} r'_{t,j} \right) \Sigma_t^{-1} \left( \sum_{j=-\max(p,q)}^{\infty} b_{t+j} r_{t,j} \right).$$

To bound the above we use the Cauchy Schwarz inequality ($\|aBb\|_1 \leq \|a\|_2 \|B\|_2$), the spec-
tral norm inequality $\|a\|_2 \|Bb\|_2 \leq \|a\|_2 \|B_{\text{spec}}\|_2$ and Minkowski’s inequality $\|\sum_{j=1}^{n} a_j\|_2 \leq \sum_{j=1}^{n} \|a_j\|_2$ we have

$$E[I^2] \leq \left\| \sum_{j=1}^{\infty} b_{t+j} r'_{t,j} \right\|^2 \|\Sigma_1^{-1}\|_\text{spec}^2 \leq \left( \sum_{j=1}^{\infty} |b_{t+j}| \cdot \|r_{t,j}\|_2 \right)^2 \|\Sigma_1^{-1}\|_\text{spec}^2. \quad (5.34)$$

We now bound each of the terms above. We note that for all $t$, using Remark 5.4.1 that $\|\Sigma_1^{-1}\|_\text{spec} \leq K$ (for some constant $K$). We now consider $r'_{t,j} = (E(X_{1}X_{-j}), \ldots, E(X_{t}X_{-j})) = (c(1-j), \ldots, c(t-j))$. By using (3.2) we have $|c(k)| \leq C \rho^k$, therefore

$$\|r_{t,j}\|_2 \leq K \left( \sum_{r=1}^{t} \rho^{2(j+r)} \right)^{1/2} \leq K \frac{\rho^j}{(1-\rho^2)^2}. $$

Substituting these bounds into (5.34) gives $E[I^2] \leq K \rho^j$. Altogether the bounds for $I$ and $II$ give

$$E(X_{t+1|t} - \hat{X}_{t+1|t})^2 \leq K \frac{\rho^j}{(1-\rho^2)^2}. $$

Thus proving (5.29).

To prove (5.30) we note that

$$E[X_t(1) - \hat{X}_{t+1|t}]^2 = E \left[ \sum_{j=0}^{\infty} b_{t+j} X_{-j} + \sum_{j=t-\max(p,q)}^{t} b_j Y_{t-j} \right]^2. $$

Using the above and that $b_{t+j} \leq K \rho^{j+x}$, it is straightforward to prove the result.

Finally to prove (5.31), we note that by Minkowski’s inequality we have

$$\left[ E \left( X_{t+1} - X_{t+1|t} \right)^2 \right]^{1/2} \leq \left[ E \left( X_t(1) \right)^2 \right]^{1/2} + \left[ E \left( X_t(1) - \hat{X}_{t+1|t} \right)^2 \right]^{1/2} + \left[ E \left( \hat{X}_{t+1|t} - X_{t+1|t} \right)^2 \right]^{1/2} \leq K \rho^{j/2} \text{ by (5.30)}.$$ 

Thus giving the desired result. \[\square\]
5.5 Forecasting for nonlinear models

In this section we consider forecasting for nonlinear models. The forecasts we construct, may not necessarily/formally be the best linear predictor, because the best linear predictor is based on minimising the mean squared error, which we recall from Chapter 4 requires the existence of the higher order moments. Instead our forecast will be the conditional expection of $X_{t+1}$ given the past (note that we can think of it as the best linear predictor). Furthermore, with the exception of the ARCH model we will derive approximation of the conditional expectation/best linear predictor, analogous to the forecasting approximation for the ARMA model, $\hat{X}_{t+1|t}$ (given in (5.24)).

5.5.1 Forecasting volatility using an ARCH($p$) model

We recall the ARCH($p$) model defined in Section 4.2

$$X_t = \sigma_t Z_t \quad \sigma_t^2 = a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2.$$ 

Using a similar calculation to those given in Section 4.2.1, we see that

$$E[X_{t+1}|X_t, X_{t-1}, \ldots, X_{t-p+1}] = E(Z_{t+1}\sigma_{t+1}|X_t, X_{t-1}, \ldots, X_{t-p+1}) = \sigma_{t+1} E(Z_{t+1}|X_t, X_{t-1}, \ldots, X_{t-p+1})$$

$$= \sigma_{t+1} E(Z_{t+1}) = 0 \cdot \sigma_{t+1} = 0.$$ 

In other words, past values of $X_t$ have no influence on the expected value of $X_{t+1}$. On the other hand, in Section 4.2.1 we showed that

$$E(X_{t+1}^2|X_t, X_{t-1}, \ldots, X_{t-p+1}) = E(Z_{t+1}^2\sigma^2_{t+1}|X_t, X_{t-1}, \ldots, X_{t-p+1}) = \sigma_{t+1}^2 E[Z_{t+1}^2] = \sigma_{t+1}^2 = \sum_{j=1}^{p} a_j X_{t+1-j}^2,$$

thus $X_t$ has an influence on the conditional mean squared/variance. Therefore, if we let $X_{t+k|t}$ denote the conditional variance of $X_{t+k}$ given $X_t, \ldots, X_{t-p+1}$, it can be derived using the following
recursion

\[ X_{t+1|t}^2 = \sum_{j=1}^{p} a_j X_{t+1-j}^2 \]

\[ X_{t+k|t}^2 = \sum_{j=k}^{p} a_j X_{t+k-j}^2 + \sum_{j=1}^{k-1} a_j X_{t+k-j|k}^2 \quad 2 \leq k \leq p \]

\[ X_{t+k|t}^2 = \sum_{j=1}^{p} a_j X_{t+k-j|t}^2 \quad k > p. \]

### 5.5.2 Forecasting volatility using a GARCH(1, 1) model

We recall the GARCH(1, 1) model defined in Section 4.3

\[ \sigma_t^2 = \sigma_0^2 + a_1 X_{t-1}^2 + b_1 \sigma_{t-1}^2 = (a_1 \sigma_{t-1}^2 + b_1) \sigma_{t-1}^2 + a_0. \]

Similar to the ARCH model it is straightforward to show that \( \mathbb{E}[X_{t+1}|X_t, X_{t-1}, \ldots] = 0 \) (where we use the notation \( X_t, X_{t-1}, \ldots \) to denote the infinite past or more precisely conditioned on the sigma algebra \( \mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots) \)). Therefore, like the ARCH process, our aim is to predict \( X_t^2 \).

We recall from Example 4.3.1 that if the GARCH the process is invertible (satisfied if \( b < 1 \)), then

\[ \mathbb{E}[X_{t+1}^2|X_t, X_{t-1}, \ldots] = \sigma_{t+1}^2 = \sigma_0^2 + a_1 X_{t-1}^2 + b_1 \sigma_{t-1}^2 = \frac{a_0}{1 - b_1} + a_1 \sum_{j=0}^{\infty} b^j X_{t-j}^2. \]

(5.35)

Of course, in reality we only observe the finite past \( X_t, X_{t-1}, \ldots, X_1 \). We can approximate \( \mathbb{E}[X_{t+1}^2|X_t, X_{t-1}, \ldots, X_1] \) using the following recursion, set \( \sigma_{1|0}^2 = 0 \), then for \( t \geq 1 \)

\[ \hat{\sigma}_{t+1|t}^2 = \sigma_0^2 + a_1 X_t^2 + b_1 \sigma_{t|t-1}^2 \]

(noticing that this is similar in spirit to the recursive approximate one-step ahead predictor defined in (5.25)). It is straightforward to show that

\[ \sigma_{t+1|t}^2 = \frac{a_0(1 - b^{t+1})}{1 - b} + a_1 \sum_{j=0}^{t-1} b^j X_{t-j}^2, \]

taking note that this is not the same as \( \mathbb{E}[X_{t+1}^2|X_t, \ldots, X_1] \) (if the mean square error existed \( \mathbb{E}[X_{t+1}^2|X_t, \ldots, X_1] \) would give a smaller mean square error), but just like the ARMA process it will
closely approximate it. Furthermore, from (5.35) it can be seen that \( \hat{\sigma}_{t+1|t}^2 \) closely approximates \( \sigma_{t+1}^2 \).

**Exercise 5.3** To answer this question you need \texttt{R install.package("tseries")} then remember library("garch").

(i) You will find the Nasdaq data from 4th January 2010 - 15th October 2014 on my website.

(ii) By taking log differences fit a GARCH(1,1) model to the daily closing data (ignore the adjusted closing value) from 4th January 2010 - 30th September 2014 (use the function \texttt{garch(x, order = c(1, 1)) fit the GARCH(1,1) model}).

(iii) Using the fitted GARCH(1,1) model, forecast the volatility \( \sigma_{t}^2 \) from October 1st-15th (noting that no trading is done during the weekends). Denote these forecasts as \( \sigma_{t|0}^2 \). Evaluate \( \sum_{t=1}^{11} \sigma_{t|0}^2 \).

(iv) Compare this to the actual volatility \( \sum_{t=1}^{11} X_t^2 \) (where \( X_t \) are the log differences).

### 5.5.3 Forecasting using a BL(1, 0, 1, 1) model

We recall the Bilinear(1, 0, 1, 1) model defined in Section 4.4

\[
X_t = \phi_1 X_{t-1} + b_{1,1} X_{t-1} \varepsilon_{t-1} + \varepsilon_t.
\]

Assuming invertibility, so that \( \varepsilon_t \) can be written in terms of \( X_t \) (see Remark 4.4.2):

\[
\varepsilon_t = \sum_{j=0}^{\infty} \left( (-b)^j \prod_{i=0}^{j-1} X_{t-i} \right) [X_{t-j} - \phi X_{t-j-1}],
\]

it can be shown that

\[
X_t(1) = \mathbb{E}[X_{t+1}|X_t, X_{t-1}, \ldots] = \phi_1 X_t + b_{1,1} X_t \varepsilon_t.
\]

However, just as in the ARMA and GARCH case we can obtain an approximation, by setting \( \tilde{X}_{1|0} = 0 \) and for \( t \geq 1 \) defining the recursion

\[
\tilde{X}_{t+1|t} = \phi_1 X_t + b_{1,1} X_t \left( X_t - \tilde{X}_{t|t-1} \right).
\]
See ? and ? for further details.

**Remark 5.5.1 (How well does \( \hat{X}_{t+1} \) approximate \( X_t(1) \)?)** We now derive conditions for \( \hat{X}_{t+1} \) to be a close approximation of \( X_t(1) \) when \( t \) is large. We use a similar technique to that used in Remark 5.4.2.

We note that \( X_{t+1} - X_t(1) = \varepsilon_{t+1} \) (since a future innovation, \( \varepsilon_{t+1} \), cannot be predicted). We will show that \( X_{t+1} - \hat{X}_{t+1} \) is ‘close’ to \( \varepsilon_{t+1} \). Subtracting \( \hat{X}_{t+1} \) from \( X_{t+1} \) gives the recursion

\[
X_{t+1} - \hat{X}_{t+1} = -b_{1,1}(X_t - \hat{X}_{t|t-1})X_t + (b\varepsilon_t X_t + \varepsilon_{t+1}).
\]

We will compare the above recursion to the recursion based on \( \varepsilon_{t+1} \). Rearranging the bilinear equation gives

\[
\varepsilon_{t+1} = -(b\varepsilon_t X_t + (X_{t+1} - \phi_1 X_t)).
\]

We observe that (5.36) and (5.37) are almost the same difference equation, the only difference is that an initial value is set for \( bX_1|0 \). This gives the difference between the two equations as

\[
\varepsilon_{t+1} - [X_{t+1} - \hat{X}_{t+1}] = (-1)^t b^t X_1 \prod_{j=1}^t \varepsilon_j + (-1)^t b^t [X_1 - \hat{X}_1|0] \prod_{j=1}^t \varepsilon_j.
\]

Thus if \( b^t \prod_{j=1}^t \varepsilon_j \overset{a.s.}{\rightarrow} 0 \) as \( t \rightarrow \infty \), then \( \hat{X}_{t+1} \overset{P}{\rightarrow} X_t(1) \) as \( t \rightarrow \infty \). We now show that if \( E[\log |\varepsilon_t|] < -\log |b| \), then \( b^t \prod_{j=1}^t \varepsilon_j \overset{a.s.}{\rightarrow} 0 \). Since \( b^t \prod_{j=1}^t \varepsilon_j \) is a product, it seems appropriate to take logarithms to transform it into a sum. To ensure that it is positive, we take absolutes and \( t \)-roots

\[
\log |b^t \prod_{j=1}^t \varepsilon_j|^{1/t} = \log |b| + \frac{1}{t} \sum_{j=1}^t \log |\varepsilon_j|.
\]

Therefore by using the law of large numbers we have

\[
\log |b^t \prod_{j=1}^t \varepsilon_j|^{1/t} = \log |b| + \frac{1}{t} \sum_{j=1}^t \log |\varepsilon_j| \overset{P}{\rightarrow} \log |b| + E \log |\varepsilon_0| = \gamma.
\]

Thus we see that \( |b^t \prod_{j=1}^t \varepsilon_j|^{1/t} \overset{a.s.}{\rightarrow} \exp(\gamma) \). In other words, \( |b^t \prod_{j=1}^t \varepsilon_j| \approx \exp(t\gamma) \), which will only
converge to zero if \( \log |\epsilon_t| < - \log |b| \).

### 5.6 Nonparametric prediction

In this section we briefly consider how prediction can be achieved in the nonparametric world. Let us assume that \( \{X_t\} \) is a stationary time series. Our objective is to predict \( X_{t+1} \) given the past. However, we don’t want to make any assumptions about the nature of \( \{X_t\} \). Instead we want to obtain a predictor of \( X_{t+1} \) given \( X_t \) which minimises the means squared error, \( \text{E}[X_{t+1} - g(X_t)]^2 \). It is well known that this is conditional expectation \( \text{E}[X_{t+1} | X_t] \). (since \( \text{E}[X_{t+1} - g(X_t)]^2 = \text{E}[X_{t+1} - \text{E}(X_{t+1} | X_t)]^2 + \text{E}[(g(X_t) - \text{E}(X_{t+1} | X_t)]^2 \)). Therefore, one can estimate

\[
\text{E}[X_{t+1} | X_t = x] = m(x)
\]

nonparametrically. A classical estimator of \( m(x) \) is the Nadaraya-Watson estimator

\[
\hat{m}_n(x) = \frac{\sum_{t=1}^{n-1} X_{t+1} K\left(\frac{x - X_t}{b}\right)}{\sum_{t=1}^{n-1} K\left(\frac{x - X_t}{b}\right)},
\]

where \( K: \mathbb{R} \to \mathbb{R} \) is a kernel function (see Fan and Yao (2003), Chapter 5 and 6). Under some ‘regularity conditions’ it can be shown that \( \hat{m}_n(x) \) is a consistent estimator of \( m(x) \) and converges to \( m(x) \) in mean square (with the typical mean squared rate \( O(b^4 + (bn)^{-1}) \)). The advantage of going the non-parametric route is that we have not imposed any form of structure on the process (such as linear/(G)ARCH/Bilinear). Therefore, we do not run the risk of misspecifying the model. A disadvantage is that nonparametric estimators tend to be a lot worse than parametric estimators (in Chapter ?? we show that parametric estimators have \( O(n^{-1/2}) \) convergence which is faster than the nonparametric rate \( O(b^2 + (bn)^{-1/2}) \)). Another possible disadvantage is that if we wanted to include more past values in the predictor, ie. \( m(x_1, \ldots, x_d) = \text{E}[X_{t+1} | X_t = x_1, \ldots, X_{t-p} = x_d] \) then the estimator will have an extremely poor rate of convergence (due to the curse of dimensionality).

A possible solution to the problem is to assume some structure on the nonparametric model, and define a semi-parametric time series model. We state some examples below:

(i) An additive structure of the type

\[
X_t = \sum_{j=1}^{p} g_j(X_{t-j}) + \epsilon_t
\]
where \( \{\varepsilon_t\} \) are iid random variables.

(ii) A functional autoregressive type structure

\[
X_t = \sum_{j=1}^{p} g_j(X_{t-d})X_{t-j} + \varepsilon_t.
\]

(iii) The semi-parametric GARCH(1,1)

\[
X_t = \sigma_t Z_t, \quad \sigma_t^2 = b\sigma_{t-1}^2 + m(X_{t-1}).
\]

However, once a structure has been imposed, conditions need to be derived in order that the model has a stationary solution (just as we did with the fully-parametric models).

See ?, ?, ?, ?, ? etc.

\section{5.7 The Wold Decomposition}

Section 5.2.1 nicely leads to the Wold decomposition, which we now state and prove. The Wold decomposition theorem, states that any stationary process, has something that appears close to an MA(\(\infty\)) representation (though it is not). We state the theorem below and use some of the notation introduced in Section 5.2.1.

\textbf{Theorem 5.7.1} Suppose that \( \{X_t\} \) is a second order stationary time series with a finite variance (we shall assume that it has mean zero, though this is not necessary). Then \( X_t \) can be uniquely expressed as

\[
X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t,
\]

(5.38)

where \( \{Z_t\} \) are uncorrelated random variables, with \( \text{var}(Z_t) = \text{E}(X_t - X_{t-1(1)})^2 \) (noting that \( X_{t-1(1)} \) is the best linear predictor of \( X_t \) given \( X_{t-1}, X_{t-2}, \ldots \)) and \( V_t \in \mathcal{X}_{-\infty} = \cap_n^{\infty} \mathcal{X}_n^{\infty} \), where \( \mathcal{X}_n^{\infty} \) is defined in (5.6).

\textbf{PROOF.} First let is consider the one-step ahead prediction of \( X_t \) given the infinite past, denoted \( X_{t-1(1)} \). Since \( \{X_t\} \) is a second order stationary process it is clear that \( X_{t-1(1)} = \sum_{j=1}^{\infty} b_j X_{t-j} \), where the coefficients \( \{b_j\} \) do not vary with \( t \). For this reason \( \{X_{t-1(1)}\} \) and \( \{X_t - X_{t-1(1)}\} \) are
second order stationary random variables. Furthermore, since \( \{X_t - X_{t-1}(1)\} \) is uncorrelated with \( X_s \) for any \( s \leq t \), then \( \{X_s - X_{s-1}(1); s \in \mathbb{R}\} \) are uncorrelated random variables. Define \( Z_s = X_s - X_{s-1}(1) \), and observe that \( Z_s \) is the one-step ahead prediction error. We recall from Section 5.2.1 that \( X_t \in \mathbb{P}(X_t - X_{t-1}(1)), (X_{t-1} - X_{t-2}(1)), \ldots) \oplus \mathbb{P}(X_{-\infty}) = \bigoplus_{j=0}^{\infty} \mathbb{P}(Z_{t-j}) \oplus \mathbb{P}(X_{-\infty}). \) Since the spaces \( \bigoplus_{j=0}^{\infty} \mathbb{P}(Z_{t-j}) \) and \( \mathbb{P}(X_{-\infty}) \) are orthogonal, we shall first project \( X_t \) onto \( \bigoplus_{j=0}^{\infty} \mathbb{P}(Z_{t-j}) \), due to orthogonality the difference between \( X_t \) and its projection will be in \( \mathbb{P}(X_{-\infty}) \). This will lead to the Wold decomposition.

First we consider the projection of \( X_t \) onto the space \( \bigoplus_{j=0}^{\infty} \mathbb{P}(Z_{t-j}) \), which is

\[
P_{Z_t,Z_{t-1},...}(X_t) = \sum_{j=0}^{\infty} \psi_j Z_{t-j},
\]

where due to orthogonality \( \psi_j = \text{cov}(X_t, (X_{t-j} - X_{t-j-1}(1))) / \text{var}(X_{t-j} - X_{t-j-1}(1)) \). Since \( X_t \in \bigoplus_{j=0}^{\infty} \mathbb{P}(Z_{t-j}) \oplus \mathbb{P}(X_{-\infty}) \), the difference \( X_t - P_{Z_t,Z_{t-1},...}X_t \) is orthogonal to \( \{Z_t\} \) and belongs in \( \mathbb{P}(X_{-\infty}) \). Hence we have

\[
X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t,
\]

where \( V_t = X_t - \sum_{j=0}^{\infty} \psi_j Z_{t-j} \) and is uncorrelated to \( \{Z_t\} \). Hence we have shown (5.38). To show that the representation is unique we note that \( Z_t, Z_{t-1}, \ldots \) are an orthogonal basis of \( \mathbb{P}(Z_t, Z_{t-1}, \ldots) \), which pretty much leads to uniqueness.

\[\square\]

**Exercise 5.4** Consider the process \( X_t = A \cos(Bt + U) \) where \( A, B \) and \( U \) are random variables such that \( A, B \) and \( U \) are independent and \( U \) is uniformly distributed on \((0, 2\pi)\).

(i) Show that \( X_t \) is second order stationary (actually it’s stationary) and obtain its means and covariance function.

(ii) Show that the distribution of \( A \) and \( B \) can be chosen in such a way that \( \{X_t\} \) has the same covariance function as the MA(1) process \( Y_t = \varepsilon_t + \phi \varepsilon_{t-1} \) (where \( |\phi| < 1 \)) (quite amazing).

(iii) Suppose \( A \) and \( B \) have the same distribution found in (ii).

(a) What is the best predictor of \( X_{t+1} \) given \( X_t, X_{t-1}, \ldots \)?

(b) What is the best linear predictor of \( X_{t+1} \) given \( X_t, X_{t-1}, \ldots \)?
It is worth noting that variants on the proof can be found in Brockwell and Davis (1998), Section 5.7 and Fuller (1995), page 94.

**Remark 5.7.1** Notice that the representation in (5.38) looks like an MA(∞) process. There is, however, a significant difference. The random variables \( \{Z_t\} \) of an MA(∞) process are iid random variables and not just uncorrelated.

We recall that we have already come across the Wold decomposition of some time series. In Section 3.3 we showed that a non-causal linear time series could be represented as a causal ‘linear time series’ with uncorrelated but dependent innovations. Another example is in Chapter 4, where we explored ARCH/GARCH process which have an AR and ARMA type representation. Using this representation we can represent ARCH and GARCH processes as the weighted sum of \( \{(Z^2_t - 1)\sigma^2_t\} \) which are uncorrelated random variables.

**Remark 5.7.2 (Variation on the Wold decomposition)** In many technical proofs involving time series, we often use results related to the Wold decomposition. More precisely, we often decompose the time series in terms of an infinite sum of martingale differences. In particular, we define the sigma-algebra \( \mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots) \), and suppose that \( \mathbb{E}(X_t|\mathcal{F}_\infty) = \mu \). Then by telescoping we can formally write \( X_t \) as

\[
X_t - \mu = \sum_{j=0}^{\infty} Z_{t,j}
\]

where \( Z_{t,j} = \mathbb{E}(X_t|\mathcal{F}_{t-j}) - \mathbb{E}(X_t|\mathcal{F}_{t-j-1}) \). It is straightforward to see that \( Z_{t,j} \) are martingale differences, and under certain conditions (mixing, physical dependence, your favourite dependence flavour etc) it can be shown that \( \sum_{j=0}^{\infty} \|Z_{t,j}\|_p < \infty \) (where \( \| \cdot \|_p \) is the \( p \)th moment). This means the above representation holds almost surely. Thus in several proofs we can replace \( X_t - \mu \) by \( \sum_{j=0}^{\infty} Z_{t,j} \). This decomposition allows us to use martingale theorems to prove results.

### 5.8 Kolmogorov’s formula (theorem)

Suppose \( \{X_t\} \) is a second order stationary time series. Kolmogorov’s(-Szegö) theorem is an expression for the error in the linear prediction of \( X_t \) given the infinite past \( X_{t-1}, X_{t-2}, \ldots \). It basically
states that
\[
E[X_n - X_n(1)]^2 = \exp\left(\frac{1}{2\pi} \int_0^{2\pi} \log f(\omega) d\omega\right),
\]
where \(f\) is the spectral density of the time series. Clearly from the definition we require that the spectral density function is bounded away from zero.

To prove this result we use (3.13);
\[
\text{var}[\hat{Y}] = \frac{\det(\Sigma)}{\det(S_{XX})},
\]
and Szegö’s theorem (see, Gray’s technical report, where the proof is given), which we state later on. Let \(P_{X_1,\ldots,X_n}(X_{n+1}) = \sum_{j=1}^n \phi_{j,n} X_{n+1-j}\) (best linear predictor of \(X_{n+1}\) given \(X_n,\ldots,X_1\)). Then we observe that since \(\{X_t\}\) is a second order stationary time series and using (3.13) we have
\[
E\left[X_{n+1} - \sum_{j=1}^n \phi_{n,j} X_{n+1-j}\right]^2 = \frac{\det(\Sigma_{n+1})}{\det(\Sigma_n)},
\]
where \(\Sigma_n = \{c(i-j); i, j = 0, \ldots, n-1\}\), and \(\Sigma_n\) is a non-singular matrix.

Szegö’s theorem is a general theorem concerning Toeplitz matrices. Define the sequence of Toeplitz matrices \(\Gamma_n = \{c(i-j); i, j = 0, \ldots, n-1\}\) and assume the Fourier transform
\[
f(\omega) = \sum_{j \in \mathbb{Z}} c(j) \exp(ij\omega)
\]
exists and is well defined (\(\sum_j |c(j)|^2 < \infty\)). Let \(\{\gamma_{j,n}\}\) denote the Eigenvalues corresponding to \(\Gamma_n\). Then for any function \(G\) we have
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^n G(\gamma_{j,n}) \to \int_0^{2\pi} G(f(\omega)) d\omega.
\]
To use this result we return to \(E[X_{n+1} - \sum_{j=1}^n \phi_{n,j} X_{n+1-j}]^2\) and take logarithms
\[
\log E[X_{n+1} - \sum_{j=1}^n \phi_{n,j} X_{n+1-j}]^2 = \log \det(\Sigma_{n+1}) - \log \det(\Sigma_n)
\]
\[
= \sum_{j=1}^{n+1} \log \gamma_{j,n+1} - \sum_{j=1}^n \log \gamma_{j,n}
\]
where the above is because \( \det \Sigma_n = \prod_{j=1}^{n} \gamma_{j,n} \) (where \( \gamma_{j,n} \) are the eigenvalues of \( \Sigma_n \)). Now we apply Szegö’s theorem using \( G(x) = \log(x) \), this states that

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \log(\gamma_{j,n}) \to \int_{0}^{2\pi} \log(f(\omega))d\omega.
\]

thus for large \( n \)

\[
\frac{1}{n+1} \sum_{j=1}^{n+1} \log \gamma_{j,n+1} \approx \frac{1}{n} \sum_{j=1}^{n} \log \gamma_{j,n}.
\]

This implies that

\[
\sum_{j=1}^{n+1} \log \gamma_{j,n+1} \approx \frac{n+1}{n} \sum_{j=1}^{n} \log \gamma_{j,n},
\]

hence

\[
\log E[X_{t+1} - \sum_{j=1}^{n} \phi_{n,j} X_{n+1-j}]^2 = \log \det(\Sigma_{n+1}) - \log \det(\Sigma_n)
\]

\[
= \sum_{j=1}^{n+1} \log \gamma_{j,n+1} - \sum_{j=1}^{n} \log \gamma_{j,n}
\]

\[
\approx \frac{n+1}{n} \sum_{j=1}^{n} \log \gamma_{j,n} - \sum_{j=1}^{n} \log \gamma_{j,n} = \frac{1}{n} \sum_{j=1}^{n} \log \gamma_{j,n}.
\]

Thus

\[
\lim_{n \to \infty} \log E[X_{t+1} - \sum_{j=1}^{n} \phi_{n,j} X_{t+1-j}]^2 = \lim_{n \to \infty} \log E[X_{n+1} - \sum_{j=1}^{n} \phi_{n,j} X_{n+1-j}]^2
\]

\[
= \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \log \gamma_{j,n} = \int_{0}^{2\pi} \log(f(\omega))d\omega.
\]

and

\[
\lim_{n \to \infty} E[X_{t+1} - \sum_{j=1}^{n} \phi_{n,j} X_{t+1-j}]^2 = \exp \left( \int_{0}^{2\pi} \log(f(\omega))d\omega \right).
\]

This gives a rough outline of the proof. The precise proof can be found in Gray’s technical report. There exists alternative proofs (given by Kolmogorov), see Brockwell and Davis (1998), Chapter 5.
This is the reason that in many papers the assumption

$$\int_0^{2\pi} \log f(\omega)d\omega > -\infty$$

is made. This assumption essentially ensures $X_t \notin \mathcal{X}_{-\infty}$.

**Example 5.8.1** Consider the AR($p$) process $X_t = \phi X_{t-1} + \varepsilon_t$ (assume wlog that $|\phi| < 1$) where $E[\varepsilon_t] = 0$ and var[$\varepsilon_t$] = $\sigma^2$. We know that $X_t(1) = \phi X_t$ and

$$E[X_{t+1} - X_t(1)]^2 = \sigma^2.$$

We now show that

$$\exp \left( \frac{1}{2\pi} \int_0^{2\pi} \log f(\omega)d\omega \right) = \sigma^2. \quad (5.39)$$

We recall that the spectral density of the AR(1) is

$$f(\omega) = \frac{\sigma^2}{|1 - \phi e^{i\omega}|^2},$$

$$\Rightarrow \log f(\omega) = \log \sigma^2 - \log |1 - \phi e^{i\omega}|^2.$$

Thus

$$\frac{1}{2\pi} \int_0^{2\pi} \log f(\omega)d\omega = \frac{1}{2\pi} \int_0^{2\pi} \log \sigma^2 d\omega - \frac{1}{2\pi} \int_0^{2\pi} \log |1 - \phi e^{i\omega}|^2 d\omega.$$

There are various ways to prove that the second term is zero. Probably the simplest is to use basic results in complex analysis. By making a change of variables $z = e^{i\omega}$ we have

$$\frac{1}{2\pi} \int_0^{2\pi} \log |1 - \phi e^{i\omega}|^2 d\omega = \frac{1}{2\pi} \int_0^{2\pi} \log(1 - \phi e^{i\omega})d\omega + \frac{1}{2\pi} \int_0^{2\pi} \log(1 - \phi e^{-i\omega})d\omega$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \sum_{j=1}^{\infty} \left[ \frac{\phi^j e^{ij\omega}}{j} + \frac{\phi^j e^{-ij\omega}}{j} \right] d\omega = 0.$$

From this we immediately prove (5.39).