Time Series Analysis

Course notes for STAT 479

Adam B Kashlak
Mathematical & Statistical Sciences
University of Alberta
Edmonton, Canada, T6G 2G1

April 20, 2021

Contents

Preface

> The understanding's in your mind. You only have to find it. But, time—Time, the creature said, is the simplest thing there is.
>
> *Time is the Simplest Thing*
> Clifford D Simak (1961)

The following are lecture notes originally produced for an undergraduate course on time series at the University of Alberta in the winter of 2020. The aim of these notes is is to introduce the main topics, applications, and mathematical underpinnings of time series analysis.

These notes were produced by consolidating two main sources being the textbook of Shumway and Stoffer, Time Series Analysis and Its Applications, and the past course notes produced by Dr. Doug Weins also at the University of Alberta.

Adam B Kashlak
Edmonton, Canada
January 2020

Chapter 1

Time Series: Overview

Introduction

In this chapter we consider different types of time series processes that we may encounter in practice. The main difference between time series and other areas of statistics like linear regression is that the noise or errors can be correlated. This arises from the fact that time implies causality; the past predicts the future. Thus, we no longer live in the independent and identically distributed setting of most other areas of statistics.

This chapter also reintroduces notions of covariance and correlation in the context of time series, which become autocovariance and autocorrelation. The critical property of stationarity is defined, which allows us to estimate such autocovariances and autocorrelations from a given time series dataset.

1.1 Types of Noise

When one is first introduced to the realm of statistics, the data on hand is treated as independent and identically distributed observations from some population. That is, the noise or errors or randomness present in the data is treated as a collection of iid random variables—typically mean zero Gaussians. Times series data breaks from the iid setting as causality becomes a key notion: the effects of random errors in the past are present in future observations as well. Thus, we consider many types of noise that can occur in real data.

1.1.1 White Noise

Let w_t denote the white noise process. This is a random variable indexed by time t such that

$$\mathrm{E}w_t = 0 \text{ and } \mathrm{Var}\,(w_t) = \sigma^2 \ \forall t \in [0, T], \text{ and } \mathrm{cov}\,(w_t, w_s) = 0 \ \forall t \neq s.$$

That is, w_t and w_s are uncorrelated but not necessarily independent.

This can be strengthened to iid noise if *uncorrelated* is replaced with *independent*. This can be further strengthened to Gaussian white noise were every $w_t \sim \mathcal{N}\left(0, \sigma^2\right)$. The intuition behind the term white noise comes from signals processing where a signal is *white* if it contains all possible frequencies. Furthermore, the white noise process will be used to generate all of the subsequent processes.

1.1.2 Autoregressive

The autoregressive (AR) process is a natural way to encode causality into the white noise process, that is, demonstrate how the past influences the future. The general formula is

$$X_t = \sum_{i=1}^{p} \theta_i X_{t-i} + w_t,$$

which is that past time observation X_{t-i} contributions $\theta_i \in \mathbb{R}$ to the present time observation X_t.

For example, if $p = 1$ and $\theta_1 = 1$, then we have the process

$$X_t = X_{t-1} + w_t,$$

which is also an example of a *Markov process* and a *Martingale* to be discussed below. This process could model, say, the price of a commodity where the previous price X_{t-1} is the best guess for the current price X_t plus or minus some noise w_t.

The AR process with $p = 1$ and $\theta_1 = 1$ can be thought of as a random walk. An interesting and useful extension of this process is the random walk with drift, which is

$$X_t = a + X_{t-1} + w_t$$

for some $a \neq 0$. In this case, the process increases (or decreases) by the fixed amount a at each time step. But there is also the addition of white noise w_t at each step. Hence, one could try to estimate the drift term a from historical data in order to ascertain if X_t (say the price of some commodity) is increasing or decreasing or remaining constant up to random noise.

Example 1.1.1. *In Figure 1.1, we have four examples of autoregressive processes where w_t is Gaussian white noise.*

1. *The white noise process:* $X_t = w_t$.
2. *Random Walk:* $X_t = X_{t-1} + w_t$.
3. *AR(2):* $X_t = X_{t-1} - 0.2X_{t-2} + w_t$
4. *AR(3):* $X_t = X_{t-1} - 0.2X_{t-2} + 0.18X_{t-3} + w_t$

Based solely on the plots, the random walk and the chosen AR(3) process look similar. Likewise, it may be hard to immediately identify that the top left plot is white noise while the bottom left is an AR(2) process.

1.1.3 Moving Average

The moving average (MA) process is a *smoother* type of noise than the white noise process. It can be expressed by the formula

$$X_t = \sum_{j=1}^{q} \phi_j w_{t-j} + w_t$$

for $\phi_j \in \mathbb{R}$. Compared to the above AR formula, the MA formula averages over the noise terms w_t as opposed to the observed values X_t. It can be thought of as ripple effects in the process. That is, if there is a shock to the process w_{t-1}, then it's effects are still felt at time t by the term $\phi_1 w_{t-1}$.

Alternatively, this model can be written as an overall average like

$$X_t = \sum_{j=-q/2}^{q/2} \phi_j w_{t+j}.$$

In this way, we can consider the MA process as a weighted averaged white noise process. A simple example is $X_t = (w_{t-1} + w_t + w_{t+1})/3$.

1.1.4 Markov Processes and Martingales

Three very useful tools in probability theory that have been extensively studied are Markov processes, martingales, and the Gaussian process. We briefly introduce them here noting that there has be extensive research on each topic.[1]

A Markov process is one where conditioning on the entire history of the process is equivalent to just conditioning on the most recent time point. More precisely, X_t is Markov if

$$\mathrm{E}\left(X_t | X_{t-1}, \ldots, X_1\right) = \mathrm{E}\left(X_t | X_{t-1}\right).$$

For example, the AR process with $p = 1$ is Markov as the present value X_t only depends on the past via X_{t-1} and no other X_{t-i}. Note that people have also studied order p Markov processes where the present depends on the more recent p time points, but the above definition is the most commonly used.

A martingale is often defined as a fair game being one where the expected winnings/looses is zero. That is, it is a stochastic process where the conditional expectation is equal to the most recent observation, i.e.

$$\mathrm{E}\left(X_t | X_{t-1}, \ldots, X_1\right) = X_{t-1}.$$

[1] See, for example, the two volume series *Diffusions, Markov Processes, and Martingales* by Rogers and Williams.

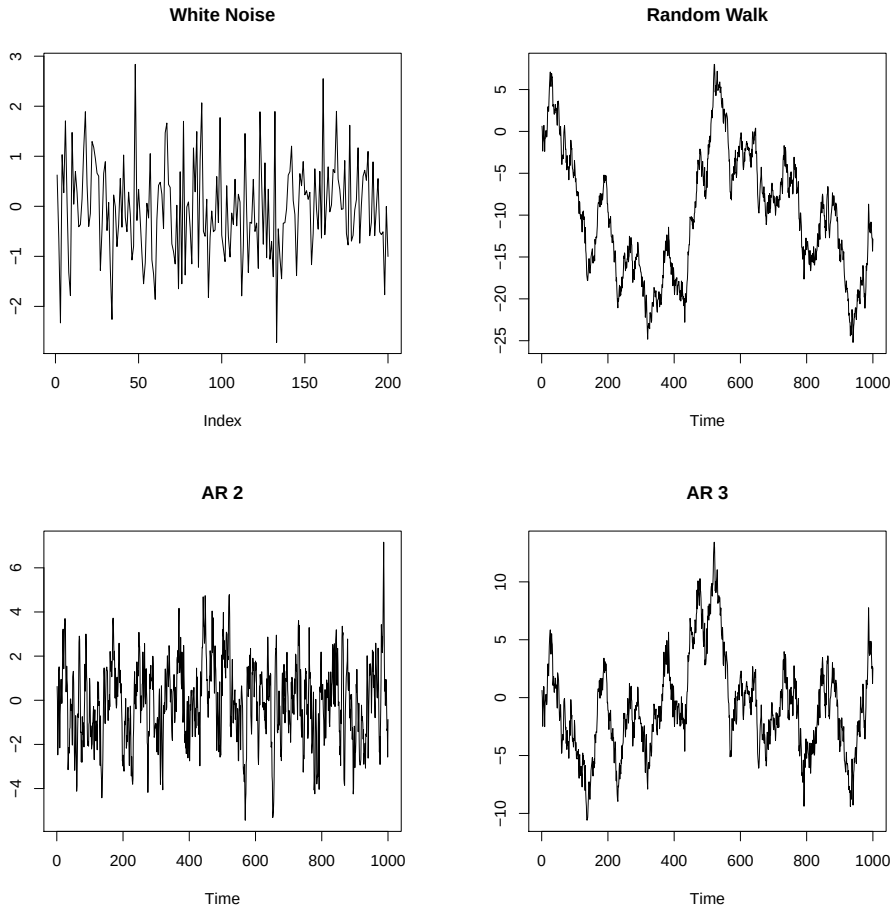

Figure 1.1: Examples of the white noise process and autoregressive versions of that white noise process of order 1, 2, and 3.

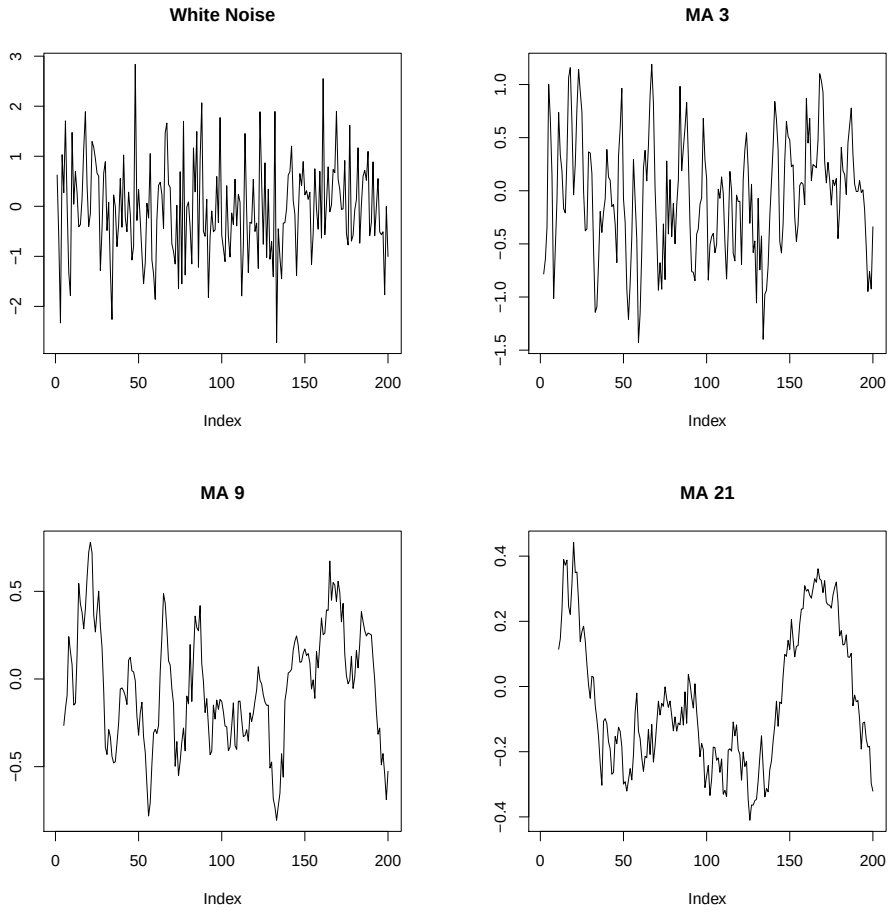

Figure 1.2: Examples of the white noise process and moving averaged versions of that white noise process averaged over windows of length 3, 9, and 21.

The AR process with $p = 1$ and $\theta_1 = 1$ is an example of a martingale. Note that supermartingales and submartingales have also been studied where the above $=$ is replaced by a $\leq$ or $\geq$, respectively.

As the normal distribution lends itself elegantly to other areas of statistics, so does it to time series. The Gaussian process is a generalization of the multivariate normal distribution. It is a stochastic process X_t where for any finite collection of time points $\{t_1, \ldots, t_k\}$, the random vector $(X_{t_1}, \ldots, X_{t_k})$ is multivariate normal. Much like the multivariate normal distribution, the Gaussian process can be defined by its mean μ_t and covariance $C_{s,t}$ where

$$\mu_t = \mathrm{E}X_t \text{ and } \Sigma_{s,t} = \operatorname{cov}(X_s, X_t).$$

Many time series fall under the category of *linear processes*. Given a white noise process w_t, a linear process is defined as

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \theta_j w_{t-j},$$

which is that every X_t is a linear combination of the terms in the white noise process with some mean μ added on. Here, we require $\sum \theta_i^2 < \infty$ in order for the process to have a finite variance. However, as we are generally interested in modelling casual processes in time—i.e. the past predicts the future and not vice versa—we can instead consider the more restricted definition

$$X_t = \mu + \sum_{j=0}^{\infty} \theta_j w_{t-j}.$$

Example 1.1.2 (The AR(1) Process)**.** *We revisit the AR(1) process, $X_t = \theta X_{t-1} + w_t$, and by using this recursive definition and assuming we can extend the series infinitely into the past, we can rewrite it as*

$$X_t = \theta(\theta X_{t-2} + w_{t-1}) + w_t = \ldots = \sum_{j=0}^{\infty} \theta^j w_{t-j}.$$

Infinite series are limits—i.e. $\sum_{j=0}^{\infty} \theta^j w_{t-j} = \lim_{N\to\infty} \sum_{j=0}^{N} \theta^j w_{t-j}$. Hence, this sum may not converge in any meaningful way. Let $S_N(\theta) = \sum_{j=0}^{N} \theta^j w_{t-j}$, then

$$\mathrm{E}S_N = 0 \text{ and } \operatorname{Var}(S_N) = \sigma^2 \sum_{j=0}^{N} \theta^{2j} = \sigma^2 \left(\frac{1 - \theta^{2N+2}}{1 - \theta^2} \right).$$

Thus, if $|\theta| < 1$, then $\operatorname{Var}(S_N(\theta)) \to \sigma^2/(1-\theta^2)$, and if w_t is Gaussian noise, then

$$S_N(\theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2/(1-\theta^2)\right).$$

In the case of the random walk, which is $\theta = 1$, the series does not converge, but by the central limit theorem, we have

$$n^{-1/2} S_N(1) \xrightarrow{d} \mathcal{N}(0, 1).$$

1.2 Properties of Times Series

1.2.1 Autocovariance

As a time series X_t can be thought of as a single entity, the covariance between two time points is referred to as the *autocovariance* and is defined as

$$K_X(s,t) = cov(X_t, X_s).$$

The notation K comes from treating the autocovariance as a kernel function for an integral transform.[2] Note that the autocovariance function is symmetric, $K(s,t) = K(t,s)$, and positive (semi) definite in the sense that for any finite collection of time points $\{t_1, \ldots, t_k\}$, we have a $k \times k$ matrix with i,jth entry $K(t_i, t_j)$ and this matrix is positive (semi) definite.

Similar to the multivariate setting, we can normalize the autocovariance into an autocorrelation by

$$\rho(s,t) = \frac{K_X(s,t)}{\sqrt{K_X(s,s)K_X(t,t)}}.$$

Example 1.2.1 (AR(1) with drift)**.** *For w_t a white noise process with variance σ^2, consider the AR(1) with drift process*

$$X_t = a + \theta X_{t-1} + w_t$$

for some real a and θ. We can use the recursive definition to get that

$$\begin{aligned} X_t &= a + \theta(a + \theta X_{t-2} + w_{t-1}) + w_t \\ &= (1+\theta)a + \theta^2 X_{t-2} + \theta w_{t-1} + w_t. \end{aligned}$$

This can be repeated m times to get

$$X_t = a\sum_{j=0}^{m} \theta^j + \theta^{m+1} X_{t-m} + \sum_{j=0}^{m} \theta^j w_{t-j}.$$

Then, assuming this process has an infinite past and that $|\theta| < 1$, we can take m to infinity to get

$$X_t = \frac{a}{1-\theta} + \sum_{j=0}^{\infty} \theta^j w_{t-j},$$

which happens to be a linear process. The mean can now be quickly calculated to be

[2] For example, $g(s) = \int f(t)K(s,t)dt$.

$\mathrm{E}X_t = a/(1-\theta)$ *as* $\mathrm{E}w_t = 0$ *for all* t*. Furthermore, the autocovariance is*

$$\begin{aligned} K_X(s,t) &= \operatorname{cov}(X_s, X_t) \\ &= \operatorname{cov}\left(\sum_{j=0}^{\infty}\theta^j w_{s-j}, \sum_{j=0}^{\infty}\theta^j w_{t-j}\right) = \mathrm{E}\left(\sum_{j,i=0}^{\infty}\theta^j w_{s-j}\theta^i w_{t-i}\right) \\ &= \sigma^2\sum_{j,i=0}^{\infty}\theta^{i+j}\mathbf{1}[s-j=t-i] = \sigma^2\theta^{|s-t|}\sum_{j=0}^{\infty}\theta^{2j} = \frac{\sigma^2\theta^{|s-t|}}{1-\theta^2}. \end{aligned}$$

This implies that the variance is $\sigma^2/(1-\theta^2)$*. Note that this process is a weakly stationary process, which will be defined below.*

1.2.2 Cross-Covariance

The cross covariance is similar to the auto covariance, but applies to multivariate time series. More simply, if we have two time series X_t and Y_t, then we can consider

$$K_{XY}(t,s) = \operatorname{cov}(X_t, Y_s)$$

and the cross-correlation

$$\rho_{XY}(t,s) = \frac{K_{XY}(t,s)}{K_X(t,t)K_Y(s,s)}.$$

1.2.3 Stationarity

In a broad sense, stationarity implies some property of the time series is invariant to shifts in time. There are two such notions we will consider.

Definition 1.2.2 (Weak Stationarity)**.** *A process* X_t *is said to be weakly stationary if its mean and autocovariance are invariant to time shifts. That is, for any* $r > 0$*,*

$$\begin{aligned} (\textit{Mean}): &\ \mathrm{E}X_t = \mathrm{E}X_{t+r} = \mu \\ (\textit{Autocovariance}): &\ K_X(t,s) = K_X(t+r, s+r) \end{aligned}$$

Definition 1.2.3 (Strong Stationarity)**.** *A process* X_t *is said to be strongly stationary if its joint distribution function is invariant to time shifts. That is, for any* $r > 0$*, and any finite collection of time points* $t_1, \ldots, t_k$*,*

$$F(X_{t_1}, \ldots, X_{t_k}) = F(X_{t_1+r}, \ldots, X_{t_k+r})$$

where $F()$ *is the joint CDF of the* k *random variables. That is,* $F(X_{t_1}, \ldots, X_{t_k}) = \mathrm{P}(X_{t_1} \le x_{t_1}, \ldots, X_{t_k} \le x_{t_k})$*.*

For a weakly (and thus for a strongly) stationary process X_t, we have that the autocovariance function is

$$K_X(s,t) = K_X(s-t,0) = K_X(\tau)$$

for some τ being the difference between two time points (the time lag). Hence, if a process is weakly stationary, the autocovariance can be treated as a univariate function.

Furthermore, this univariate function is both symmetric and bounded in τ. This can be seen by noting, for symmetry, that

$$\begin{aligned} K_X(\tau) = K_X((\tau+t)-t) = K_X(\tau+t,t) \\ = K_X(t,\tau+t) = K_X(t-(\tau+t)) = K_X(-\tau), \end{aligned}$$

and, for boundedness, that $K_X(0) = \mathrm{Var}\,(X_t)$ for all t and by applying the Cauchy-Schwarz inequality that, for any r,

$$K_X(0)^2 = \mathrm{Var}\,(X_t)\,\mathrm{Var}\,(X_{t+r}) \geq \mathrm{cov}\,(X_t, X_{t+r})^2 = K_X(r)^2.$$

This implies that $|K_X(\tau)| \leq K_X(0)$.

Stationarity also helps in the statistical context with estimation. Specifically, we cannot estimate the autocovariance for a single non-stationary series, but can estimate it for a stationary series. With that in mind, often we can modify a time series to make it stationary.

Example 1.2.4. *Consider the time series*

$$X_t = a + bt + Y_t$$

where Y_t is a mean zero stationary process. The mean $\mu_t = a + bt$ is a function of t. However, subtracting off this linear trend leaves

$$X_t - a - bt = Y_t,$$

which is stationary.

As with cross-covariance, we can consider joint stationarity of two series X_t and Y_t when dealing with multiple time series at once.

Definition 1.2.5 (Joint Stationarity)**.** *The processes X_t and Y_t are said to be jointly stationary if both are individually stationary and also if the cross covariance function is also stationary—i.e. $K_{XY}(t,s) = K_{XY}(t+r, s+r)$ for any r.*

1.3 Estimation

Estimating parameters in a time series model is harder than it is in standard statistics where we often assume that the observations are iid. Now, we are faced with a single sequence of points $X_1, X_2, \ldots, X_T$ which are not iid. To estimate the mean and autocovariance we require the process to be weakly stationary. If it isn't, then, for example, every X_t will have its own mean and we cannot estimate it.

Note as indicated above, we will consider time series with T total observations observed at equally spaced intervals. If the observations are irregularly spaced, more work has to be done.

1.3.1 Estimating the mean

For a weakly stationary process, $\mathrm{E}X_t = \mu$ for all $t = 1, \ldots, T$, so we can consider the usual sample average, $\bar{X} = T^{-1} \sum_{t=1}^{T} X_t$, as an estimator for μ. Specifically, due to the linearity of expectation, we have $\mathrm{E}\bar{X} = \mu$. However, as these X_t are not uncorrelated, the variance calculation is a bit more involved than usual.

$$\begin{aligned}
\mathrm{Var}\left(\bar{X}\right) &= \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} X_t\right) \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathrm{cov}\left(X_t, X_s\right) \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} K_X(|t-s|) \\
&= \frac{1}{T}K_X(0) + \frac{2}{T}\sum_{z=1}^{T-1}\left(1 - \frac{z}{T}\right)K_X(z).
\end{aligned}$$

Note that in the uncorrelated case $K_X(0) = \sigma^2$ and $K_X(z) = 0$ for $z > 0$. Thus, the formula reduces to the usual σ^2/T.

1.3.2 Estimating the autocovariance

For a weakly stationary process, we can define the sample autocovariance function to be

$$\hat{K}_X(h) = \frac{1}{T}\sum_{t=1}^{T-h}(X_{t+h} - \bar{X})(X_t - \bar{X}).$$

As h gets bigger, the number of terms in the sum decreases giving less accurate estimation. Similarly, the sample autocorrelation function is defined to be

$$\hat{\rho}(h) = \hat{K}_X(h)/\hat{K}(0),$$

and the sample cross covariance and cross correlation are

$$\hat{K}_{XY}(h) = \frac{1}{T}\sum_{t=1}^{T-h}(X_{t+h} - \bar{X})(Y_t - \bar{Y}) \text{ and } \hat{\rho}_{XY}(h) = \frac{\hat{K}_{XY}(h)}{\sqrt{\hat{K}_X(0)\hat{K}_Y(0)}}.$$

Examples of estimated autocovariance functions are in Figure 1.3 for the white noise process, the random walk, the moving average process with a window of length 9, and the autoregressive process $X_t = X_{t-1} - 0.2X_{t-2} + w_t$. However, we note that the random walk and this AR(2) process are not stationary. Thus, even though we can compute the autocovariance in R with the `acf` function, we must consider its validity.

The sample autocovariance is defined in such a way to make the function positive semi-definite. This ensures that estimates of variances will never be negative as, for any real vector $a = (a1, \ldots, a_T) \in \mathbb{R}^T$, the variance estimate for $a \cdot X = \sum_{t=1}^{T} a_t X_t$ is

$$\sum_{t=1}^{T}\sum_{s=1}^{T} a_t a_s \hat{K}(|t-s|),$$

which must be non-negative.

1.3.3 Detecting White Noise

A main goal of time series analysis is to transform the data into a white noise process. That is, we aim to identify trends and patterns in the process. Once those have been removed, what remains is random noise. Hence, we need to determine if a process has been transformed into a white noise process.

One way to do this is to look at the estimated autocorrelation function for a given time series as displayed in Figure 1.3. Note that for the white noise process, we see a spike at lag 0 referring to an estimate of the variance of the process $\hat{K}_X(0)$, which, in this case, is 1. At the remaining lags, $\hat{K}_X(h)$ is *small* for $h \neq 0$. The question is what does *small* mean?

In the plots of Figure 1.3, R includes blue dashed lines at the value $2/\sqrt{T}$. This is because for w_t iid mean zero white noise with variance $\mathrm{E}w_t^2 = \sigma^2$ and finite fourth moment, $\mathrm{E}w_t^4 < \infty$, we have a sample autocovariance of approximately

$$\hat{K}_w(h) \approx \frac{1}{T}\sum_{t=1}^{T-h} w_{t+h} w_t.$$

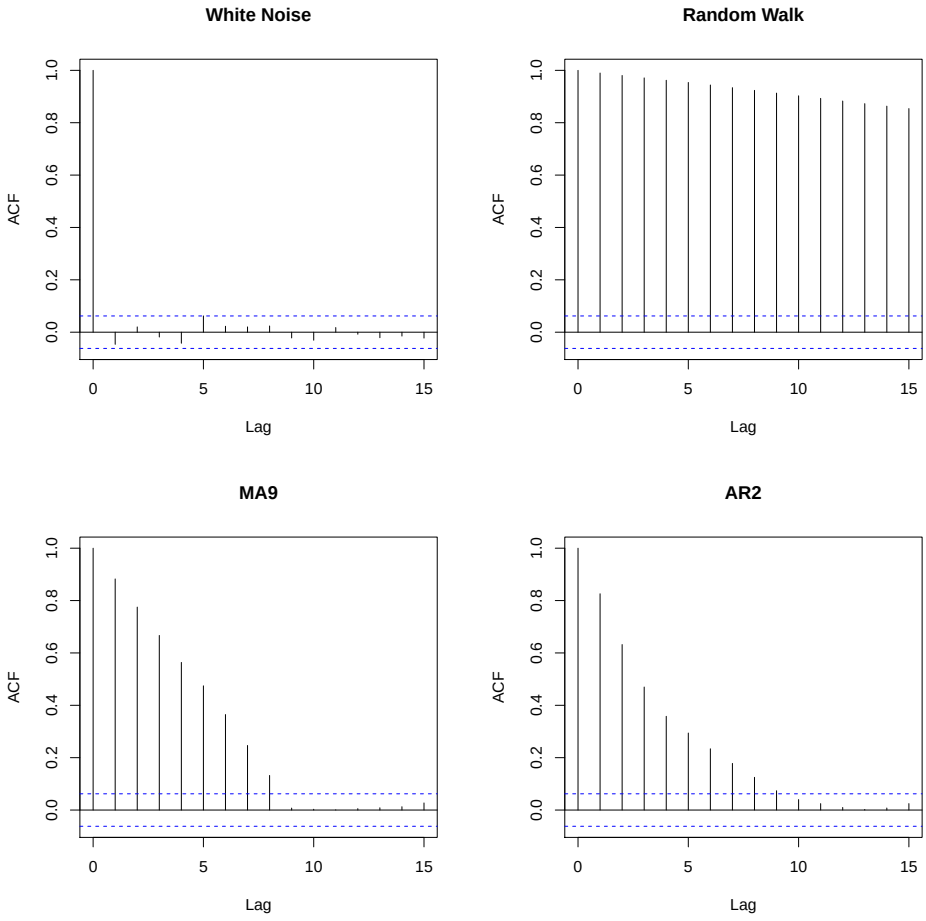

Figure 1.3: Estimated autocorrelations from the processes from Figures 1.1 and 1.2 using the `acf` function in R.

For $h \neq 0$, this has zero mean as $\mathrm{E}\,(w_t w_s) = 0$ for $s \neq t$, and has a variance of

$$\frac{1}{T^2}\mathrm{Var}\left(\sum_{t=1}^{T-h} w_{t+h}w_t\right) = \frac{1}{T^2}\sum_{s,t=1}^{T-h}\mathrm{E}[w_{s+h}w_s w_{t+h}w_t] =$$

$$\frac{1}{T^2}\sum_{t=1}^{T-h}\mathrm{E}[w_{t+h}^2 w_t^2] = \frac{T-h}{T^2}\sigma^4 \approx \frac{\sigma^4}{T}$$

for large T. Thus, the variance of the autocorrelation is approximately $1/T$, and we would expect the values $\hat{\rho}_X(h)$ to be within two standard deviations from the mean, or $\pm 2/\sqrt{T}$, when X_t is a white noise process. Hence, we can examine the autocorrelations to determine whether or not the process looks like white noise.

Chapter 2

Statistical Models for Time Series

Introduction

Two main goals that traverse much of statistics are fitting models to data and using such models for prediction or forecasting. In this chapter, we consider classic linear regression (briefly) to discuss its uses and its shortcomings with respect to time series data. Then, we discuss more sophisticated models for time series of the ARIMA family of models.

As mentioned in the previous chapter, we need to transform a time series X_t into a time series Y_t that is stationary. This is because stationarity allows us to do estimation from a single series. The hard part it to determine how to extract such a stationary Y_t from the series X_t. Before continuing, we need some definitions.

Definition 2.0.1 (Backshift Operator)**.** *The backshift operator B acts on time series by $BX_t = X_{t-1}$. This can be iterated to get $B^k X_t = X_{t-k}$. This can also be inverted to get the forward shift operator $B^{-1}X_t = X_{t+1}$. Thus, $B^{-1}B = BB^{-1} = I$ the identity operator.*

Definition 2.0.2 (Difference Operator)**.** *The difference operator ∇ acts on time series by $\nabla X_t = X_t - X_{t-1}$. Note that $\nabla X_t = (1-B)X_t$. This operator can also be iterated to get*

$$\nabla^k X_t = (1-B)^k X_t = \sum_{i=0}^{k} (-1)^i \binom{k}{i} X_{t-i}.$$

For example, the second difference operator is $\nabla^2 X_t = X_t - 2X_{t-1} + X_{t-2}$.

Differencing a series can be used to remove periodic trends. Fun fact: using the gamma function, we can also consider fractional differencing for $\kappa \in \mathbb{R}^+$, which is

$$\nabla^\kappa X_t = \sum_{i=0}^{\kappa} (-1)^i \frac{\Gamma(\kappa+1)}{i!\Gamma(\kappa-i+1)} X_{t-i}.$$

2.1 Regression

2.1.1 Linear Regression in Brief

In linear regression, we consider modelling a response variable X_t by some independent variables or predictors $z_{t,1}, \ldots, z_{t,p}$ as

$$X_t = \beta_0 + \beta_1 z_{t,1} + \ldots + \beta_p z_{t,p} + \varepsilon_t$$

where the β_i are unknown fixed parameters and the ε_t are iid mean zero uncorrelated errors. Written in vector-matrix form, we have $X = Z\beta + \varepsilon$ where $X \in \mathbb{R}^T$ and $Z \in \mathbb{R}^{T\times(p+1)}$. Given this setup, the Gauss–Markov theorem tells us that the least squares estimator, $\hat{\beta} = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}X$, is the minimal variance unbiased estimator for β.

2.1.2 Linear Regression for Time Series

The big challenge for time series is that the error process ε_t may not be uncorrelated white noise, but in fact, a correlated process. Thus, we consider the time series model

$$X_t = \mu_t + Y_t \tag{2.1.1}$$

where μ_t is a deterministic process such as $\mu_t = \beta_0 + \beta_1 z_{t,1} + \ldots + \beta_p z_{t,p}$ and where Y_t is a stationary stochastic process. Starting with an observed series X_t, this leaves us with the goal of identifying the deterministic *trend* μ_t and the stochastic piece Y_t.

When diagnosing the fit of the linear regression model to the data, we often consider the vector residuals being $r = X - \hat{\beta}Z$. Plotting fitted values against the residuals, we would expect in linear regression to have residuals that look like random noise. However, with time series data, the residuals will often not be white noise but some other stochastic process.

Example 2.1.1 (Prescription Price Data)**.** *The* `TSA` *package in* `R` *contains a dataset called* `prescrip`*, which charts monthly US average prescription drug costs for 1986 to 1992. Looking at the time series in Figure 2.1, we notice that there is a steady increase over the time span of the data. A quadratic regression model to this data yielded a fitted model*

$$(\text{cost}) = 505000 + 510.5(\text{time}) - 0.13(\text{time})^2$$

where (time) *ranges from 1986.6 to 1992.17. The F-statistic of 1612 with degrees of freedom* $(2, 65)$ *is very significant (p-value* $< 2 \times 10^{-16}$*), and the* `R` *summary for the fitted model is*

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*505000*	*1.344e+05*	*3.757*	*0.0004*
time	*-510.5*	*1.351e+02*	*-3.778*	*0.0003*
time^2	*0.129*	*3.396e-02*	*3.799*	*0.0003*

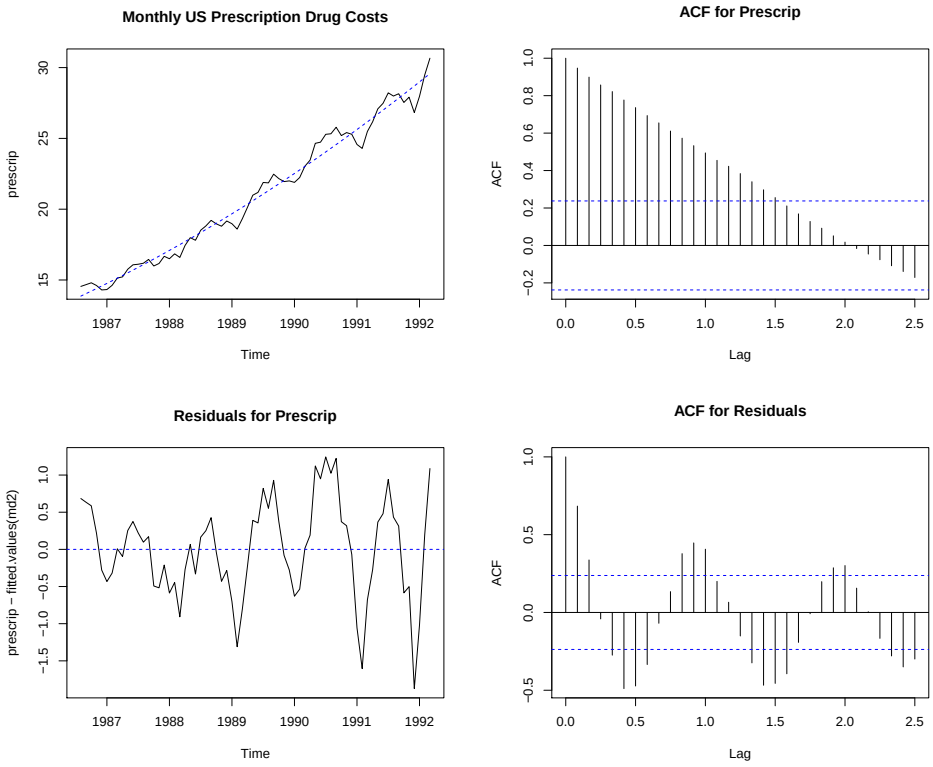

Figure 2.1: Plotted time series of prescription drug costs with fitted regression line (top left). The residuals for the series plotted (bottom left). The estimated autocorrelation functions for the original series (top right) and the residual series (bottom right).

By removing this trend and focusing on the residuals, we see the bottom two plots from Figure 2.1. Here, the estimated autocorrelation does not seem as extreme. However, there are still some periodic patterns that emerge and that will need to be dealt with.

To the residual process, we can fit a linear model using sines and cosines, which gives the fitted model

$$(\text{residuals}) = 0.2 - 0.09\sin(2\pi(\text{time})) - 0.65\cos(2\pi(\text{time})).$$

The F-statistic of 33.36 with degrees of freedom $(2, 65)$ *is also very significant (p-value* $\approx 10^{-10}$*), and the* `R` *summary for the fitted model is as follows.*

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*0.02*	*0.06*	*0.34*	*0.732*
sin	*-0.09*	*0.08*	*-1.07*	*0.290*
cos	*-0.65*	*0.08*	*-8.07*	*2.24e-11*

In Figure 2.2, we have the residuals for this trigonometric model as well as the estimated autocorrelation. In the ACF plot, we see a large spike at lag = 0.1. Hence, considering the first difference operator applied to this process, we have the bottom two plots in Figure 2.2. Now the estimated autocorrelation is looking much more like white noise.

Regression with time series on time series

Note also that we can consider regression of one series with respect to another. For example, given two time series X_t and Y_t, we could fit a model

$$X_t = \hat{\beta}_0 + \hat{\beta}_1 Y_t.$$

We could also consider fitting a model with a fixed lag h

$$X_t = \hat{\beta}_0 + \hat{\beta}_1 Y_{t-h}.$$

As a hypothetical example, X_t could be monthly rainfall and Y_{t-1} could be the average temperature of the previous month.

2.2 Smoothing

Smoothing methods are a large class of statistical tools that can be applied to noisy data. While much theoretical work has gone into understanding these methods, we will just present the main idea briefly.

Moving Average Smoothing

For a time series X_t for $t = 1, \ldots, T$ and coefficients $\theta_{-r}, \ldots, \theta_0, \ldots, \theta_r$ with $\sum_{j=-r}^{r} \theta_j = 1$, we can define a new process

$$M_t = \sum_{j=-r}^{r} \theta_j X_{t-j}$$

for $t = r+1, \ldots, T-r$. The simplest example is to set $\theta_j = (2r+1)^{-1}$, which just compute the sample average with a window of length $2r+1$. This can be performed easily in `R` with the `filter` function.

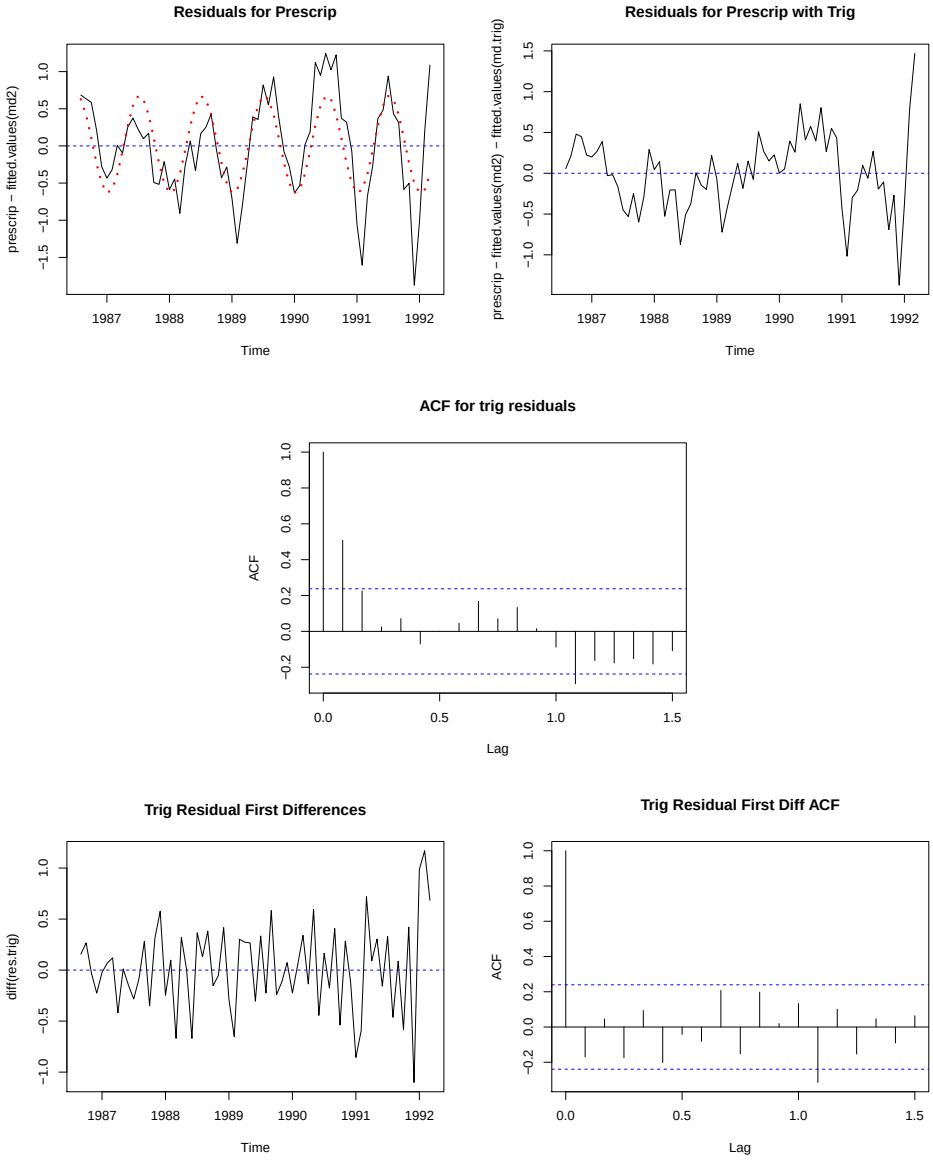

Figure 2.2: The trig model fitted to the residuals (top left). The residuals for the trig model (top right). The ACF for the trig residual model (middle). The first differences of the trig residuals (bottom left). The ACF for the first differences of the trig residuals (bottom right).

Kernel Smoothing

A kernel function $\kappa_h(x, x_0)$ is a non-negative function that is decreasing in $|x - x_0|$ and has a *bandwidth* parameter h. We also require that $\int_{\mathbb{R}} \kappa_h(x, x_0)dx < \infty$ for all $x_0 \in \mathbb{R}$. Examples include

Gaussian:	$\kappa_h(x, x_0) = \exp\left(-\lvert x - x_0\rvert^2/2h^2\right)$
Triangular:	$\kappa_h(x, x_0) = (1 - \lvert x - x_0\rvert/h)_+$
Epanechnikov:	$\kappa_h(x, x_0) = (1 - \lvert x - x_0\rvert^2/h^2)_+$

where the notation $(\ldots)_+$ means take the positive part and set the rest to zero.

In the time series context, we can use a kernel to construct weights to be used in the previously mentioned moving average smoother. Specifically, for some $r \in \mathbb{N}$, we can define θ_i for $i = -r, \ldots, 0, \ldots, r$ as

$$\theta_i = \kappa_h(i, 0) / \sum_{j=-r}^{r} \kappa_h(j, 0).$$

Kernel-based methods also occur in probability density estimation (the kernel density estimator) and in linear regression (Nadaraya-Watson) as well as others. Note further that kernel based estimators typically are biased estimators and as the bandwidth h increases, the bias increases while the variance decreases. As a result, much research has gone into bandwidth selection. This can be implemented in `R` via the `ksmooth` function.

Lowess

Lowess is an acronym for locally weighted scatterplot smoothing. This method combines nearest neighbours and weighted linear regression. Effectively, it takes a window of data points, say $X_{t-r}, \ldots, X_{t+r}$, and applies a low degree polynomial regression to it. This can be implemented in `R` by the function `lowess`. The `R` manual page states that lowess "is defined by a complex algorithm".

Cubic Spline Smoothing

As discussed in the context of linear regression, fitting polynomials to data can be a powerful tool. However, it typically unwise to fit one high degree polynomial to your data. Instead, spline models split the T data points into k pieces by defining the partition

$$1 = t_1 < t_2 < \ldots < t_{k+1} = T.$$

These points are referred to as the *knots*. Then, a separate polynomial—typically cubic but could have another degree instead—is fit to each subinterval of approximately T/k data points. That is, we fit a linear model

$$M_t^{(i)} = \beta_{i,0} + \beta_{i,1}t + \beta_{i,2}t^2 + \beta_{i,3}t^3.$$

to the ith interval. As a result, this can be written as a least squares estimation problem where the $\hat{M}_t$ are the fitted values that minimize

$$\sum_{t=1}^{T}(X_t - M_t)^2.$$

Here, if we assume $t = 1, \dots, T$ and $f_i(t)$ are spline basis functions for $i = 1, \dots, T$, then the design matrix for the regression is F with ijth entry is $F_{i,t} = f_i(t)$. Thus, we can write $\hat{M} = F\hat{\beta} = F(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}X$.

If we want to upgrade our spline model into a *smoothing* spline we fit the same polynomial but with a penalty term to enforce more smoothing. That is, we would find the $\hat{M}_t$ that minimizes

$$\sum_{t=1}^{T}(X_t - M_t)^2 + \lambda \int (M_s'')^2 ds$$

where $\lambda \geq 0$ is a smoothing parameter. Note that this minimization problem is taken over all possible twice continuously differentiable functions M_t, which is referred to as the Sobolev space H^2. When $\lambda = 0$, so smoothing is applied and we simply have a least squares estimator for our data. Taking $\lambda \to \infty$ imposes that the second derivative of M_t must be zero. Hence, we have a straight line fit to our data in the limit.

Using the notation from before, we can also work out an explicit solution reminiscent of ridge regression.[1] From the penalty term, we can define the matrix Ω to have entries

$$\Omega_{i,j} = \int f_i''(s) f_j''(s) ds.$$

Then, we have

$$\int (M_s'')^2 ds = \beta^{\mathrm{T}}\Omega\beta,$$

and consequently, we are finding the coefficients $\hat{\beta}$ that minimize

$$\|X_t - F\beta\|_2^2 + \lambda\beta^{\mathrm{T}}\Omega\beta,$$

which is $\hat{\beta} = (F^{\mathrm{T}}F + \lambda\Omega)^{-1}F^{\mathrm{T}}X$.

[1] Recall that the ridge estimator for the model $Y = X\beta + \varepsilon$ is $\hat{\beta} = (X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}Y$.

2.3 ARIMA Models for Times Series

In the previous section, we considered regression models like

$$X_t = f(t) + w_t$$

where $f(t)$ is deterministic and w_t is white noise. The goal was to estimate the deterministic piece by some $\hat{f}(t)$. But such an approach cannot handle time series models like the AR(1) process $X_t = X_{t-1} + w_t$. In this section, we take a closer look at fitting such time series models to observed data.

2.3.1 Autoregressive Processes

We reintroduce the autoregressive process formally as

Definition 2.3.1 (Autoregressive Process)**.** *The time series X_t is an $AR(p)$ process if X_t has zero mean and if we can write it as*

$$X_t = w_t + \sum_{i=1}^{p} \phi_i X_{t-i}$$

where w_t is white noise with variance σ^2 and $\phi_1, \ldots, \phi_p \in \mathbb{R}$ are constants with $\phi_p \neq 0$. Using the backshift operator B, we can write the $AR(p)$ process as $\Phi(B)X_t = w_t$ where

$$\Phi(B) = \left(1 - \sum_{i=1}^{p} \phi_i B^i\right).$$

We will refer to $\Phi(B)$ as the **autoregressive operator**.

Note that we choose X_t to have zero mean for convenience. If X_t has mean $\mu \neq 0$, then we can rewrite $X_t = \tilde{X}_t + \mu$ where $\tilde{X}_t$ is mean zero to get

$$\begin{aligned}
\tilde{X}_t + \mu &= w_t + \sum_{i=1}^{p} \phi_i (\tilde{X}_t + \mu) \\
\tilde{X}_t &= -\mu \left(1 - \sum_{i=1}^{p} \phi_i\right) + w_t + \sum_{i=1}^{p} \phi_i \tilde{X}_{t-i}.
\end{aligned}$$

That is, we can rewrite X_t as a mean zero AR(p) process with an added constant.

A major aspect to consider when analyzing AR processes is *causality*. We have seen that the AR(1) process with $|\phi_1| < 1$ has a causal representation as a linear process. Specifically,

$$X_t = \phi_1 X_{t-1} + w_t = \sum_{i=0}^{\infty} \phi_1^i w_{t-i}.$$

This process was also shown to be stationary. However, when $\phi_1 = 1$, we have a random walk which is not stationary. Writing the random walk as a linear process gives a series that does not converge.

Similarly, we can consider the setting of an AR(1) process with $|\phi_1| > 1$, which will grow exponentially fast. However, we can still write this process in the form of a non-causal linear process.

$$\text{if } X_{t+1} = \phi_1 X_t + w_{t+1} \text{ then } X_t = \phi_1^{-1} X_{t+1} - \phi_1^{-1} w_{t+1}.$$

Continuing in this fashion and noting that $|\phi_1^{-1}| < 1$, we have that

$$\begin{aligned} X_t &= \phi_1^{-1}(\phi_1^{-1} X_{t+2} - \phi_1^{-1} w_{t+2}) - \phi_1^{-1} w_{t+1} \\ &= \phi_1^{-2} X_{t+2} - \phi_1^{-2} w_{t+2} - \phi_1^{-1} w_{t+1} = \ldots = -\sum_{i=1}^{\infty} \phi_1^{-i} w_{t+i}, \end{aligned}$$

which is a linear process with reverse causality.[2] Using this linear process representation, we can compute the stationary autocovariance to be

$$\begin{aligned} K_X(\tau) &= \mathrm{E}\left(\sum_{i,j=1}^{\infty} \phi_1^{-(i+j)} w_{t+\tau+i} w_{t+j} \right) \\ &= \sigma^2 \sum_{i,j=1}^{\infty} \phi_1^{-(i+j)} \mathbf{1}[\tau + i = j] = \sigma^2 \phi_1^{-|\tau|} \sum_{i=1}^{\infty} \phi_1^{-2i} = \frac{\sigma^2 \phi_1^{-|\tau|} \phi_1^{-2}}{1 - \phi_1^{-2}} \end{aligned}$$

where the extra ϕ_1^{-2} comes from the fact that the above sums begin at 1 instead of at 0.

If we strengthen the white noise process w_t to be iid Gaussian with variance σ^2, then we have that the process X_t is Gaussian. Consequently, it is completely characterized by its mean—which is zero—and its autocovariance above. Now consider the causal AR(1) process

$$Y_t = \phi^{-1} Y_{t-1} + v_t$$

where v_t is iid Gaussian white noise with variance $\sigma^2 \phi^{-2}$. This is a mean zero process with autocovariance

$$K_Y(\tau) = \frac{\sigma^2 \phi_1^{-|\tau|} \phi_1^{-2}}{1 - \phi_1^{-2}}.$$

Thus, these two processes are stochastically equivalent—i.e. for any finite collection of time points $t_1, \ldots, t_k$, the vectors $(X_{t_1}, \ldots, X_{t_k})$ and $(Y_{t_1}, \ldots, Y_{t_k})$ are equal in distribution. Thus the non-causal $AR(1)$ process with $|\phi| > 1$ has an equivalent causal representation.

[2] Note: the summation starts from $i = 1$ instead of $i = 0$.

We can extend this idea to general AR(p) processes in order to rewrite a recursively defined AR(p) process as stationary linear process. For some AR operator, we have the general form

$$\Phi(B)X_t = w_t.$$

If the operator $\Phi(B)$ is invertible, then we can simply write the linear process form

$$X_t = \Phi^{-1}(B)w_t.$$

But then we have to determine if the inverse operator exists and what its form is.

Reconsidering the AR(1) process above, we write it as $(1 - \phi_1 B)X_t = w_t$. Considering the complex polynomial $\Phi(z) = 1 - \phi_1 z$ for $z \in \mathbb{C}$, we note that

$$\Phi^{-1}(z) = \frac{1}{1 - \phi_1 z} = 1 + \sum_{j=1}^{\infty} \phi_1^j z^j,$$

which has a radius of convergence of $|\phi_1^{-1}|$. In the case that $|\phi_1| < 1$, we can use this to quickly write

$$X_t = \left(1 + \sum_{j=1}^{\infty} \phi_1^j B^j\right) w_t.$$

For the general AR(p) process, consider the complex polynomial $\Phi(z) = 1 - \phi_1 z - \ldots \phi_p z^p$ and recall that this can be factored on $\mathbb{C}$ into $\Phi(z) = \phi_p(z - r_1)\ldots(z - r_p)$ where $r_1, \ldots, r_p$ are the roots.[3] Then, noting that $(-1)^p \phi_p \prod_{j=1}^p r_j = 1$, we can write

$$\Phi^{-1}(z) = \frac{1}{(1 - r_1^{-1}z)\ldots(1 - r_p^{-1}z).}$$

Now, assuming further that all of the roots $|r_i| > 1$, we can write $\Phi(B)X_t = w_t$ as a causal linear process

$$X_t = \Phi^{-1}(B)w_t = \prod_{j=1}^{p} \left(1 + \sum_{i=1}^{\infty} r_i^{-1} B\right) w_t.$$

2.3.2 Moving Average Process

Next, we reintroduce the moving average process formally as

Definition 2.3.2 (Moving Average Process)**.** *The time series X_t is an MA(q) process if X_t has zero mean and if we can write it as*

$$X_t = w_t + \sum_{j=1}^{q} \theta_j w_{t-j}$$

[3] From the fundamental theorem of algebra, https://en.wikipedia.org/wiki/Fundamental_theorem_of_algebra

where w_t is white noise with variance σ^2 and $\theta_1, \ldots, \theta_q \in \mathbb{R}$ are constants with $\theta_q \neq 0$. Using the backshift operator B, we can write the MA(q) process as $X_t = \Theta(B)w_t$ where

$$\Theta(B) = \left(1 + \sum_{j=1}^{q} \theta_j B^j\right).$$

We will refer to $\Theta(B)$ as the **moving average operator**.

Unlike for autoregressive processes, we already have the MA(q) process written in the form of a linear process. Hence, it will be stationary for any choice of the θ_j. However, similar to how we were able to find a causal AR process that is equivalent to a non-causal one, there is a uniqueness problem with the MA process that needs to be addressed.

For simplicity, consider the MA(1) process $X_t = w_t + \theta_1 w_{t-1}$ with w_t white noise with variance σ^2. This has mean zero and stationary autocovariance

$$K_X(\tau) = \begin{cases} (1+\theta_1^2)\sigma^2 & \text{for } \tau = 0 \\ \theta_1\sigma^2 & \text{for } \tau = 1 \\ 0 & \text{for } \tau \geq 2 \end{cases}$$

Alternatively, we note that the process $Y_t = v_t + \theta_1^{-1} v_t$, with v_t white noise with variance $\theta_1^2\sigma^2$, is also mean zero with the same autocovariance at X_t. Hence, if the white noise processes are Gaussian, then X_t and Y_t are stochastically equivalent. This can certainly cause trouble in a statistics context as if we were to estimate the parameters for the MA(1) model, which parameters would we be estimating?

To choose a specific representation for the MA process, we consider which one is *invertible*. That is, which process can be written as a causal AR process for white noise in terms of X_t? Starting with the general form, we have $X_t = \Theta(B)w_t$. If $\Theta(B)$ is invertible, then we can write $w_t = \Theta^{-1}(B)X_t$. Using the above MA(1) process as an example, we can express the white noise process as

$$w_t = \sum_{i=0}^{\infty} (-1)^i \theta_1^i X_{t-i} \quad \text{or} \quad v_t = \sum_{i=0}^{\infty} (-1)^i \theta_1^{-i} Y_{t-i}.$$

Thus, as only one of θ_1 and θ_1^{-1} can be less than 1 in magnitude, only one of the above series is convergent in the mean squared sense. That process will be the invertible one. Note that w_t is equal in distribution to $\theta_1^{-1} v_t$. Note also that if $\theta_1 = 1$ then we do not have invertibility.

2.3.3 Auto Regressive Moving Average Processes

Now we can combine the AR and MA processes in to the autoregressive moving average process (ARMA), which simply is as follows.

Definition 2.3.3 (Autoregressive Moving Average Process)**.** *The time series X_t is an $ARMA(p,q)$ process if X_t has zero mean and if we can write it as*

$$X_t = w_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j}$$

where w_t is white noise with variance σ^2 and $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q \in \mathbb{R}$ are constants with $\phi_p \neq 0$ and $\theta_q \neq 0$. Using the backshift operator B, we can succinctly write this process as $\Phi(B)X_t = \Theta(B)w_t$ where as before

$$\Phi(B) = \left(1 + \sum_{i=1}^{p} \phi_i B^i\right), \quad \text{and} \quad \Theta(B) = \left(1 + \sum_{j=1}^{q} \theta_j B^j\right).$$

The first thing to note is that similar to the introduction of the AR process, we assume in the definition that X_t has zero mean. If instead it has a mean $\mu \neq 0$, we can subtract off the mean to get

$$\begin{aligned} \Phi(B)(X_t - \mu) &= \Theta(B)w_t \\ \Phi(B)X_t &= \mu\left(1 - \sum_{i=1}^{p} \phi_i\right) + \Theta(B)w_t \end{aligned}$$

and consider the mean zero process.

The second thing to note is that the model is not unique as written. That is, for some invertible operator $\eta(B)$, we can consider the equivalent process

$$\eta(B)\Phi(B)X_t = \eta(B)\Theta(B)w_t.$$

For example, we can consider the white noise process $X_t = w_t$ and, for some $|\theta| < 1$, the equivalent process

$$\begin{aligned} (1 - \theta B)X_t &= (1 - \theta B)w_t \\ X_t &= \theta X_{t-1} - \theta w_{t-1} + w_t. \end{aligned}$$

This may look like a more complex ARMA process, but is in fact just white noise in disguise. To address this problem, we only want to consider AR and MA operators that are relatively prime. That is, for $z \in \mathbb{C}$, we want the polynomials

$$\begin{aligned} \Phi(z) &= 1 - \phi_1 z - \ldots - \phi_p z^p, \text{ and} \\ \Theta(z) &= 1 + \theta_1 z + \ldots + \theta_q z^q \end{aligned}$$

to not have any common roots. In the case that Θ is invertible, we can write the ARMA process as

$$\frac{\Phi(B)}{\Theta(B)} X_t = w_t.$$

Thus, in this form, we see that common factors in Φ and Θ can be cancelled out.

When we write X_t in this way, it is said to be invertible if we have

$$w_t = \frac{\Phi(B)}{\Theta(B)} X_t = X_t + \sum_{j=1}^{\infty} \pi_j X_{t-j}$$

where $\sum_{j=1}^{\infty} |\pi_j| < \infty$. Hence, returning the the previous discussion on the MA processes, we want to write out the process as a convergent series. Considering the MA polynomial $\Theta(z)$ for $z \in \mathbb{C}$, the ARMA process X_t is invertible if and only if all of the roots of $\Theta(z)$ lie outside of the unit disk $\mathcal{D} = \{z : |z| \leq 1\}$.

Similarly, we can write the ARMA process in the form of a stationary linear process

$$X_t = \frac{\Theta(B)}{\Phi(B)} w_t.$$

However, this process may not be causal. A causal process as discussed before can be written as

$$X_t = w_t + \sum_{j=1}^{\infty} \psi_j w_{t-j}$$

with $\sum_{j=1}^{\infty} |\psi_j| < \infty$. A necessary and sufficient condition for causality in an ARMA process is to have an autoregressive polynomial $\Phi(z)$ such that all of its roots lie outside of the unit disk $\mathcal{D} = \{z : |z| \leq 1\}$.

In summary, an ARMA process X_t is

1. causal if $r_1, \ldots, r_p$, the roots of $\Phi(z)$ are such that $|r_i| > 1$;
2. invertible if $r_1, \ldots, r_q$, the roots of $\Theta(z)$ are such that $|r_i| > 1$.

Note that the proof of invertibility is mostly identical to that for causality expect focusing on Θ instead of Φ.

Proof of Causality. Let the roots of $\Phi(z)$ be $r_1, \ldots, r_p$. First, assume that the roots are all outside of the unit disk, and without loss of generality, are ordered so that $1 < |r_1| \leq \ldots \leq |r_p|$. Then, let $|r_1| = 1 + \varepsilon$ for some $\varepsilon > 0$. This implies that $\Phi^{-1}(z)$ exists and has a power series expansion

$$\Phi^{-1}(z) = \sum_{j=0}^{\infty} a_j z^j$$

with a radius of convergence of $|z| < 1 + \varepsilon$.

If we choose a δ such that $0 < \delta < \varepsilon$ then the point $z = 1 + \delta$ lies within the radius of convergence, so

$$\Phi^{-1}(1 + \delta) = \sum_{j=0}^{\infty} a_j (1 + \delta)^j < \infty.$$

As this series converges, we know that there exists a constant $c > 0$ such that $|a_j(1+\delta)^j| < c$ for all j. Hence, $|a_j| < c(1+\delta)^{-j}$. Thus,

$$\sum_{j=0}^{\infty} |a_j| < c \sum_{j=0}^{\infty} (1+\delta)^{-j} < \infty,$$

and the sequence of a_j is absolutely summable. This implies that for the ARMA process $\Phi(B)X_t = \Theta(B)w_t$, we can write

$$X_t = \Phi^{-1}(B)\Theta(B)w_t = w_t + \sum_{j=1}^{\infty} \psi_j w_{t-j}.$$

Since the a_j are absolutely summable, so are the coefficients ψ_j. Thus, we have that X_t is a causal process.

For the reverse, let's assume that X_t, defined by $\Phi(B)X_t = \Theta(B)w_t$, be a causal process. That is, we can write

$$X_t = w_t + \sum_{j=1}^{\infty} \psi_j w_{t-j} \quad \text{and} \quad \sum_{j=1}^{\infty} |\psi_j| < \infty.$$

As a result, we can write $X_t = \Psi(B)w_t$ and also $\Phi(B)X_t = \Theta(B)w_t$. Equating the two right hand expressions, we have

$$\Theta(B)w_t = \Phi(B)\Psi(B)w_t.$$

Writing the complex polynomial $\Phi(z)\Psi(z) = \sum_{i=1}^{\infty} a_j z^j$, we know that this series has a radius of convergence of at least $|z| \leq 1$ as Ψ also does and Φ is a finite polynomial. Hence, it makes sense to write

$$\sum_{j=1}^{q} \theta_j w_{t-j} = \sum_{j=1}^{\infty} a_j w_{t-j}.$$

If we consider computing the covariance of each sum with w_{t-i} for $i = 1, 2, 3, \ldots$, we get a sequence of equations $\theta_j = a_j$ for $j \leq q$ and $a_j = 0$ for $j > q$. That is, we can equate matching coefficients in the two series. Thus, $\Theta(z) = \Phi(z)\Psi(z)$ for $|z| \leq 1$.

We know that none of the roots of the polynomial Ψ can lie on or within the unit disk. Hence, if there exists a $z_0 \in \mathcal{D}$ such that $\Phi(z_0) = 0$, then $\Theta(z_0) = 0$ and the two polynomials have a common root. As they are assumed to have no common factors, this implies that all roots of Φ lie outside of the unit disk. □

2.3.4 ARIMA

Often, we do not have an ARMA process but an ARMA process with some deterministic trend. Thus, the process is not stationary, but often can be transformed into a

stationary process via the differencing operator. For example, if X_t is a stationary process, and we have Y_t defined by

$$Y_t = \beta_0 + \beta_1 t + X_t,$$

then by applying the first difference operator, we have

$$\nabla Y_t = \beta_1 + X_t - X_{t-1},$$

which is stationary. This motivates the following definition.

Definition 2.3.4 (Autoregressive Moving Integrated Average Process)**.** *The time series X_t is an ARIMA(p, d, q) process if the dth difference process,*

$$\nabla^d X_t = (1 - B)^d X_t,$$

is an ARMA(p, q) process. We can write it in terms of the backshift operator as $\Phi(B)(1 - B)^d X_t = \Theta(B) w_t$ where as before

$$\Phi(B) = \left(1 + \sum_{i=1}^{p} \phi_i B^i\right), \quad \text{and} \quad \Theta(B) = \left(1 + \sum_{j=1}^{q} \theta_j B^j\right).$$

As before, we assume in the definition that $\nabla^d X_t$ has zero mean. If instead it has a mean $\mu \neq 0$, we write $\Phi(B)(1 - B)^d X_t = \mu \left(1 - \sum_{i=1}^{p} \phi_i\right) + \Theta(B) w_t$. For example, if we have $X_t = \beta_0 + \beta_1 t + \phi X_{t-1} + w_t$ for $\beta_0, \beta_1 \in \mathbb{R}$ and $|\phi| < 1$, then

$$\begin{aligned} \nabla X_t &= X_t - X_{t-1} \\ &= [\beta_0 + \beta_1 t + \phi X_{t-1} + w_t] - [\beta_0 + \beta_1 (t-1) + \phi X_{t-2} + w_{t-1}] \\ &= \beta_1 + \phi \nabla X_{t-1} + w_t - w_{t-1}. \end{aligned}$$

This is an ARMA$(1, 1)$ process with non-zero mean that can be written as

$$(1 - \phi B)\delta X_t = \beta_1 + (1 - B) w_t.$$

Note that the mean is not β_1 but in fact $\beta_1/(1 - \phi)$.

2.4 Testing for Stationarity and Autocorrelation

When presented with time series data, we often want to know if the series is stationary as it stands or after applying some difference operators. Similarly, we may be interested in knowing if there are any significant autocorrelations at various lags. These questions motivate a large collection of statistical tests.

2.4.1 Box-Pierce and Ljung-Box Tests

The R function `Box.test()` in the `stats` package performs both the Box-Pierce and Ljung-Box tests.

For a stationary time series X_t, we denote the estimated autocorrelations to be $\hat{\rho}_X(h)$ at lag h. If we assume that the true autocorrelations are zero—i.e. $\rho_X(h) = 0$ for all $h \neq 0$—then we have white noise. Instead of visually looking at a plot of the autocorrelation, we can use the Box-Pierce test to test for non-zero correlations by combining the estimated autocorrelations at lags $1, \ldots, h$ for some user chosen value h. We aim to test the hypotheses

$$H_0 : \rho_X(1) = \ldots = \rho_X(h) = 0, \quad H_1 : \exists j \text{ s.t. } \rho_X(j) \neq 0.$$

Under H_0, we have that $\sqrt{n}\hat{\rho_X}(j)$ is approximately $\mathcal{N}(0,1)$ for $j = 1, \ldots, h$. The test statistic for the Box-Pierce test is

$$Q_{\text{BP}} = n \sum_{j=1}^{h} \hat{\rho}_X(j)^2,$$

which will be approximately $\chi^2(h)$ under H_0. Recall, however, that the ability to estimate $\hat{\rho}_X$ becomes harder for large lags especially if the data size is small. Hence, h should not be set to be too large in practice.

Another version of this test is the Ljung-Box test, which has a similar form to the Box-Pierce test and the same approximate $\chi^2(()\,h)$ distribution. The alternative form is supposed to give a distribution under H_0 that is closer to the desired $\chi^2(h)$. The test statistic is

$$Q_{\text{LB}} = n(n+2) \sum_{j=1}^{h} \frac{\hat{\rho}_X(j)^2}{n-j}.$$

In the function `Box.test()`, there is a `fitdf` argument. The point of this argument is to reduce the chi-squared degrees of freedom in the case that you are fitting a model first. In particular, if you first fit an ARMA(p,q) model to X_t and then apply `Box.test` to the residual process, you should reduce the number of degrees of freedom by $p + q$. In this case, we require $h > p + q$ to be able to perform these tests. Ljung and Box study Q_{LB} in a 1978 research article[4] and look at the first and second moments of their statistic for how closely it coincides with the $\chi^2(h)$ distribution.

2.4.2 Durbin–Watson Test

As mentioned above, the Box-Pierce and Ljung-Box tests can be applied to the residuals of an ARMA model with the goal of determining if there are non-zero

[4] Ljung, Greta M., and George EP Box. "On a measure of lack of fit in time series models." Biometrika 65, no. 2 (1978): 297-303.

autocorrelations among the residuals. Similarly, the Durbin-Watson test tests for autocorrelations of order 1 among the residuals of a linear model.

Considering the linear model

$$X_t = \beta_0 + \beta_1 t + \ldots + \beta_p t^p + r_t$$

with $t = 1, \ldots, T$, we can compute the least squares estimator $\hat{\beta}$ and then compute the residuals $\hat{r}_t = X_t - \left\langle \hat{\beta}, (1, t, \ldots, t^p) \right\rangle$. The Durbin–Watson Test assumes the following model for the residuals:

$$\hat{r}_t = \rho \hat{r}_{t-1} + w_t$$

where w_t is white noise. Then, it tests the hypotheses $H_0 : \rho = 0,\ H_1 : \rho \neq 0$. It does this by computing the test statistics

$$Q_{\mathrm{DW}} = \frac{\sum_{t=2}^{T} (\hat{r}_t - \hat{r}_{t-1})^2}{\sum_{t=1}^{T} \hat{r}_t^2}.$$

If this test statistic is close to zero, it implies that $\hat{r}_t$ and $\hat{r}_{t-1}$ are close in value indicating a strong positive autocorrelation of order 1. In contrast, if the test statistic is large (close to the max of 4), then it indicates that there is a strong negative autocorrelation of order 1. Otherwise, a test statistic near 2 indicates no autocorrelation of order 1. In the R function, `dwtest()` in the `lmtest` package, p-values are computed for this statistic. The documentation claims that *Under the assumption of normally distributed [errors], the null distribution of the Durbin-Watson statistic is the distribution of a linear combination of chi-squared variables.* Furthermore, for large sample sizes, this code apparently switches to a normal approximation for the p-value computation.

2.4.3 Breusch–Godfrey test

The Breusch–Godfrey test is similar to the Durbin-Watson test in the sense that it applies to the residuals of a linear model. In this case, it can test for higher order autocorrelations AR(p) than just AR(1) processes. In this case, the R^2 value is computed for a regression model, being the ratio of the regression sum of squares to the total sum of squares. Then nR^2 is compared to a $\chi^2(p)$ where p is the order of the AR process to be tested. Alternatively, the documentation for the R implementation, `bgtest` in the `lmtest` package, says that the user can use the Chi-Squared distribution or can switch to the F distribution if desired.

Note that there is also a Breusch–Pagan test, which tests for heteroskedasticity in the residuals of a linear model. An R implementation can be found in `bptest` in the `lmtest` package.

2.4.4 Augmented Dickey-Fuller Test

Switching away from testing for non-zero autocorrelations, we now consider testing for stationarity or non-stationarity of a time series. These tests are often referred to as *unit root tests*, because—recalling the previous sections—if the autoregressive operator Φ has a unit root, then the process is not stationary. Hence, these tests aim to determine whether or not a unit root exists based on some observed data.

The Dickey-Fuller Test performs such a unit root test for AR(1) models. In this case, the null hypothesis is that $\Phi(z)$ has a root $|r| = 1$, and the alternative is that $|r| > 1$ for all i. If

$$X_t = \phi X_{t-1} + w_t$$

then the first difference can be written as

$$\nabla X_t = (\phi - 1)X_{t-1} + w_t.$$

Denoting $\phi' = \phi - 1$, we want to test the null $H_0 : \phi' = 0$ against the alternative $H_1 : \phi \neq 0$. This is done by estimating $\hat{\phi}'$ and the standard error for $\hat{\phi}'$. This null hypothesis is equivalent to testing $H_0 : \phi = 1$ or that the polynomial $\Phi(z) = 1 - \phi z$ has a unit root.

The Augmented Dickey-Fuller Test extends this idea to AR(p) models. If we have

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + w_t,$$

then the first difference can be written as

$$\nabla X_t = \phi_1' X_{t-1} + \sum_{i=1}^{p-1} \phi_{i+1}' \nabla X_{t-i} + w_t$$

where the coefficients are $\phi_1' = \sum_{j=1}^{p} \phi_j - 1$ and $\phi_i' = -\sum_{j=i}^{p} \phi_j$ for $j > 1$. Thus, if 1 is a root—i.e. if $\Phi(1) = 0$—then that implies that $\phi_1' = 0$. Hence, we can perform a similar test to the Dickey-Fuller test above.

In R, the Augmented Dickey-Fuller test is implemented in the function `adf.test()` in the `tseries` package. In this version of the test, a constant and a linear term are first assumed and the residual process is run through the above test. That is, we consider the model

$$X_t = \beta_0 + \beta_1 t + \sum_{i=1}^{p} \phi_i X_{t-i} + w_t$$

with a deterministic linear trend. The R function also requires the user to choose how many lags to use when estimating the parameters $\hat{\phi}$.

2.4.5 Phillips–Perron test

An alternative to the Augmented Dickey-Fuller test is the Phillips-Perron test. The set up is the same, but the test is designed to be more robust to deviations in the assumptions. In R, this test can be performed by the function `pp.test()` in the `tseries` package. The test statistic is more complicated, and the p-value is computed via a table of values and linear interpolation. The documentation also points out that the Newey-West estimator is used for the variance, which is a robust estimator of the covariance in linear regression when the classic assumptions of homoscedastic uncorrelated errors is violated.

2.5 Autocorrelation and Partial Autocorrelation

Given some time series data, we often wish to diagnose the type time series process that produced the data. Two tools we can use are the estimated autocorrelation function and the estimated partial autocorrelation function.

First, we recall the notion of partial correlation outside of the time series context. Given two random variables X and Y, we may compute the correlation $\text{corr}(X, Y)$, which measures the linearity between the two variables. That is, the closer to 1 the magnitude of the correlation is, the closer X and Y are to being linearly dependent. Often in statistics, we are reminded that correlation does not imply causation. In fact, given two correlated random variables, there may be a third random variable Z influencing both of them. Hence, consider iid observations $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$. We can define the partial correlation between X and Y given Z to be the correlation of the residuals of X and Y each regressed on Z. That is, let

$$\hat{X}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_i \quad \text{and} \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i$$

be the ith fitted values for X and Y, respectively. Then, the partial correlation is

$$\text{corr}(X - \hat{X}, Y - \hat{Y})$$

where $X - \hat{X}$ and $Y - \hat{Y}$ are the residuals for X and Y, respectively.

The idea of partial correlation is to remove the dependency of the confounding variable Z. Hence, if $\text{corr}(X - \hat{X}, Y - \hat{Y}) = 0$, then X and Y are said to be conditionally uncorrelated—or conditionally independent in the case that X, Y, and Z are jointly normal. An hypothetical example is if X is the price of a subjects car and Y is the price of a subjects house, we may expect X and Y to be positively correlated. However, conditioning on Z, the subjects income, the house value and car value may be conditionally uncorrelated. Note that in the case that the random variables are jointly normal, we have that

$$\text{corr}(X - \hat{X}, Y - \hat{Y}) = \text{corr}(X, Y|Z) = \frac{\text{E}[(X - \text{E}X)(Y - \text{E}Y)|Z]}{\text{E}[(X - \text{E}X)^2|Z]\text{E}[(Y - \text{E}Y)^2|Z]}.$$

In the time series context we define

Definition 2.5.1 (Partial Autocorrelation)**.** *Let X_t be a stationary process, then the partial autocorrelation at lag h is*

$$\varphi_X(h) = \begin{cases} \rho_X(1) & \text{if } h = 1 \\ \text{corr}(X_{t+h} - \hat{X}_{t+h}, X_t - \hat{X}_t) & \text{if } h > 1 \end{cases}$$

where $\hat{X}_{t+h}$ and $\hat{X}_t$ are the result of regressing each respective term on all of the intermediate terms. That is,

$$\begin{aligned} \hat{X}_{t+h} &= \beta_1 X_{t+h-1} + \ldots + \beta_{h-1} X_{t+1} \\ \hat{X}_t &= \beta_1 X_{t+1} + \ldots + \beta_{h-1} X_{t+h-1} \end{aligned}$$

where the intercept term is excluded as X_t is assumed to be mean zero. Note that due to stationarity of X_t and the symmetry of the autocorrelation function, the β's above are the same coefficients. Lastly, if X_t is a Gaussian process, we can write

$$\varphi_X(h) = \text{corr}(X_{t+h}, X_t | X_{t+h-1}, \ldots, X_{t+1})$$

for $h > 1$.

2.5.1 ACF for AR(p)

To begin, we consider the causal mean zero AR(1) process $X_t = \phi X_{t-1} + w_t$ where $|\phi| < 1$. Then, we can multiply by X_{t-h} and note that

$$\begin{aligned} \text{E}\left(X_t X_{t-h}\right) &= \text{E}\left(\phi X_{t-1} X_{t-h}\right) + \text{E}\left(w_t X_{t-h}\right) \\ K_X(h) &= \phi K_X(h-1) + 0. \end{aligned}$$

Therefore, the autocovariance is defined by a first order difference equation

$$f(h) = \phi f(h-1).$$

As the characteristic polynomial is $1 - \phi z$ with root $z_1 = \phi^{-1}$, we can solve this difference equation to get the solution $f(h) = c z_1^{-h}$ for some constant c corresponding to the *initial condition* $f(0) = c$. This can be checked by plugging in the solution to the equation to get

$$c\left(\phi^{-1}\right)^{-h} = c\phi\left(\phi^{-1}\right)^{-h+1}.$$

Revisiting the autocorrelation, we have $K_X(0) = \text{Var}\left(X_t\right)$, so $\rho_X(h)$ follows the same difference equation with $c = 1$. Hence, if we have an AR(1) process as above,

$$\rho_X(h) = \phi^{-h}.$$

For the causal AR(2) process, we have $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + w_t$. Proceeding as before, we note that

$$\begin{aligned} \text{E}\left(X_t X_{t-h}\right) &= \text{E}\left(\phi_1 X_{t-1} X_{t-h}\right) + \text{E}\left(\phi_2 X_{t-2} X_{t-h}\right) + \text{E}\left(w_t X_{t-h}\right) \\ K_X(h) &= \phi_1 K_X(h-1) + \phi_2 K_X(h-2). \end{aligned}$$

The roots of the characteristic polynomial will tell us about the behaviour of the process X_t. For $1 - \phi_1 z - \phi_2 z^2$, we denote the two roots as z_1 and z_2. Recall that $|z_i| > 1$ as we assume X_t is causal. There are three possible settings to consider[5]

1. if $z_1 \neq z_2$ and the roots are real, then we have the solution to the second order difference equation
$$\rho(h) = c_1 z_1^{-h} + c_2 z_2^{-h}$$
where c_1 and c_2 are two constants such that $c_1 + c_2 = 1$.

2. if $z_1 = z_2$ and necessarily real, then we have $\rho(h) = z_1^{-h}(c_1 + c_2 h)$.

3. if $z_1 = \bar{z}_2$ are complex conjugate roots, then
$$\begin{aligned}\rho(h) &= c_1 z_1^{-h} + \bar{c}_1 \bar{z}_1^{-h} \\ &= |c_1||z_1|^{-h} \left(\mathrm{e}^{-ib}\mathrm{e}^{-i\theta h} + \mathrm{e}^{ib}\mathrm{e}^{i\theta h}\right) \\ &= 2|c_1||z_1|^{-h}\cos(\theta h + b)\end{aligned}$$

In all three cases, we have the autocorrelation $\rho(h)$ decaying exponentially to zero. The rate of decay depends on the magnitude of the roots. Furthermore, if the roots are complex, then there is periodic behaviour in the process.

This can be extended to the AR(p) process where we have a pth order difference equation for ρ. The resulting solution will look like

$$\rho(h) = z_1^{-h} f_1(h) + \ldots + z_r^{-h} f_r(h)$$

where $z_1, \ldots, z_r$ are the unique roots with multiplicities $m_1, \ldots, m_r$ with $\sum_{i=1}^{r} m_i = p$ and where $f_i(h)$ is a polynomial in h of degree m_i.

2.5.2 ACF for MA(q)

The exposition of the autocorrelation for the general MA(q) process is much simpler relative to the previous discussion of the AR(p) process. Let $X_t = \sum_{j=0}^{q} \theta_j w_j$ with $\theta_0 = 1$, then

$$K_X(h) = \sum_{j=0}^{q-h} \theta_j \theta_{j+h}$$

for $h = 0, \ldots, q$ and with $K_X(h) = 0$ otherwise. Noting the variance is $K_X(0) = 1 + \theta_1^2 + \ldots + \theta_q^2$, we have an autocorrelation of

$$\rho_X(h) = \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{\sum_{j=0}^{q} \theta_j^2} \quad \text{for } h \leq q.$$

Thus, unlike the AR process, the autocorrelation for the MA(q) process is zero for $h > q$. Thus, it can be used to identify the order of the process.

[5] For more details, see a textbook on difference equations.

2.5.3 PACF for AR(p)

To introduce why the partial correlation is of interest, we first consider the causal AR(1) process $X_t = \phi X_{t-1} + w_t$. From before, we saw that

$$\begin{aligned}\operatorname{corr}(X_t, X_{t-2}) &= \operatorname{corr}(\phi X_{t-1} + w_t, X_{t-2}) \\ &= \operatorname{corr}(\phi^2 X_{t-2} + \phi w_{t-1} + w_t, X_{t-2}) \\ &= \operatorname{corr}(\phi^2 X_{t-2}, X_{t-2}) + \operatorname{corr}(\phi w_{t-1}, X_{t-2}) + \operatorname{corr}(w_t, X_{t-2}) \\ &= \phi^2 + 0 + 0.\end{aligned}$$

In contrast, if we consider

$$\operatorname{corr}(X_t - \phi X_{t-1}, X_{t-2} - \phi X_{t-1}) = \operatorname{corr}(w_t, X_{t-2} - \phi X_{t-1}) = 0.$$

Hence when taking X_{t-1} into account, the autocorrelation between X_t and X_{t-2} is zero.

To properly consider the partial autocorrelation, we need to compute the least squares estimator $\hat{X}_t$ for X_t based on previous time points. For example, to compute $\varphi(2)$, we take $\hat{X}_t = \hat{\beta} X_{t-1}$ where $\hat{\beta}$ is chosen to minimize

$$\begin{aligned}&\mathrm{E}(X_t - \hat{X}_t)^2 = \mathrm{E}(X_t - \beta X_{t-1})^2 \\ &\quad = \mathrm{E}\left(X_t^2\right) - 2\beta \mathrm{E}\left(X_t X_{t-1}\right) + \beta^2 \mathrm{E}\left(X_{t-1}^2\right) = (1+\beta^2) K_X(0) - 2\beta K_X(1).\end{aligned}$$

By taking the derivative with respect to β, we can find the critical point $\hat{\beta} = K_X(1)/K_X(0)$. Similarly, for $\hat{X}_{t-2} = \hat{\beta} X_{t-1}$, we have

$$\begin{aligned}&\mathrm{E}(X_{t-2} - \beta X_{t-1})^2 \\ &\quad = \mathrm{E}\left(X_{t-2}^2\right) - 2\beta \mathrm{E}\left(X_{t-2} X_{t-1}\right) + \beta^2 \mathrm{E}\left(X_{t-1}^2\right) = (1+\beta^2) K_X(0) - 2\beta K_X(1)\end{aligned}$$

as before. In the case of the AR(1) model, we have $\hat{\beta} = \phi$. Thus, we have from before that $\varphi(1) = \phi$ and $\varphi(2) = 0$ and, in fact, $\varphi(h) = 0$ for $h \geq 2$.

For the general AR(p) process, $X_t = w_t + \sum_{i=1}^{p} \phi_i X_{t-i}$, we have a similar set up. For lags $h > p$, if we assume for now that the least squares estimator is

$$\hat{X}_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p},$$

then we get a similar calculation as above. Namely that

$$\operatorname{corr}(X_t - \hat{X}_t, X_{t-h} - \hat{X}_{t-h}) = \operatorname{corr}(w_t, X_{t-h} - \hat{X}_{t-h}) = 0.$$

In the case that the lag is less than or equal to p, we need to determine how the estimate the coefficients β_i before computing the PACF.

2.5.4 PACF for MA(1)

For the invertible MA(1) model $X_t = w_t + \theta w_{t-1}$, which is with $|\theta| < 1$, we can write it as a convergent infinite series

$$X_t = w_t + \sum_{i=1}^{\infty} \theta^i X_{t-i}$$

in terms of the X_{t-i}. Then, applying similar tricks as above gives a least squares estimator for $\hat{X}_t = \hat{\beta} X_{t-1}$ to be $\hat{\beta} = K_X(1)/K_X(0)$. In the case of the MA(1) process, we have $\hat{\beta} = \theta/(1+\theta^2)$. Hence,

$$\begin{aligned}\operatorname{cov}&\left(X_t - \frac{\theta X_{t-1}}{1+\theta^2}, X_{t-2} - \frac{\theta X_{t-1}}{1+\theta^2}\right) \\ &= K_X(2) - \frac{2\theta}{1+\theta^2} K_X(1) + \left(\frac{\theta}{1+\theta^2}\right)^2 K_X(0) = \frac{-\theta^2}{1+\theta^2}.\end{aligned}$$

Also, the variance is

$$\begin{aligned}\operatorname{Var}\left(X_t - \frac{\theta X_{t-1}}{1+\theta^2}\right) &= K_X(0)\left(1 + \left(\frac{\theta}{1+\theta^2}\right)^2\right) - \frac{2\theta}{1+\theta^2} K_X(1) \\ &= 1 + \theta^2 + \frac{\theta^2}{1+\theta^2} - \frac{2\theta^2}{1+\theta^2} = \frac{1+\theta^2+\theta^4}{1+\theta^2}.\end{aligned}$$

Thus, the partial autocorrelation at lag 1 is $\varphi(1) = -\theta^2/(1+\theta^2+\theta^4)$. This can be extended to lags greater than one to show that the partial autocorrelation for the MA(1) process decreases but does not vanish as the lag increases.

Hence, we have the following table:

	AR(p)	MA(q)
ACF	decreases geometrically	= zero for lags $> q$
PACF	= zero for lags $> p$	decreases geometrically

This means that we can use the ACF and PACF to try to understand the behaviour of a time series process.

Chapter 3

Estimation and Forecasting

Introduction

Thus far, we have considered many types of time series models, but have performed little in the way of actual statistics. In this chapter, we consider to main goals of time series models: estimating parameters and forecasting/predicting. For the first topic, we will consider different methods for estimating parameters as well as model selection methods to determine the best fit to the data. For the second part, we consider the task of prediction in time series.

For a time series observed at $X_1, \ldots, X_T$, we may want to fit a causal invertible ARMA(p,q) process,

$$X_t = w_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j},$$

to the data by estimating the coefficients $\hat{\phi}_i$ and $\hat{\theta}_j$.

3.1 The AR process

3.1.1 Estimation for AR processes

There are two approaches to estimating the parameters of the AR(p) process that we will consider: Using (1) the Yule-Walker equations or (2) the maximum likelihood estimator. More methods are available in the `R` function `ar()`.

The Yule-Walker Estimator

We first consider the causal AR(p) process and the autocovariance function. We first assume that the mean $\mathrm{E}X_t = 0$. In practise, we can estimate the mean by

$\bar{X} = T^{-1}\sum_{t=1}^{T} X_t$ and then consider the centred time series $X_i - \bar{X}$. For estimating the variance for the white noise process σ^2,

$$\begin{aligned} X_t &= w_t + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} \\ K_X(0) = \operatorname{cov}(X_t, X_t) &= \operatorname{cov}(X_t, w_t + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p}) \\ &= \sigma^2 + \phi_1 K_X(1) + \ldots + \phi_p K_X(p) \\ \sigma^2 &= K_X(0) - \phi_1 K_X(1) - \ldots - \phi_p K_X(p). \end{aligned}$$

Hence, we can use the estimates for the autocovariance to estimate σ^2.

$$\hat{\sigma}^2 = \hat{K}_X(0) - \phi_1 \hat{K}_X(1) - \ldots - \hat{\phi}_p K_X(p).$$

However, we need estimators for the parameters ϕ_i. To estimate these ϕ_i, we can consider more equations based on the autocovariance at lags 1 through p.

$$\begin{aligned} K_X(1) &= \operatorname{cov}(X_{t-1}, X_t) = \phi_1 K_X(0) + \phi_2 K_X(1) + \ldots + \phi_p K_X(p-1) \\ K_X(2) &= \operatorname{cov}(X_{t-2}, X_t) = \phi_1 K_X(1) + \phi_2 K_X(0) + \ldots + \phi_p K_X(p-2) \\ &\vdots \\ K_X(p) &= \operatorname{cov}(X_{t-p}, X_t) = \phi_1 K_X(p-1) + \phi_2 K_X(p-2) + \ldots + \phi_p K_X(0) \end{aligned}$$

Here, we have p linear equations with p unknowns, which can be written as $K = \Gamma\phi$ where

$$K = \begin{pmatrix} K_X(1) \\ \vdots \\ K_X(p) \end{pmatrix}, \quad \phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}, \quad \Gamma = \begin{pmatrix} K_X(0) & K_X(1) & \cdots & K_X(p-1) \\ K_X(1) & K_X(0) & \cdots & K_X(p-2) \\ \vdots & \ddots & \ddots & \vdots \\ K_X(p-1) & K_X(p-2) & \cdots & K_X(0) \end{pmatrix}.$$

We can also write $\sigma^2 = K_X(0) - \phi^{\mathrm{T}} K$. This system of $p+1$ equations is known as the *Yule-Walker Equations*. We can solve for the coefficients $\phi = \Gamma^{-1}K$ as the matrix Γ is positive definite. As a result, $\sigma^2 = K_X(0) - K^{\mathrm{T}}\Gamma^{-1}K$, because Γ is symmetric.

We can use the estimator for the autocovariance from Chapter 1 to get a data driven estimate for the parameters for this time series:

$$\hat{\phi} = \hat{\Gamma}^{-1}\hat{K}, \quad \text{and} \quad \hat{\sigma}^2 = \hat{K}_X(0) - \hat{K}^{\mathrm{T}}\hat{\Gamma}^{-1}K.$$

These estimators can be shown to converge in distribution to a multivariate normal distribution.

Theorem 3.1.1 (Asymptotic Normality for Yule-Walker)**.** *Given $\hat{\phi}_i$ and $\hat{\sigma}^2$ as above for a causal AR(p) process, we have as $T \to \infty$,*

$$\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\Gamma^{-1}\right) \quad \textit{and} \quad \hat{\sigma}^2 \xrightarrow{P} \sigma^2.$$

Corollary 3.1.1 (Asymptotic Normality for PACF)**.** *For the causal AR(p) process, as* $T \to \infty$*,*

$$\sqrt{T}\hat{\varphi}(h) \xrightarrow{d} \mathcal{N}(0,1).$$

for lags $h > p$*.*

In the standard `stats` package in `R`, the function `ar()` fits an autoregressive model to time series data, which can implement many ways to estimate the parameters, but defaults to the Yule-Walker equations. To demonstrate it, we can use the `arima.sim()` function to simulate $T = 100$ data points from the AR(1) process

$$X_t = 0.7X_{t-1} + w_t.$$

Using the Yule-Walker equations, we get $\hat{\phi}_1 = 0.73$. Note that the `ar()` function fits models for AR(1) up to AR(20) and then chooses the best with respect to AIC. In the case of data from the AR(3) process

$$X_t = 0.7X_{t-1} - 0.3X_{t-3} + w_t$$

we get the fitted model

$$X_t = 0.752X_{t-1} - 0.002X_{t-2} - 0.285X_{t-3}.$$

Plots of these two processes with the estimated ACF and PACF are displayed in Figure 3.1.

Maximum Likelihood Estimation

Given that w_t is Gaussian white noise, we can write down the likelihood and solve the maximum likelihood estimator. There is actually more than one MLE approach in time series. Also, this can be made more tractible by using conditional probability.

Begining with the causal AR(1) process $X_t = \mu + \phi(X_{t-1} - \mu) + w_t$ with some mean $\mu \in \mathbb{R}$, we use the recursive definition of the process to write the likelihood as

$$\begin{aligned} L(\mu, \phi, \sigma^2; X_1, \ldots, X_T) &= f(X_1, \ldots, X_T; \mu, \phi, \sigma^2) \\ &= f(X_1) f(X_2|X_1) \ldots f(X_T|X_{T-1}). \end{aligned}$$

Assuming the white noise is Gaussian, the term $X_1 \sim \mathcal{N}\left(\mu, \sigma^2/(1-\phi^2)\right)$ as we have solved before.[1] Meanwhile, recalling that normality is preserved under conditioning, the conditional distribution of $X_t|X_{t-1}$ is $\mathcal{N}\left(\mu + \phi(X_{t-1} - \mu), \sigma^2\right)$. Hence, putting it all together gives

$$\begin{aligned} L(\mu, \phi, \sigma^2) = &\frac{(1-\phi^2)^{1/2}}{(2\pi\sigma^2)^{T/2}} \\ &\times \exp\left[-\frac{1}{2\sigma^2}\left\{(X_1 - \mu)^2(1-\phi^2) + \sum_{t=2}^{T}((X_t - \mu) - \phi(X_{t-1} - \mu))^2\right\}\right] \end{aligned}$$

[1] Note that here, we are considering X_1 as an infinite causal linear process.

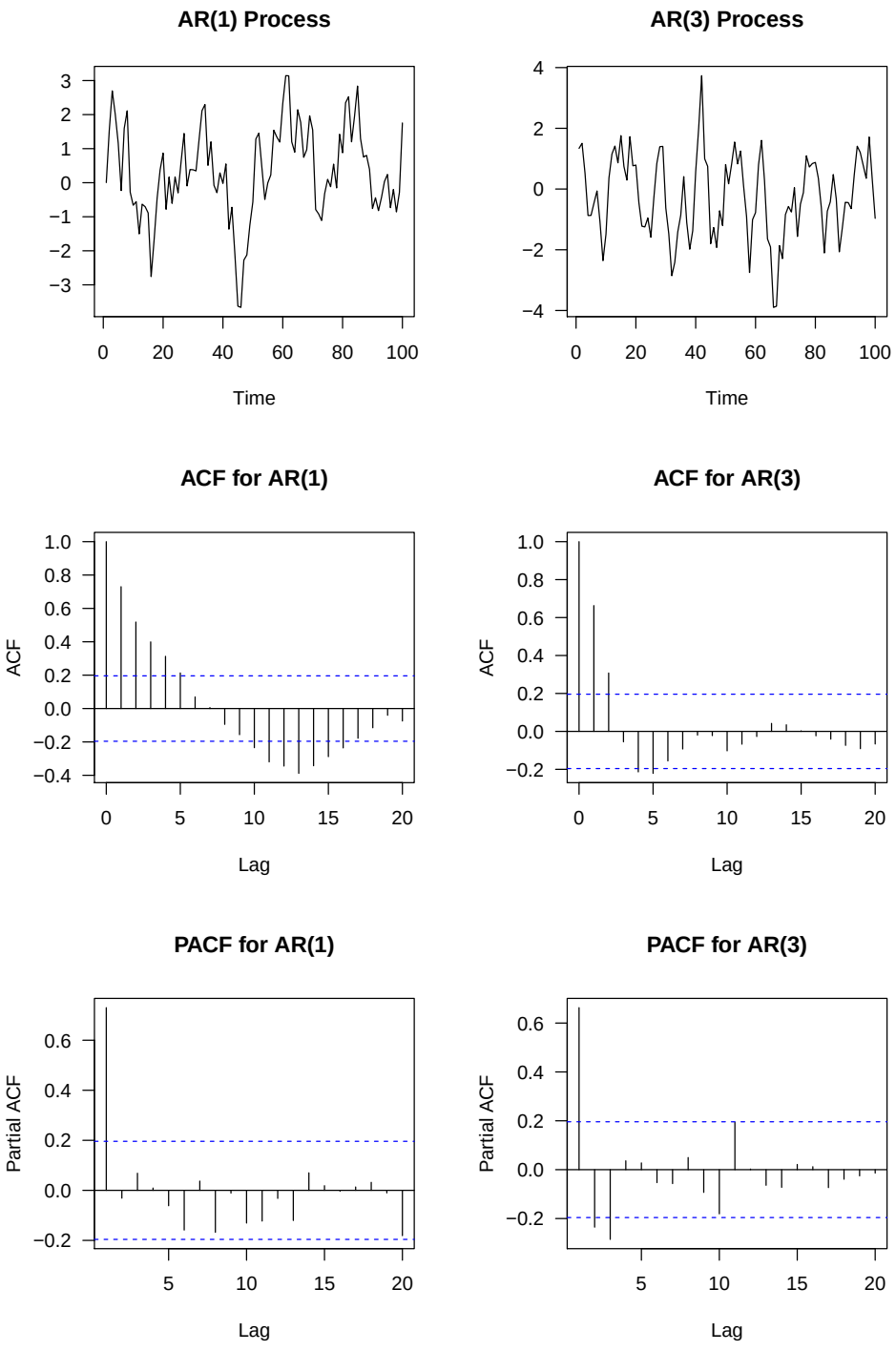

Figure 3.1: Simulated AR(1) and AR(3) processes with estimated ACF and PACF.

Writing the *unconditional* sum of squares in the exponent as

$$S_u(\mu,\phi) = (X_1-\mu)^2(1-\phi^2) + \sum_{t=2}^{T}((X_t-\mu)-\phi(X_{t-1}-\mu))^2$$

we can take derivatives of the log likelihood to solve for the MLEs. For the variance,

$$\begin{aligned}
&\frac{\partial \log(L)}{\partial \sigma^2} = \\
&= \frac{\partial}{\partial \sigma^2}\left\{\frac{1}{2}\log(1-\phi^2) - \frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{S_u(\mu,\phi)}{2\sigma^2}\right\} \\
&= -\frac{n}{2\sigma^2} + \frac{S_u(\mu,\phi)}{2\sigma^4},
\end{aligned}$$

which gives $\hat{\sigma}^2 = S_u(\mu,\phi)/T$. However, solving for MLEs $\hat{\mu}$ and $\hat{\phi}$ are not as straight forward, because we would have to solve the nonlinear system of equations

$$\begin{aligned}
0 &= -2(1-\phi^2)(X_1-\mu) + 2(1-\phi)\sum_{t=2}^{T}(X_t - \phi X_{t-1} - \mu(1-\phi)) \\
0 &= \frac{-\phi}{1-\phi^2} - \frac{1}{2\sigma^2}\left\{2\phi(X_1-\mu)^2 - 2\sum_{t=2}^{T}(X_t-\phi X_{t-1}-\mu(1-\phi))(X_{t-1}-\mu)\right\}.
\end{aligned}$$

This headache arises due to the starting point X_1. If we condition the likelihood on X_1, we can simplify the problem.[2]

Conditioning on X_1, we have

$$L(\mu,\phi,\sigma^2|X_1) = \frac{1}{(2\pi\sigma^2)^{(T-1)/2}}\exp\left[\sum_{t=2}^{T}((X_t-\mu)-\phi(X_{t-1}-\mu))^2\right].$$

Thus, the MLE for the variance is $\hat{\sigma}^2 = S_c(\mu,\phi)/(T-1)$ where similarly to above S_c is the *conditional* sum of squares in the exponent. We can rewrite S_c as

$$S_c(\mu,\phi) = \sum_{t=2}^{T}(X_t - (\alpha + \phi X_{t-1}))^2$$

where $\alpha = \mu(1+\phi)$, which coincides with simple linear regression. Hence, for the design matrix $X \in \mathbb{R}^{(T-1)\times 2}$ with columns 1 and X_t for $t = 1,\ldots,T-1$, the least squares estimator is

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\phi}\end{pmatrix} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(X_2,\ldots,X_T)^{\mathrm{T}},$$

[2] What we just did above is the *unconditional likelihood*. What follows is the *conditional likelihood* as we condition on X_1 to remove the nonlinearity.

which after some computation can be reduced to

$$\hat{\phi} = \frac{\sum_{t=2}^{T}(X_t - \bar{X}_{(2)})(X_{t-1} - \bar{X}_{(1)})}{\sum_{t=2}^{T}(X_{t-1} - \bar{X}_{(1)})^2}$$

where $\bar{X}_{(1)} = (T-1)^{-1}\sum_{t=1}^{T-1} X_t$ and $\bar{X}_{(2)} = (T-1)^{-1}\sum_{t=2}^{T} X_t$.[3] Given $\hat{\phi}$, we can determine $\hat{\alpha} = \bar{X}_{(2)} - \hat{\phi}\bar{X}_{(1)}$ and finally

$$\hat{\mu} = \frac{\bar{X}_{(2)} - \hat{\phi}\bar{X}_{(1)}}{1 - \hat{\phi}}.$$

To compare these estimators to the Yule-Walker estimator, we note that for the AR(1) process that

$$\hat{\phi}^{\mathrm{YW}} = \frac{\hat{K}_X(1)}{\hat{K}_X(0)} = \frac{\sum_{t=2}^{T}(X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_{t=1}^{T}(X_t - \bar{X})^2},$$

which is very similar to the MLE estimator except that the MLE uses $\bar{X}_{(1)}$ and $\bar{X}_{(2)}$ that are adjusted for the end points of the time series. In the limit, the two are equivalent.

Similarly, for the mean, the Yule-Walker equations chooses $\hat{\mu}^{\mathrm{YW}} = \bar{X}$. In contrast, for the MLE, we have

$$\frac{\bar{X}_{(2)} - \hat{\phi}\bar{X}_{(1)}}{1 - \hat{\phi}} \approx \frac{\bar{X} - \hat{\phi}\bar{X}}{1 - \hat{\phi}} = \bar{X}.$$

For AR(p) processes, the MLE estimators can still be computed in a similar manor, but the equations are more complex. Still, conditioning on the starting values $X_1, \ldots, X_p$ allows for a reduction to linear regression.

Proof of Asymptotic Normality for Yule-Walker

3.1.2 Forecasting for AR processes

Another significant problem in time series analysis is that of forecasting. That is, given data $X_1, \ldots, X_T$, we want to compute the best predictions for subsequent time points $X_{T+1}, X_{T+2}, \ldots, X_{T+m}$. We won't initially assume that the process is autoregressive, but we will assume that X_t is stationary.

First, we can consider linear predictors, which are those of the form

$$\hat{X}_{T+m} = \alpha_0 + \sum_{t=1}^{T} \alpha_t X_t$$

[3] Note that this is just $\sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$ from linear regression.

where we want to make a good choice of parameters α_t. To do that, we minimize the squared error as usual:

$$\underset{\alpha_1,\ldots,\alpha_T}{\arg\min}\ \mathrm{E}\left\{\left(X_{T+m}-\alpha_0-\sum_{t=1}^{T}\alpha_t X_t\right)^2\right\}.$$

Taking the derivative with respect to each α_t gives a system of equations

$$\begin{aligned}
0&=\mathrm{E}\left(X_{T+m}-\hat{X}_{T+m}\right)\\
0&=\mathrm{E}\left((X_{T+m}-\hat{X}_{T+m})X_1\right)\\
&\vdots\\
0&=\mathrm{E}\left((X_{T+m}-\hat{X}_{T+m})X_T\right)
\end{aligned}$$

Let the mean of the process be μ. By the first equation,

$$\mu=\mathrm{E}\left(X_{T+m}\right)=\mathrm{E}\left(\hat{X}_{T+m}\right)=\mathrm{E}\left(\alpha_0+\sum_{t=1}^{T}\alpha_t X_t\right)=\alpha_0+\sum_{t=1}^{T}\alpha_t\mu$$

$$\alpha_0=\mu\left(1-\sum_{t=1}^{T}\alpha_t\right)$$

Hence, we have $\hat{X}_{t+m}=\mu+\sum_{t=1}^{T}\alpha_t(X_t-\mu)$. Thus, we can centre the process and consider time series with $\mu=0$ and $\alpha_0=0$.

For a one-step-ahead prediction, which is to estimate $\hat{X}_{T+1}$, we solve the above equations to get

$$\begin{aligned}
0&=\mathrm{E}\left((X_{T+1}-\hat{X}_{T+1})X_1\right)=K_X(T)-\sum_{t=1}^{T}\alpha_t K_X(t-1)\\
0&=\mathrm{E}\left((X_{T+1}-\hat{X}_{T+1})X_2\right)=K_X(T-1)-\sum_{t=1}^{T}\alpha_t K_X(t-2)\\
&\vdots\\
0&=\mathrm{E}\left((X_{T+m}-\hat{X}_{T+m})X_T\right)=K_X(1)-\sum_{t=1}^{T}\alpha_t K_X(T-t)
\end{aligned}$$

If similar to before we let K be the T-long vector with entries $K_X(T),\ldots,K_X(1)$ and let

$$\Gamma=\begin{pmatrix}
K_X(0) & K_X(1) & \cdots & K_X(T-1)\\
K_X(1) & K_X(0) & \cdots & K_X(T-2)\\
\vdots & \ddots & \ddots & \vdots\\
K_X(T-1) & K_X(T-2) & \cdots & K_X(0)
\end{pmatrix}.$$

Then, the above equations can be written as $K = \Gamma\alpha$ or $\alpha = \Gamma^{-1}K$ in the case that the inverse exists. Thus, for $X = (X_1, \ldots, X_T)^{\mathrm{T}}$, our one-step prediction can be written as

$$\hat{X}_{T+1} = \alpha^{\mathrm{T}} X = K^{\mathrm{T}}\Gamma^{-1}X.$$

As like estimation for the AR process with the Yule-Walker equations, our prediction is based on the autocovariances. If we knew what the autocovariance is—i.e. we use K and Γ instead of $\hat{K}$ and $\hat{\Gamma}$—then the mean squared prediction error is

$$\begin{aligned}
\mathrm{E}\left(X_{T+1} - \hat{X}_{T+1}\right)^2 &= \mathrm{E}\left(X_{T+1} - K^{\mathrm{T}}\Gamma^{-1}X\right)^2 \\
&= \mathrm{E}\left(X_{T+1}^2 - 2K^{\mathrm{T}}\Gamma^{-1}XX_{T+1} + K^{\mathrm{T}}\Gamma^{-1}XX^{\mathrm{T}}\Gamma^{-1}K\right) \\
&= K_X(0) - 2K^{\mathrm{T}}\Gamma^{-1}K + K^{\mathrm{T}}\Gamma^{-1}\Gamma\Gamma^{-1}K \\
&= K_X(0) - K^{\mathrm{T}}\Gamma^{-1}K.
\end{aligned}$$

For the AR(p) process,

$$X_t = w_t + \sum_{i=1}^{p} \phi_i X_{t-i},$$

the one-step-ahead prediction comes precisely from estimating the coefficents as in the Yule-Walker equations to get

$$\hat{X}_{T+1} = \sum_{i=1}^{p} \hat{\phi}_i X_{T+1-i}.$$

However, if we do not know a priori that we have an order-p process, then we would have to estimate α_i for all $i = 1, \ldots, T$, which could require the inversion of a very large matrix. Thus, for general ARMA models, we have to work harder.

The Durbin-Levinson Algorithm

The system of equations $\alpha = \Gamma^{-1}K$ and computation of the mean squared prediction error for the one-step ahead estimate,

$$P_{T+1}^{T} := \mathrm{E}\left(X_{T+1} - \hat{X}_{T+1}\right)^2 = K_X(0) - K^{\mathrm{T}}\Gamma^{-1}K,$$

can be solved iteratively. This is because, Γ is a *Toeplitz* Matrix, and the Durbin-Levinsion Algorithm can be used to solve a system of equations involving a Toeplitz

matrix. To do this, we need to consider a sequence of one-step-ahead predictors

$$\begin{aligned}
\hat{X}_2^1 &= \alpha_{1,1}X_1 \\
\hat{X}_3^2 &= \alpha_{2,1}X_1 + \alpha_{2,2}X_2 \\
\hat{X}_4^3 &= \alpha_{3,1}X_1 + \alpha_{3,2}X_2 + \alpha_{3,3}X_3 \\
&\vdots \\
\hat{X}_{T+1}^T &= \alpha_{T,1}X_1 + \alpha_{T,2}X_2 + \dots \alpha_{T,T}X_T.
\end{aligned}$$

We begin recursively with $\alpha_{0,0} = 0$ and $P_0^1 = K_X(0)$, which is that the MSPE given no information is just the variance. Then, given coefficients $\alpha_{t,1}, \dots, \alpha_{t,t}$, we can compute

$$\alpha_{t+1,1} = \frac{\rho_X(t) - \sum_{i=1}^{t-1} \alpha_{t-1,i}\rho_X(i)}{1 - \sum_{i=1}^{t-1} \alpha_{t-1,i}\rho_X(t-i)}$$

and $P_{t+1}^t = P_t^{t-1}(1 - \alpha_{t+1,1}^2)$ and for the remaining coefficients $\alpha_{t+1,i} = \alpha_{t,t-i-1} - \alpha_{t+1,1}\alpha_{t,i}$

The Innovations Algorithm

The *innovations* for a time series are the residuals for the one-step-ahead estimate, $X_t - \hat{X}_t^{t-1}$. First, note that $X_t - \hat{X}_t^{t-1}$ and $X_s - \hat{X}_s^{s-1}$ are uncorrelated for $s \neq t$. The the innovations algorithm for calculating the one-step-ahead prediction $\hat{X}_{T+1}^T$ is as follows.

First initialize $X_1^0 = 0$ and $P_1^0 = K_X(0)$. Then, given the past observations $X_t, \dots, X_1$ and past one-step predictions $X_t^{t-1}, \dots, X_1^0$, we compute

$$\begin{aligned}
X_{t+1}^t &= \sum_{j=1}^{t} \theta_{t,j}(X_{t+1-j} - X_{t+1-j}^{t-j}) \\
P_{t+1}^t &= K_X(0) - \sum_{j=0}^{t-1} \theta_{t,t-j}^2 P_{j+1}^j
\end{aligned}$$

where

$$\theta_{t,t-j} = \left(K_X(t-j) - \sum_{k=0}^{j-1} \theta_{j,j-k}\theta_{t,t-k}P_{k+1}^k \right) (P_{j+1}^j)^{-1}$$

The innovations algorithm is useful for computing predictions for ARMA(p,q) processes specifically for the MA part. To see this, we consider a simple example: the MA(1) process.

Let $X_t = \theta w_{t-1} + w_t$. Then, we know that the autocovariance is $K_X(0) = \sigma^2(1 + \theta^2)$, $K_X(1) = \sigma^2\theta$, and $K_X(h) = 0$ for $h \geq 2$. Then, we have that $\theta_{0,0} = 1$

and

$$\begin{aligned}
\theta_{1,1} &= [K_X(1) - 0]/[K_X(0) - 0] = \theta/(1 + \theta^2) = \sigma^2\theta/P_1^0 \\
\theta_{2,2} &= [K_X(2) - 0]/[K_X(0) - 0] = 0 \\
\theta_{2,1} &= [K_X(1) - \theta_{1,1}\theta_{2,2}P_1^0]/[P_2^1] = [\sigma^2\theta - 0]/[P_2^1] \\
&\vdots \\
\theta_{t,j} &= 0, \quad \text{for } j > 1 \\
\theta_{t,1} &= [K_X(1) - 0]/[P_t^{t-1}] = \sigma^2\theta/[P_t^{t-1}]
\end{aligned}$$

We can also update the MSPEs as $P_{t+1}^t = (1 + \theta^2 - \theta\theta_{t,1})\sigma^2$. Therefore, the one-step-ahead prediction is

$$X_{t+1}^t = \theta_{t,1}(X_t - X_t^{t-1}) = \theta(X_t - X_t^{t-1})\sigma^2/P_t^{t-1}$$

3.2 The ARMA Process

3.2.1 Estimation for ARMA processes

For an ARMA(p,q) process, we have parameters $\mu, \sigma^2, \phi_1, \ldots, \phi_p, \theta_1, \ldots \theta_q$ to estimate. To estimate terms for an ARMA process, we revisit the maximum likelihood approach from above for the AR process. Note first that to do this, we assume that the white noise is Gaussian so that we have a distribution for the likelihood equation. Similar to before, we need to carefully rewrite the likelihood to make it tractible. This time, we consider conditioning the tth time point on the previous $t-1$ time points. That is,

$$L(\mu, \sigma^2, \phi, \theta) = \prod_{t=1}^{T} f(X_t \,|\, X_{t-1}, \ldots, X_1),$$

which means that we will consider each X_t predicted by the previous observations $X_{t-1}, \ldots, X_1$.

As we assume we have a causal invertiable ARMA(p.q) process, we can write it as a linear process in the form

$$X_t = \sum_{i=0}^{\infty} \psi_i w_{t-i}$$

where the infinite past will be convenient to assume even if the data is finite. The distribution of $X_t|X_{t-1}\ldots X_1$ will be normal with mean $\hat{X}_t^{t-1}$, the one-step-ahead prediction. The variance with by $\mathrm{E}\left((X_t - \hat{X}_t^{t-1})^2\right)$, which is also the mean squared

prediction error P_t^{t-1}. In the case of $t = 1$, we just use the variance for the linear process

$$K_X(0) = \sigma^2 \sum_{i=1}^{\infty} \psi_i^2.$$

From there, we can use the Durbin-Levinson Algorithm to update the MSPE by $P_{t+1}^t = P_t^{t-1}(1-\alpha_{t+1,1}^2)$. The precise computation is not important for our purposes, but we can write $P_t^{t-1} = \sigma^2 r_t$ where r_t does not depend on σ^2. This allows us to write the likelihood as

$$L(\mu, \sigma^2, \phi, \theta) = (2\pi\sigma^2)^{-T/2} \left[\prod_{t=1}^{T} r_t\right]^{-1/2} \exp\left(-\frac{S(\mu, \phi, \theta)}{2\sigma^2}\right)$$

with the sum of squares $S(\mu, \phi, \theta) = \sum_{t=1}^{T}(X_t - \hat{X}_t^{t-1})^2/r_t$.

From all of this, we can get the MLE for the variance $\hat{\sigma}^2 = S(\hat{\mu}, \hat{\phi}, \hat{\theta})/n$ as a function of the other estimators. To find those estimators, we can maximize the *concentrated* likelihood, which is when we replace $\hat{\sigma}^2$ with $S(\hat{\mu}, \hat{\phi}, \hat{\theta})/n$ and solve for the MLEs for $\hat{\mu}, \hat{\phi}, \hat{\theta}$. That is, for some constant C,

$$\log L(\mu, \phi, \theta) = C - \frac{T}{2}\log\hat{\sigma}^2 - \frac{1}{2}\sum_{t=1}^{T}\log r_t, \quad \text{or}$$

$$\ell(\mu, \phi, \theta) = \log\left(\frac{S(\mu, \phi, \theta)}{T}\right) + \frac{1}{T}\sum_{t=1}^{T}\log r_t.$$

We could check that for AR(p) processes—that is, without any MA part—that we recover the MLE estimates from before.

Asymptotic Distribution

Similar to the Yule-Walker equations for the AR process, we have a central limit-like theorem for the MLE estimator for the ARMA process. If we let $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ then as $T \to \infty$

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{\text{d}} \mathcal{N}\left(0, \sigma^2\Gamma_{p,q}^{-1}\right)$$

where $\Gamma_{p,q}$ is a $(p+q) \times (p+q)$ matrix with block form

$$\Gamma_{p,q} = \begin{pmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta} \end{pmatrix}$$

where the i, jth entry in $\Gamma_{\phi\phi}$ is $K_Y(i-j)$ for the AR process $\Phi(B)Y_t = w_t$, and the i, jth entry in $\Gamma_{\theta\theta}$ is $K_{Y'}(i-j)$ for the AR process $\Theta(B)Y_t' = w_t$, and the i, jth entry in $\Gamma_{\phi\theta}$ is the cross covariance between Y and Y'.

Example 3.2.1 (AR(1))**.** *For the casual AR(1) process $X_t = \phi X_{t-1} + w_t$, we recall that $K_X(0) = \sigma^2/(1-\phi^2)$. Therefore, $\Gamma_{1,0} = (1-\phi^2)^{-1}$ and we have that*

$$\hat{\phi} \xrightarrow{d} \mathcal{N}\left(\phi, \frac{1}{T}(1-\phi^2)\right).$$

Example 3.2.2 (MA(1))**.** *Similar to the AR(1) prcess, consider the invertible MA(1) process $X_t = \theta w_{t-1} + w_t$. The MA polynomial is $\Theta(B) = 1 + \theta B$, so the AR(1) process $\Theta(B)Y_t = w_t$ has a variance of $K_Y(0) = \sigma^2/(1-\theta^2)$. Thus, we have that*

$$\hat{\theta} \xrightarrow{d} \mathcal{N}\left(\theta, \frac{1}{T}(1-\theta^2)\right).$$

Note that similar to linear regression and many other areas of statistics, if we include too many terms when fitting an ARMA process to a dataset, the standard errors of the estimate will be larger than necessary. Thus, it is typically good to keep the number of parameters as small as possible. Hence, the use of AIC and BIC in the R function `arima()`.

3.2.2 Forecasting for ARMA processes

We already discussed approaches to forecasting that can be applied to the ARMA(p,q) process. In this section, we take a closer look at forecasting for the ARMA process and the behaviour of these predicted values. As usual, we will assume that the ARMA(p,q) process written in operator form as

$$\Phi(B)X_t = \Theta(B)w_t$$

is both causal and invertible as well as mean zero.[4] We will also assume that we are making a prediction based on observed data $X_1, \ldots, X_T$.

Given a sample size T, there are two possible estimators for the future point X_{T+h} to consider. The prediction that minimizes the mean squared error is

$$\hat{X}_{T+h} = \mathrm{E}\left(X_{T+h} \,|\, X_T, \ldots, X_1\right).$$

However, it is often mathematically convenient to consider the estimator based on the infinite past, which is

$$\tilde{X}_{T+h} = \mathrm{E}\left(X_{T+h} \,|\, X_T, \ldots, X_1, X_0, X_{-1}, \ldots\right).$$

As the data size T gets large, these two predictions are very close.

[4] As usual, we can subtract the mean to achieve this last assumption.

As we assume that that ARMA process is both causal and invertible, we can rewrite it in two different forms:

$$X_{T+h} = w_{T+h} + \sum_{j=1}^{\infty} \psi_j w_{T+h-j} \qquad \text{(Causal)}$$

$$w_{T+h} = X_{T+h} + \sum_{j=1}^{\infty} \pi_j X_{T+h-j} \qquad \text{(Invertible)}$$

and we can consider the above conditional expectations applied to each of these equations.

First, we note that

$$\mathrm{E}\left(w_t | X_T, \ldots, X_0, \ldots\right) = \begin{cases} w_t & \text{if } t \leq T \\ 0 & \text{if } t > T \end{cases}$$

This is because (1) if $t > T$ then w_t is independent of the sequence of $X_T, \ldots$ and (2) if $t \leq T$ then based on the casual and invertible representations, we have a one to one correspondence between the X's and the w's. Similarly,

$$\mathrm{E}\left(X_t | X_T, \ldots, X_0, \ldots\right) = \begin{cases} X_t & \text{if } t \leq T \\ \tilde{X}_t & \text{if } t > T \end{cases}$$

Applying this idea to the causal representation, we get that

$$\tilde{X}_{T+h} = \sum_{j=h}^{\infty} \psi_j w_{T+h-j}$$

as the first $h-1$ terms in the sum become zero. Then, subtracting this from X_{t+h} gives

$$X_{T+h} - \tilde{X}_{T+h} = \sum_{j=0}^{h-1} \psi_j w_{T+h-j}.$$

Hence, the mean squared prediction error is

$$P_{T+h}^{T} = \mathrm{E}\left((X_{T+h} - \tilde{X}_{T+h})^2\right) = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2.$$

Note that we can also apply the conditional expectation to the invertible representation. In that case, we get

$$0 = \tilde{X}_{T+h} + \sum_{j=1}^{h-1} \pi_i \tilde{X}_{T+h-j} + \sum_{j=h}^{\infty} \pi_i X_{T+h-j}$$

$$\tilde{X}_{T+h} = -\sum_{j=1}^{h-1} \pi_i \tilde{X}_{T+h-j} - \sum_{j=h}^{\infty} \pi_i X_{T+h-j}.$$

This shows that the $T+h$ predicted value is a function of the data $X_T, \ldots$ and the previous $h-1$ predicted values $\tilde{X}_{T+h-1}, \ldots, \tilde{X}_{T+1}$.

Long Range Forecast Behaviour

What happens if we try to predict far into the future? If we consider an ARMA(p,q) process with mean μ, then the h-step-ahead estimator based on the infinite past is

$$\tilde{X}_{T+h} = \mu + \sum_{j=h}^{\infty} \psi_j w_{t+h-j}.$$

Now, we know from before that the coefficients ψ_i tend to zero fast enough to be absolutely summable—i.e. $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Therefore, $\sum_{j=h}^{\infty} |\psi_j| \to 0$ as $h \to \infty$. This implies that

$$\tilde{X}_{T+h} \xrightarrow{\text{P}} \mu \text{ as } h \to \infty.$$

We can prove this first by noting that the variance of $\tilde{X}_{T+h}$ is $\sum_{j=h}^{\infty} \psi_j^2$. Then, for any $\varepsilon > 0$, we can use Chebyshev's inequality to get that

$$\text{P}\left(|\tilde{X}_{T+h} - \mu| > \varepsilon\right) = \text{P}\left(\left|\sum_{j=h}^{\infty} \psi_j w_{t+h-j}\right| > \varepsilon\right) \leq \varepsilon^{-2} \sum_{j=h}^{\infty} \psi_j^2 \to 0$$

as $h \to \infty$.

Meanwhile, the MSPE from above is

$$P_{T+h}^{T} = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2.$$

Therefore, as $h \to \infty$, we have that the MSPE tends to $K_X(0)$, which is just the variance of the process X_t.

Hence, in the long run, the forecast for an ARMA(p,q) process tends towards its mean, and the variance tends to the variance of the process.

Truncating the infinite past

For a small sample size T, we can forecast by solving the system of equations presented above by inverting the $T \times T$ matrix Γ. For a large sample size T, we can use the recursive approach to forecast. However, it is worth considering what the effect is of not having access to the past time points $X_0, X_{-1}, \ldots$ before the dataset was collected.

Using the invertible respresentation of the time series, we have the truncated h-step-ahead prediction

$$\tilde{X}_{T+h}^{T} = -\sum_{j=1}^{h-1} \pi_i \tilde{X}_{T+h-j}^{T} - \sum_{j=h}^{T} \pi_i X_{T+h-j}.$$

Given the coefficients ϕ_i and θ_i, we can write this truncated prediction as

$$\tilde{X}_{T+h}^T = \sum_{i=1}^{p} \phi_i \tilde{X}_{T+h-i}^T + \sum_{j=1}^{q} \theta_j \tilde{w}_{T+h-j}^T$$

where we replace the predicted value $\tilde{X}_{T+h-i}^T$ with the observed value X_{T+h-i} if $i \in [h, T+h-1]$ and with 0 if $i \geq T+h$. Similarly, $\tilde{w}_t^T = 0$ if $t < 1$ or if $t > T$. Otherwise,

$$\tilde{w}_t^T = \Phi(B)\tilde{X}_t^T - \sum_{j=1}^{q} \theta_j \tilde{w}_{t-j}.$$

To see all of this in action, we can consider a few simple examples.

Example 3.2.3 (ARMA(1,1))**.** *For the causal invertible ARMA(1,1) process $X_{t+1} = \phi X_t + w_{t+1} + \theta w_t$, we can consider the one-step-ahead prediction*

$$\tilde{X}_{T+1}^T = \phi X_T + \theta \tilde{w}_T^T$$

and the h-step-ahead prediction $\tilde{X}_{T+h}^T = \phi X_{T+h-1}^T$ for $h \geq 2$.

Hence, to forecast for the ARMA(1,1) process, we only need X_T and the estimate $\tilde{w}_T^T$. For the error term $\tilde{w}_T^T$, we have that

$$\begin{aligned} w_{t+1} &= X_{t+1} - \phi X_t - \theta w_t \\ \tilde{w}_{t+1}^T &= X_{t+1} - \phi X_t - \theta \tilde{w}_t^T. \end{aligned}$$

Hence, we can start from $\tilde{w}_0^T = 0$ and $X_0 = 0$ and compute the $\tilde{w}_{t+1}^T$ iteratively.

We can also compute the variance of the prediction (the MSPE). For this, we note that the ARMA(1,1) process can be writen in a causal form as

$$X_t = w_t + \sum_{i=1}^{\infty} \psi_i w_{t-i} = w_t + \sum_{i=1}^{\infty} (\phi + \theta)\phi^{i-1} w_{t-i}.$$

Then, we have that

$$\begin{aligned} P_{T+h}^T = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2 &= \sigma^2 \left[1 + (\phi+\theta)^2 \sum_{j=1}^{h-1} \phi^{2j-2} \right] \\ &= \sigma^2 \left[1 + (\phi+\theta)^2 \left(\frac{1-\phi^{2h-2}}{1-\phi^2} \right) \right] \to \sigma^2 \left[1 + \frac{(\phi+\theta)^2}{1-\phi^2} \right] \end{aligned}$$

as $h \to \infty$.

Backcasting

We can also consider forecasting into the past or *backcasting*. That is, we can predict backwards h time units into the past by

$$\hat{X}_{1-h}^{T} = \sum_{i=1}^{T} \alpha_i X_i$$

for some coefficients α_t. To do this, we proceed as before by considering for $t = 1, \ldots, T$

$$\mathrm{E}\left(X_{1-h} X_t\right) = \sum_{i=1}^{T} \alpha_i \mathrm{E}\left(X_i X_t\right)$$
$$K_X(t+h-1) = \sum_{i=1}^{T} \alpha_i K_X(t-i).$$

This means we can compute the coefficients α_i by solving the system of equations $K = \Gamma\alpha$ where

$$K = \begin{pmatrix} K_X(h) \\ K_X(h+1) \\ \vdots \\ K_X(T+h-1) \end{pmatrix}, \quad \Gamma = \begin{pmatrix} K_X(0) & K_X(1) & \cdots & K_X(T-1) \\ K_X(1) & K_X(0) & \cdots & K_X(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ K_X(T-1) & K_X(T-2) & \cdots & K_X(0) \end{pmatrix}$$

just as we did before for forecasting.

Remark 3.2.4 (Fun Fact)**.** *For a stationary Gaussian process, the vector* $(X_{T+1,X_T,\ldots,X_1})$ *is equal in distribution to* $(X_{0,X_1,\ldots,X_T})$*, so forecasting and backcasting are equivalent.*

3.3 Seasonal ARIMA

Very often with time series, there is a strong seasonal component to the data. For example, temperatures and other climate measurements have annual cycles. Financial data may have annual or quarterly cycles. For another example, electricity consumption may have both annual cycles and daily cycles—we use more electricity when we are awake than when we are asleep, and we use more electricity in the winter when it is dark and we are prone to staying inside, for example.

Thus, it is often beneficial to consider modelling time series data are certain lags based on the seasonality of the data. For example, instead of considering an AR(1) process

$$X_t = \phi X_{t-1} + w_t,$$

we could consider an annual AR(1) process

$$X_t = \phi X_{t-12} + w_t$$

or more generally, $X_t = \phi X_{t-s} + w_t$ which we will call an AR$(1)_s$ process for some value of $s > 1$.

In general, we can consider a seasonal ARMA process denoted AMRA$(p', q')_s$ which is

$$\Phi_s(B^s)X_t = \Theta_s(B^s)w_t$$

where the polynomials Φ_s and Θ_s are

$$\begin{aligned}\Phi_s(B^s) &= 1 - \varphi_1 B^s - \varphi_2 B^{2s} - \ldots - \varphi_{p'} B^{p's} \\ \Theta_s(B^s) &= 1 + \vartheta_1 B^s + \vartheta_2 B^{2s} + \ldots + \vartheta_{q'} B^{q's}\end{aligned}$$

The reason for the notation is to combine the seasonal ARMA with the regular ARMA process to get an ARMA$(p, q) \times (p', q')_s$ process, which can be written as

$$\Phi_s(B^s)\Phi(B)X_t = \Theta_s(B^s)\Theta(B)w_t.$$

If we were to include a differencing operator as in the ARIMA model to account for non-stationarity, we get the Seasonal ARIMA or SARIMA$(p, d, q) \times (p', d', q')_s$ model. This takes on the form

$$\Phi_s(B^s)\Phi(B)\nabla_s^{d'}\nabla^d X_t = \Theta_s(B^s)\Theta(B)w_t.$$

where $\nabla^d = (1 - B)^d$ and $\nabla_s^{d'} = (1 - B^s)^{d'}$. The large number of parameters is why the `auto.arima()` function in the R package `forecast` takes so long to run when the seasonal component is included.

3.3.1 Seasonal Autoregressive Processes

We can consider a purely seasonal AR process such as an annual AR(1) like

$$\begin{aligned}(1 - \varphi B^{12})X_t &= w_t \\ X_t &= \varphi X_{t-12} + w_t\end{aligned}$$

where $|\phi| < 1$. To compute the autocovariance quickly, we can rewrite this as a linear process

$$\begin{aligned}X_t &= \varphi X_{t-12} + w_t \\ &= \varphi(\varphi X_{t-24} + w_{t-12}) + w_t \\ &= \varphi^2 X_{t-24} + \varphi w_{t-12} + w_t \\ &\vdots \\ &= \sum_{j=0}^{\infty} \varphi^j w_{t-12j}.\end{aligned}$$

Therefore, the variance is as usual $K_X(0) = \sigma^2/(1-\phi^2)$. For lags $h = 1, \ldots, 11$, we have

$$K_X(h) = \text{cov}\left(\sum_{j=0}^{\infty} \varphi^j w_{t-12j}, \sum_{i=0}^{\infty} \varphi^i w_{t-h-12i}\right)$$
$$= \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} \varphi^{j+i}\text{cov}\left(w_{t-12j}, w_{t-h-12i}\right) = 0,$$

because the indices $t-12j$ and $t-h-12i$ will never be equal unless h is a multiple of 12. In that case, we have

$$K_X(12) = \sigma^2\varphi\sum_{j=0}^{\infty}\varphi^{2j} = \frac{\sigma^2\varphi}{1-\varphi^2}.$$

Note that this is the same as $K_Y(1)$ for the AR(1) process $Y_t = \varphi Y_{t-1} + w_t$. Hence, the above seasonal AR process is effectively 12 uncorrelated AR processes running in parallel to each other. This is why we often include a seasonal and non-seasonal component in the SARIMA models.

Note also that the $\text{AR}(1,0)_{12}$ could also be thought of as an $\text{AR}(12,0)$. However, trying to estimate or forecast with an $\text{AR}(12,0)$ process will include many parameters that are unnecessary, which will increase the variance for our estimators and predictions.

3.3.2 Seasonal ARMA Processes

We can take the above process and add in an MA(1) process to get the $\text{ARMA}(0,1)\times(1,0)_{12}$ which is

$$(1-\varphi B^{12})X_t = \theta w_{t-1} + w_t$$

with $|\theta| < 1$ and $|\varphi| < 1$. We can compute the variance as usual by noting that the MA piece is based on w_{t-1}, which is uncorrelated with the $\text{AR}(1)_{12}$ piece to get

$$K_X(0) = \text{Var}\left(X_t\right) = \text{Var}\left(\varphi X_{t-12} + \theta w_{t-1} + w_t\right) = \varphi^2 K_X(0) + \sigma^2(\theta^2+1),$$

which gives that

$$K_X(0) = \sigma^2\frac{1+\theta^2}{1-\varphi^2}.$$

For the autocovariances at other lags, we first note that if $h = 12m$ is a multiple of 12, then

$$K_X(h) = K_X(12m) = \frac{\sigma^2\varphi^m}{1-\varphi^2}$$

as above, since the MA piece will not affect the calculation. However, if $h = 1 \mod 12$ or $h = 11 \mod 12$, we have to work harder. First, consider that

$$\begin{aligned} K_X(1) &= \mathrm{E}\left(X_{t-1}\left[\varphi X_{t-12} + \theta w_{t-1} + w_t\right]\right) \\ &= \varphi K_X(11) + \sigma^2\theta, \quad \text{and} \\ K_X(11) &= \mathrm{E}\left(X_{t-11}\left[\varphi X_{t-12} + \theta w_{t-1} + w_t\right]\right) \\ &= \varphi K_X(1). \end{aligned}$$

From this, we have that

$$K_X(1) = \frac{\sigma^2\theta}{1-\varphi^2} \quad \text{and} \quad K_X(11) = \frac{\sigma^2\theta\varphi}{1-\varphi^2}$$

Continuing on, we have that

$$K_X(13) = \varphi K_X(1) = \frac{\sigma^2\theta\varphi}{1-\varphi^2}$$

as well. Hence, we can generalize this to

$$K_X(h) = K_X(12m \pm 1) = \frac{\sigma^2\theta\varphi^m}{1-\varphi^2}.$$

Lastly, for any lags $h \neq -1, 0, 1 \mod 12$, we have $K_X(h) = 0$ as none of the indices line up in the autocovariance computation.

Chapter 4

Analysis in the Frequency Domain

Introduction

Time series data often exhibits cyclic behaviour as we saw with SARIMA models in the previous chapter. Furthermore, a time series may have more than one cycle occurring simultaneously. In this chapter, we will consider spectral methods for identifying the cyclic behaviour of time series data.

In general, we are interested in the frequency ω of the time series. For example, a time series that repeats every 12 months, the frequency is $\omega = 1/12$.

4.1 Periodic Processes

We will consider the periodic process

$$X_t = A\cos(2\pi\omega t + \phi)$$

where A is the amplitude, ω is the frequency, and ϕ is the phase. This process can be rewritten as linear combination of trig functions as

$$X_t = U_1\cos(2\pi\omega t) + U_2\sin(2\pi\omega t)$$

where $U_1 = A\cos(\phi)$ and $U_2 = -A\sin(\phi)$. This is due to the identity $\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y)$.

If we take U_1 and U_2 to be uncorrelated mean zero random variables with vari-

ance σ^2, then we can compute the autocovariance as

$$\begin{aligned}
K_X(h) &= \\
&= \operatorname{cov}\left(U_1\cos(2\pi\omega(t+h)) + U_2\sin(2\pi\omega(t+h)), U_1\cos(2\pi\omega t) + U_2\sin(2\pi\omega t)\right) \\
&= \operatorname{cov}\left(U_1\cos(2\pi\omega(t+h)), U_1\cos(2\pi\omega t)\right) \\
&\quad + \operatorname{cov}\left(U_2\sin(2\pi\omega(t+h)), U_2\sin(2\pi\omega t)\right) \\
&= \sigma^2\cos(2\pi\omega(t+h))\cos(2\pi\omega t) + \sigma^2\sin(2\pi\omega(t+h))\sin(2\pi\omega t) \\
&= \sigma^2\cos(2\pi\omega h),
\end{aligned}$$

so depending on the lag h, the autocovariance with rise or fall.

We can combine q different frequencies $\omega_1, \ldots, \omega_q$ to get a more general period process

$$X_t = \sum_{j=1}^{q} \left\{U_{j1}\cos(2\pi\omega_j t) + U_{j2}\sin(2\pi\omega_j t)\right\}$$

where all of the U_{j1} and U_{j2} are uncorrelated mean zero random variance with potentially different variances σ_j^2. The autocovariance in this case is

$$K_X(h) = \sum_{j=1}^{q} \sigma_j^2\cos(2\pi\omega_j h).$$

Remark 4.1.1 (Aliasing)**.** *Aliasing is a problem that can occur when taking a discrete sample from a continuous signal. Since we have to sample a certain rate, high frequency behaviour in the signal may look like low frequency patterns in our sample. This is displayed in Figure 4.1 where the red dots are sampled too infrequently making it appear as if there is a low frequency oscilation in the data instead of the actually high frequency oscilation in black.*

4.1.1 Regression, Estimation, and the FFT

Given some time series data $X_1, \ldots, X_T$ with T odd, then we can exactly represent the data as

$$X_t = \beta_0 + \sum_{j-1}^{(T-1)/2} \left\{\beta_{j1}\cos(2\pi t\, j/T) + \beta_{j2}\sin(2\pi t\, j/T)\right\}.$$

This is because the sin's and cos's form a basis, and similarly to how a $T-1$ degree polynomial can pass through T points, these $T-1$ parameters $\beta_\star$ can be used to fit the data exactly. Note that for T even, we can also do this with

$$X_t = \beta_0 + \sum_{j-1}^{T/2-1} \left\{\beta_{j1}\cos(2\pi t\, j/T) + \beta_{j2}\sin(2\pi t\, j/T)\right\} + \beta_{T/2}\cos(\pi t).$$

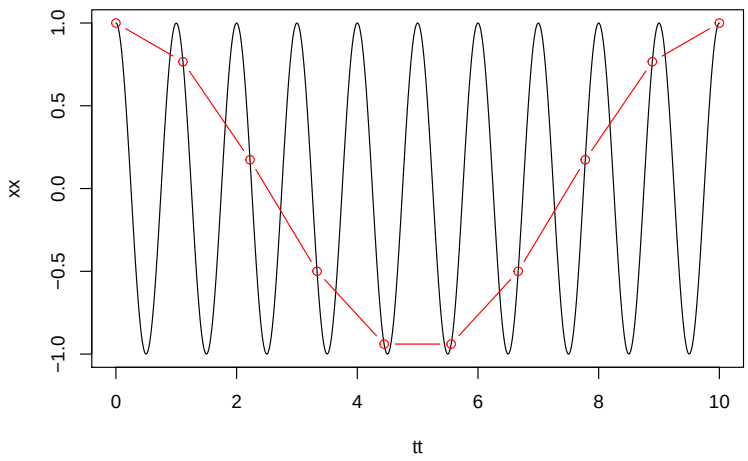

Figure 4.1: Aliasing occurs when we sample too infrequently to capture high frequency oscilations.

Of course, in practice, we do not want to include $T-1$ parameters to fit our data exactly. Instead, we assume that most of these β will be zero and the only non-zero β's will correspond to prominent frequencies in the time series.

We can estimate all of the β's by treating this as a linear regression problem.[1]

[1] Recall that for $Y = X\beta + \varepsilon$, if $X^{\mathrm{T}}X = cI_n$, then $\hat{\beta} = X^{\mathrm{T}}Y/c$.

First, with a little work, we can show that

$$\begin{aligned}
\sum_{t=1}^{T} \cos^2(2\pi tj/T) = \sum_{t=1}^{T} \sin^2(2\pi tj/T) = n/2 \qquad & \text{for } j = 1, \dots T/2 - 1 \\
\sum_{t=1}^{T} \cos^2(2\pi tj/T) = n \qquad & \text{for } j = 0, T/2 \\
\sum_{t=1}^{T} \sin^2(2\pi tj/T) = 0 \qquad & \text{for } j = 0, T/2 \\
\sum_{t=1}^{T} \cos(2\pi tj/T)\cos(2\pi tk/T) = 0 \qquad & \text{for } j \neq k \\
\sum_{t=1}^{T} \cos(2\pi tj/T)\cos(2\pi tk/T) = 0 \qquad & \text{for } j \neq k \\
\sum_{t=1}^{T} \cos(2\pi tj/T)\sin(2\pi tk/T) = 0 \qquad & \text{for any } j, k.
\end{aligned}$$

Hence, our design matrix for linear regression has orthongal columns, so computing each $\hat{\beta}$ becomes, for $j \neq 0, T/2$,

$$\begin{aligned}
\hat{\beta}_{j1} &= \frac{2}{T} \sum_{t=1}^{T} X_t \cos(2\pi tj/T) \\
\hat{\beta}_{j2} &= \frac{2}{T} \sum_{t=1}^{T} X_t \sin(2\pi tj/T)
\end{aligned}$$

and $\hat{\beta}_0 = \bar{X}$ and, if T is even, $\hat{\beta}_{T/2} = T^{-1} \sum_{t=1}^{T} (-1)^t X_t$.

Given these estimates, we can define the *scaled periodogram* $P(j/T) = \hat{\beta}_{j1}^2 + \hat{\beta}_{j2}^2$, which can be used to determine which frequencies are the most prominent in the time series X_t. Note that the variance of for $\beta_{j1}\cos(2\pi tj/T) + \beta_{j2}\sin(2\pi tj/T)$ is $\beta_{j1}^2 + \beta_{j2}^2$, so the periodogram is the sample variance for frequency j/T.

Computing all of these β as above is computationally infeasable for large T. However, if T is a highly composite integer—i.e. one with a lot of small integer factors like 2^m—then, we can use the *discrete Fourier transform*,

$$\begin{aligned}
d(j/T) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \exp(-2\pi itj/T) \\
&= \frac{1}{\sqrt{T}} \left\{ \sum_{t=1}^{T} X_t \cos(2\pi tj/T) - i \sum_{t=1}^{T} X_t \sin(2\pi tj/T) \right\}.
\end{aligned}$$

The squared magnitude of the coefficients

$$|d(j/T)|^2 = \frac{1}{T}\left\{\sum_{t=1}^{T} X_t \cos(2\pi tj/T)\right\}^2 + \frac{1}{T}\left\{\sum_{t=1}^{T} X_t \sin(2\pi tj/T)\right\}^2$$

is the (unscaled) periodogram. The scaled periodogram is $P(j/T) = (4/T)|d(j/T)|^2$, which follows from the equations for the $\hat{\beta}$ above. Noting that $\cos(2\pi - \theta) = \cos(\theta)$ and that $\sin(2\pi - \theta) = -\sin(\theta)$, we have that $|d(1 - j/T)|^2 = |d(j/T)|^2$ and likewise $P(1 - j/T) = P(j/T)$. Hence, we only consider frequencies $j/T < 1/2$.

The DFT can be computed quickly via the *Fast Fourier Transform* (FFT). The DFT is just a linear transformation of the data X_t, which can be written at $d = WX$ for some matrix W. This type of transformation would take $O(T^2)$ time to compute. However, the FFT uses a sparse representation of W to reduce the time to $O(T \log_2 T)$. There are many algorithms for the FFT, but the most common takes a divide-and-conquer approach. That is, if $T = 2^m$, then it breaks the data in half based on odd and even indices $X_1, X_3, \ldots, X_{T-1}$ and $X_2, X_4, \ldots, X_T$ and computes the Fourier transform of each separately. However, since $T/2 = 2^{m-1}$ is also divisible by 2, this idea can be applied recursively to get 4 partitions of the data, then 8, and so on.

If we rewrite the DFT as

$$\begin{aligned}\sqrt{T}d(j/T) &= \sum_{t=1}^{T/2} X_{2t} \mathrm{e}^{-\frac{2\pi itj}{T/2}} + \mathrm{e}^{-\frac{2\pi ij}{T}} \sum_{t=1}^{T/2} X_{2t-1} \mathrm{e}^{-\frac{2\pi itj}{T/2}} \\ &= E_j + \mathrm{e}^{-\frac{2\pi ij}{T}} O_j,\end{aligned}$$

then we can decompose it into even and odd parts E_j and O_j, respectively. These two pieces are each DFTs of size $T/2$. We also note that there is a redundency in the calculations, which is for $j < T/2$, we have

$$\sqrt{T}d(j/T) = E_j + \mathrm{e}^{-\frac{2\pi ij}{T}} O_j$$

and that

$$\sqrt{T}d(j/T + 1/2) = E_j - \mathrm{e}^{-\frac{2\pi ij}{T}} O_j.$$

Remark 4.1.2. *Scaling and the FFT In Fourier analysis and for different FFT implementations in code, there are often different scaling factors included. Hence, to make sure we are estimating what we want to estimate, one needs to take care when using FFT algorithms.*

4.2 Spectral Distribution and Density

We begin by presenting a way to represent the autocovariance function as described in the Wiener-Khintchine Theorem.[2] The theorem applies to continuous time pro-

[2] https://en.wikipedia.org/wiki/Wiener\OT1\textendashKhinchin_theorem

cesses. However, in this course, we only consider discrete time processes.

Theorem 4.2.1 (Wiener-Khintchine Theorem I)**.** *Let X_t be stationary with autocovariance $K_X(h) = \text{cov}(X_{t+h}, X_t)$. Then, there exists a unique monotonically increasing function $F_X(\omega)$, called the spectral distribution function, with $F_X(-1/2) = 0$ and $F_X(1/2) = K_X(0)$ such that*

$$K_X(h) = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h} dF_X(\omega)$$

where this is a Riemann-Stieltjes integral.

Given slightly stronger conditions, we can also define the spectral density. That is, if the autocovariance is absolutely summable, then the spectral distribution is absolutely continuous in turn implying that the derivative exists almost everywhere: $dF_X(\omega) = f_X(\omega)d\omega$.

Theorem 4.2.2 (Wiener-Khintchine Theorem II)**.** *Let X_t be stationary with autocovariance $K_X(h) = \text{cov}(X_{t+h}, X_t)$ such that*

$$\sum_{h=-\infty}^{\infty} |K_X(h)| < \infty$$

Then, we can write

$$K_X(h) = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h} f_X(\omega)d\omega.$$

Furthermore, we have the inverse transformation

$$f_X(\omega) = \sum_{h=-\infty}^{\infty} K_X(h)\mathrm{e}^{-2\pi i\omega h}$$

for $\omega \in [-1/2, 1/2]$.

From here we see that if $f_X(\omega)$ exists, then it is an even function—i.e. $f_X(\omega) = f_X(-\omega)$. Also, since $K_X(0) = \int_{-1/2}^{1/2} f_X(\omega)d\omega$, the variance of the process can be thought of as the integral of the spectral density over all frequencies. In a way, this is similar to how the total sum of squares can be decomposed into separate sums of squares in a ANOVA.

As a simple example, consider the period process $X_t = U_1 \cos(2\pi\omega_0 t) + U_2 \sin(2\pi\omega_0 t)$ from before. Then, the autocovariance is

$$K_X(h) = \sigma^2 \cos(2\pi\omega_0 h) = \frac{\sigma^2}{2}\left(\mathrm{e}^{2\pi\omega_0 h} + \mathrm{e}^{-2\pi\omega_0 h}\right) = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h} dF_X(\omega)$$

where $F_X(\omega) = 0$ for $\omega < -\omega_0$, $F_X(\omega) = \sigma^2/2$ for $\omega \in [-\omega_0, \omega_0]$, and $F_X(\omega) = \sigma^2$ for $\omega > \omega_0$. Note that in this case, the autocovariance is not absolutely summable.

As a second example, we consider the white noise process w_t. In this case, the autocovariance is simply σ^2 at lag $h = 0$ and 0 for all other lags h. Hence, it is absolutely summable and the spectral density is just $f_X(\omega) = \sigma^2$ for all $\omega \in [-1/2, 1/2]$. Hence, as mentioned in Chapter 1, white noise in a sense contains every frequency at once with equal power.

4.2.1 Filtering and ARMA

Given the spectral density for one time series X_t, we can find the spectral density for another related time series $y_t = \sum_{j=-\infty}^{\infty} a_j X_{t-j}$ for some fixed sequence a_j with $\sum_{j=-\infty}^{\infty} |a_j| < \infty$. Treating $a : \mathbb{Z} \to \mathbb{R}$ as a function, then $a(j) = a_j$ is called the *impulse response function* and its Fourier transform

$$A(\omega) = \sum_{j=-\infty}^{\infty} a_j \mathrm{e}^{-2\pi i \omega j}$$

is the *frequency response function.* Given all of this, we have the following theorem.

Theorem 4.2.3. *Let X_t be a time series with spectral density $f_X(\omega)$ and let $\sum_{j=-\infty}^{\infty} |a_j| < \infty$, then the spectral density for $y_t = \sum_{j=-\infty}^{\infty} a_j X_{t-j}$ is*

$$f_Y(\omega) = |A(\omega)|^2 f_X(\omega)$$

for $A(\omega)$ the frequency response function from above.

Proof. To prove the above theorem, we just compute the autocovariance directly.

$$\begin{aligned}
K_Y(h) &= \operatorname{cov}\left(\sum_{j=-\infty}^{\infty} a_j X_{t+h-j}, \sum_{l=-\infty}^{\infty} a_l X_{t-l} \right) \\
&= \sum_{j=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} a_j a_l K_X(h - j + l) \\
&= \sum_{j=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} a_j a_l \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i \omega (h-j+l)} f_X(\omega) d\omega \\
&= \int_{-1/2}^{1/2} \left[\sum_{j=-\infty}^{\infty} a_j \mathrm{e}^{-2\pi i \omega j} \right] \left[\sum_{l=-\infty}^{\infty} a_l \mathrm{e}^{2\pi i \omega l} \right] \mathrm{e}^{2\pi i \omega h} f_X(\omega) d\omega \\
&= \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i \omega h} |A(\omega)|^2 f_X(\omega) d\omega.
\end{aligned}$$

Therefore, by the uniqueness of the Fourier transform, we have that $f_Y(\omega) = |A(\omega)|^2 f_X(\omega)$. □

We can apply this result to a causal ARMA(p,q) process. For $\Phi(B)X_t = \Theta(B)w_t$, we have rewrite it as

$$X_t = \frac{\Theta(B)}{\Phi(B)} w_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}.$$

Writing $\Psi(z) = \Theta(Z)/\Phi(z) = \sum_{j=0}^{\infty} \psi_j z^j$ and using this as the a_j from the above theorem, we have that

$$A(\omega) = \sum_{j=-\infty}^{\infty} \psi_j \mathrm{e}^{-2\pi i \omega j} = \Psi(\mathrm{e}^{-2\pi i\omega}) = \frac{\Theta(\mathrm{e}^{-2\pi i\omega})}{\Phi(\mathrm{e}^{-2\pi i\omega})}.$$

Using the fact that $f_w(\omega) = \sigma^2$ for all ω, we have finally that

$$f_X(\omega) = |A(\omega)|^2 f_w(\omega) = \sigma^2 \left| \frac{\Theta(\mathrm{e}^{-2\pi i\omega})}{\Phi(\mathrm{e}^{-2\pi i\omega})} \right|^2.$$

4.3 Spectral Statistics

Thus far, our discussion of spectral analysis for time series has not considered the issue of working with noisy data. Given a finite set of data $X_1, \ldots, X_T$, we can compute the DFT

$$d(\omega_j) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \mathrm{e}^{-2\pi i \omega_j t}$$

for frequencies $\omega_j = j/T$. We can also compute the real and imaginary part separately being the sine and cosine transformations:

$$d_c(\omega_j) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \cos 2\pi\omega_j t$$

$$d_s(\omega_j) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \sin 2\pi\omega_j t$$

so that $d(\omega_j) = d_c(\omega_j) - i d_s(\omega_j)$.

We can compute the inverse DFT by

$$X_t = \frac{1}{\sqrt{T}} \sum_{j=0}^{T-1} d(\omega_j) \mathrm{e}^{2\pi i \omega_j t}$$

for $t = 1, \ldots, T$. This gives us the periodogram defined to be

$$I(\omega_j) = |d(\omega_j)|^2 = d_c(\omega_j)^2 + d_s(\omega_j)^2.$$

We can also centre the DFT when $j \neq 0$ to get

$$d(\omega_j) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (X_t - \bar{X}) \mathrm{e}^{-2\pi i \omega_j t}$$

as $\sum_{t=1}^{T} \mathrm{e}^{-2\pi i \omega_j t} = 0$ for any $\omega_j \neq 0$. This is useful as it allows us to write the periodogram for $j \neq 0$ as

$$\begin{aligned} I(\omega_j) &= \left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (X_t - \bar{X}) \mathrm{e}^{-2\pi i \omega_j t} \right|^2 \\ &= \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} (X_t - \bar{X})(X_s - \bar{X}) \mathrm{e}^{-2\pi i \omega_j (t-s)} \\ &= \frac{1}{T} \sum_{h=-T+1}^{T-1} \mathrm{e}^{-2\pi i \omega_j h} \sum_{t=1}^{T-|h|} (X_{t+|h|} - \bar{X})(X_t - \bar{X}) \\ &= \frac{1}{T} \sum_{h=-T+1}^{T-1} \hat{K}_X(h) \mathrm{e}^{-2\pi i \omega_j h} \end{aligned}$$

Hence, the periodogram can be written in terms of the Fourier transform of the estimated autocovariance as we might have expected from the previous discussion. The problem we face here is that that the estimator $\hat{K}_X(h)$ is very poor for large h as there are relatively few pairs of time points to consider. Hence, we often truncate this summation by only summing over $|h| \leq m$ for some $m \ll T$.

4.3.1 Spectral ANOVA

We can consider the spectral approach to time series as an ANOVA problem. That is, we can consider how much variation in the time series is due to a certain frequency much like the sum of squares decomposition from classic ANOVA.

For simplicity, let T be odd. We consider

$$X_t = \beta_0 + \sum_{j=1}^{(T-1)/2} \{\beta_{j1} \cos(2\pi t\, j/T) + \beta_{j2} \sin(2\pi t\, j/T)\}$$

where we found before that

$$\hat{\beta}_{j1} = \frac{2}{T} \sum_{t=1}^{T} X_t \cos(2\pi t j/T) = \frac{2}{\sqrt{T}} d_c(\omega_j)$$

$$\hat{\beta}_{j2} = \frac{2}{T} \sum_{t=1}^{T} X_t \sin(2\pi t j/T) = \frac{2}{\sqrt{T}} d_s(\omega_j)$$

and $\hat{\beta}_0 = \bar{X}$. Therefore, we have

$$X_t - \bar{X} = \frac{2}{\sqrt{T}} \sum_{j=1}^{(T-1)/2} \{d_c(\omega_j)\cos(2\pi t\, j/T) + d_s(\omega_j)\sin(2\pi t\, j/T)\}$$

and $\sum_{t=1}^{T}(X_t - \bar{X})^2 = 2\sum_{j=1}^{(T-1)/2}\left\{d_c(\omega_j)^2 + d_s(\omega_j)^2\right\} = 2\sum_{j=1}^{(T-1)/2} I(\omega_j)$. This is because $\sum_{t=1}^{T}\cos(2\pi tj/T)^2 = T/2$ and similarly for the sine series.

Thus, we have decomposed the total sum of squares over T data points into the sum of $(T-1)/2$ terms, $2I(\omega_j)$, each with 2 degrees of freedom. Thus, the periodogram $I(\omega_j)$ can directly be thought of as the variation due to frequency ω_j in the time series. Note that one would never want to use the `aov()` function in R to compute this. Instead, the `fft()` is much more efficient.

4.3.2 Large Sample Behaviour

In this section, we assume X_t is a stationary process with mean μ, absolutely summable autocovariance function $K_X(h)$ and spectral density $f_X(\omega)$. If we write the periodogram using the true mean μ—this makes the calculations easier than using $\bar{X}$—we find that

$$I(\omega_j) = \frac{1}{T}\sum_{h=-T+1}^{T-1} \mathrm{e}^{-2\pi i\omega_j h} \sum_{t=1}^{T-|h|} (X_{t+|h|} - \mu)(X_t - \mu)$$

$$\mathrm{E}[I(\omega_j)] = \sum_{h=-T+1}^{T-1} \left(\frac{T-|h|}{T}\right) K_X(h)\mathrm{e}^{-2\pi i\omega_j h}.$$

For taking the limit as $T \to \infty$, we have to consider a sequence of frequencies $\omega_j^{(T)}$ that tends towards some ω as the sample size grows. For example, if we want $\omega = 1/3$, we could consider

$$\omega_1^{(2)} = 1/2,\ \omega_1^{(4)} = 1/4,\ \omega_3^{(8)} = 3/8,\ \omega_5^{(16)} = 5/16,\ \omega_{11}^{(32)} = 11/32\ \ldots$$

In this case, if we have $\omega_j^{(T)} \to \omega$ as $T \to \infty$, then

$$\mathrm{E}[I(\omega_j^{(T)})] \to f_X(\omega) = \sum_{h=-\infty}^{\infty} K_X(h)\mathrm{e}^{-2\pi i\omega_j h}.$$

Going further, if we strengthen the absolute summability condition $\sum_{h=-\infty}^{\infty}|K_X(h)| < \infty$ to the condition

$$c = \sum_{h=-\infty}^{\infty} |h||K_X(h)| < \infty,$$

then we have that

$$\operatorname{cov}\left(d_c(\omega_j), d_c(\omega_k)\right) = \begin{cases} f_X(\omega_j)/2 + \varepsilon_T & \text{for } \omega_j = \omega_k \\ \varepsilon_T & \text{for } \omega_j \neq \omega_k \end{cases}$$

and similarly for d_s where ε_T is an error term bound by $|\varepsilon_T| \leq c/T$. Hence, the estimated covariance matrix should have a strong diagonal with smaller noisy off-diagonal entries.

We can use this to find via the central limit theorem that if our process X_t is just iid white noise with variance σ^2, then

$$d_c(\omega_j^{(T)}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2/2\right)$$
$$d_s(\omega_j^{(T)}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2/2\right)$$

Thus, recalling that $I(\omega_j) = d_c(\omega_j)^2 + d_s(\omega_j)^2$, we have that

$$2I(\omega_j^{(T)})/\sigma^2 \xrightarrow{\mathrm{d}} \chi^2(2)$$

and this $I(\omega_j^{(T)})$ will be asymptotically independent with some other $I(\omega_k^{(T)})$.

For the general linear process, we have

Theorem 4.3.1. *If $X_t = \sum_{j=\infty}^{\infty} \psi_j w_{t-j}$ with the ψ_j absolutely summable and with w_t being iid white noise with variance σ^2 and with*

$$\sum_{h=\infty}^{\infty} |h||K_X(h)| < \infty,$$

then for any collection of m frequencies $\omega_j^{(T)} \to \omega_j$, we have jointly that

$$2I(\omega_j^{(T)})/f(\omega_j) \xrightarrow{d} \chi^2(2)$$

given that $f(\omega_j) > 0$ for $j = 1, \ldots, m$.

Thus, we can use this result for many statistical applications like constructing a $1-\alpha$ confidence interval for the spectral density f_x at some frequency ω by

$$\frac{2I(\omega_j^{(T)})}{\chi^2_{2,1-\alpha/2}} \leq f_X(\omega) \leq \frac{2I(\omega_j^{(T)})}{\chi^2_{2,\alpha/2}}.$$

4.3.3 Banding, Tapering, Smoothing, and more

In the previous section, it is noted that the periodogram is a natural estimator for the spectral density. However, there are many aspects of such estimation to consider. In this section, we will consider methods of averaging, smoothing, banding, and tapering, which will, if used correctly, give us a better estimator for the spectral density. Ultimately, there is no one right way to estimate a given spectral density from some time series data. One must explore some of the following techniques as appropriate.

Bartlett's and Welch's Methods

For a time series $X_1, \ldots, X_T$, we are able to compute $I(\omega_j)$ for any frequency $\omega_j = j/T$. However, such fine granularity is often not necessary. Instead, computing the periodogram for fewer frequencies with better accuracy is often preferrable.

Bartlett's method take this into consideration by splitting the time series into K separate disjoint series of equal length $m = T/K$. That is,

$$\{X_1, \ldots, X_K\}, \{X_{K+1}, \ldots, X_{2K}\}, \ldots \{X_{mK-m+1}, \ldots, X_T\}.$$

Then, for each of these K time series pieces, we can compute periodograms $I^{(1)}(\omega_j), \ldots, I^{(K)}(\omega_j)$ and average them to get

$$I(\omega_j) = K^{-1} \sum_{i=1}^{K} I^{(i)}(\omega_j).$$

In this case, we only have periodogram values for $\omega_j = j/m$ for $j = 1, \ldots, m$. But the variance of the estimate decreases. It is also faster to compute as performing K DFTs of size m is faster than performing one DFT of size $mK = T$.

Welch's method is nearly identical to Bartlett's method. However, this new approach allows for the time series to be partitioned into overlaping pieces that overlap by a fixed number of data point.

Banding

Instead of partitioning in the time domain as Barlett's method does, we can instead partition the frequencies into *bands*. In this case, we can define a frequency band of $2m + 1$ frequencies to be

$$\mathcal{B} = \left\{ \omega \, : \, \omega_j - \frac{m}{T} \leq \omega \leq \omega_j + \frac{m}{T} \right\}.$$

Here, we say that $(2m + 1)/T$ is the bandwidth of $\mathcal{B}$. The idea is that if locally $f_X(\omega)$ is constant for all frequencies in the band $\mathcal{B}$, then the spectal density can be estimated to be

$$\bar{I}(\omega) = \frac{1}{2m + 1} \sum_{i=-m}^{m} I(\omega_j + i/T)$$

for any $\omega \in \mathcal{B}$. Considering the previous result that $2I(\omega_j^{(T)})/f(\omega_j) \xrightarrow{\text{d}} \chi^2(2)$, we have the extension that

$$\frac{2(2m + 1)\bar{I}(\omega_j^{(T)})}{f(\omega_j)} \xrightarrow{\text{d}} \chi^2(4m - 4)$$

as long as T is large and m is small. Note that there typically is no optimal bandwidth and many can be tried when analyzing a time series in the spectal domain.

The above notion of banding simply weights all frequencies in the band $\mathcal{B}$ equally, which is $1/(2m+1)$. Instead, we can use a weighted average of the frequencies of the form

$$\tilde{I}(\omega) = \sum_{i=-m}^{m} c_i I(\omega_j + i/T)$$

where the weights c_i sum to 1. The mathematically get convergence for this object to a chi-squared distribution as before, we require that as $T \to \infty$ and $m \to \infty$ such that $m/T \to 0$ that $\sum_{i=-m}^{m} c_i^2 \to 0$. Then, it can be shown that

$$\begin{aligned} \mathrm{E}[\tilde{I}(\omega)] &\to f_X(\omega) \\ \frac{\mathrm{cov}\left(\tilde{I}(\omega_1), \tilde{I}(\omega_2)\right)}{\sum_{i=-m}^{m} c_i^2} &\to \begin{cases} 0 & \omega_1 \neq \omega_2 \\ f_X(\omega)^2 & \omega_1 = \omega_2 \neq 0 \neq 1/2 \\ 2f_X(\omega)^2 & \omega_1 = \omega_2 = 0 \text{ or } = 1/2 \end{cases} \end{aligned}$$

In this case, $\tilde{I}$ is asymptotically a weighted sum of chi-squared random variables which is hard to work with directly. Instead, we can approximate the "length" of the band by

$$L = \left[\sum_{i=-m}^{m} c_i^2 \right]^{-1}$$

to get roughly that

$$\frac{2L\tilde{I}(\omega_j^{(T)})}{f(\omega_j)} \xrightarrow{\mathrm{d}} \chi^2(2L).$$

Note that this works perfectly if we replace c_i with $(2m+1)^{-1}$ recovering the equally weighted scenario from above.

One way to choose specific weights is via the *Daniell kernel*. In this case, we begin simply with $c_i = 1/3$ for $i = -1, 0, 1$. Applying this to a sequence u_t results in

$$u_t^{(1)} = \frac{u_{t-1}}{3} + \frac{u_t}{3} + \frac{u_{t+1}}{3}.$$

If we apply this kernel a second time, we get

$$\begin{aligned} u_t^{(2)} &= \frac{u_{t-1}^{(1)}}{3} + \frac{u_t^{(1)}}{3} + \frac{u_{t+1}^{(1)}}{3} \\ &= \frac{u_{t-2}}{9} + \frac{2u_{t-1}}{9} + \frac{3u_t}{9} + \frac{2u_{t+1}}{9} + \frac{u_{t+2}}{9}. \end{aligned}$$

Also, sometimes the *modified Daniell kernel* is applid, which is the same idea as above be starting with

$$u_t^{(1)} = \frac{u_{t-1}}{4} + \frac{u_t}{2} + \frac{u_{t+1}}{4}.$$

Tapering

Tapering a time series is another way to focus in on estimating the spectral density for a certain range of frequencies. To discuss this, we begin in the time domain. For a mean zero stationary process X_t with spectral density $f_X(\omega)$, we construct a tapered process with $Y_t = a_t X_t$ for some coefficients a_t. Thus, the DFT for Y_t gives us

$$d_Y(\omega_j) = \frac{1}{T^{1/2}} \sum_{i=1}^{T} a_t X_t \mathrm{e}^{-2\pi i \omega_j t}.$$

Then, the expected value of the periodogram is

$$\mathrm{E}\left[I_Y(\omega_j)\right] = \int_{-1/2}^{1/2} W_T(\omega_j - \omega) f_X(\omega) d\omega$$

with $W_T(\Omega) = |A_T(\omega)|^2$ with $A_T(\omega) = T^{-1/2} \sum_{i=1}^{T} a_t \mathrm{e}^{-2\pi i \omega t}$. Here, we refer to $W_T(\omega)$ as the spectral window.

Example 4.3.1. *The Fejér or modified Bartlett kernel is*

$$W_T(\omega) = \frac{\sin(n\pi\omega)^2}{n \sin(\pi\omega)^2}$$

with $W_T(0) = n$, which comes from $a_t = 1$ for all t. When averaging over a band $\mathcal{B}$ as above, the spectral window is similarly averaged. That is, for $\bar{I}(\omega) = \frac{1}{2m+1} \sum_{i=-m}^{m} I(\omega_j + i/T)$, we have

$$W_T(\omega) = \frac{1}{2m+1} \sum_{i=-m}^{m} \frac{\sin(n\pi(\omega + i/T))^2}{n \sin(\pi(\omega + i/T))^2}.$$

Example 4.3.2. *Other tapers that live up to the name "tapering" include the cosine taper, which sets the coefficients $a_t = [1 + \cos(2\pi(t - (T+1)/2)/T)]/2$.*

4.3.4 Parametric Estimation

4.4 Filtering