Time Series Analysis

Yong Yoon

November 8, 2022

1 Introduction

1.1 What is a Time Series?

Definition: A time series is a realization of a random variable indexed by time.

$$X_t = \{x_1, x_2, \dots, x_T\}$$

Our task is to analyze the above stochastic process.

- CDF is $P(X \leq x) = F_X(x)$
  - ◇ $F_x(x)$ gives us the full description of of random variable X.
- For two random variables, $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$
- What if we have many variables?
  - ◇ How do we estimate their joint distribution?
  - ◇ It would be virtually impossible, so we make some assumptions to make things manageable
- Time series can be viewed as a stochastic process of $X_1, X_2, \dots, X_t, \dots$

1.2 White Noise

[Example 1.8]

- A white noise process $w_0, w_1, \dots$ (sometimes denoted $a_0, a_1, \dots$)
  - ◇ $Cov(w_t, w_s) = 0$ for $t \neq s$
  - ◇ $E[w_t] = 0$
  - ◇ $Var[w_t] = \sigma_w^2$, which is a constant (w.r.t. t)
- Often denoted $w_t \sim wn(0, \sigma_w^2)$

- We often require the noise to be independent and identically (iid) random variables with mean 0 and variance σ_w^2
- A particular case is the Gaussian white noise:
  - ◇ $w_t \sim N(0, \sigma_w^2)$
  - ◇ This is a rather strict condition, but allows us to manage things statistically

1.3 More Statistics

- Suppose X and Y are random variables
- Expectation $\mu_X = E[X]$ and $\mu_Y = E[Y]$
- Variance $\sigma_X^2 = Var[X]$ and $\sigma_Y^2 = Var[Y]$
  Notice that $Var[X] = E[(X - \mu_X)^2] = E[X^2] - \mu_X^2$
- Covariance

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$$

- Correlation

$$\rho_{X,Y} = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

We have $-1 \leq \rho \leq 1$ (you should be able to proof this through the Cauchy-Swartz Inequality). Note also that uncorrelated does not necessarily mean independent. However, independence guarantees no correlation.

Example: For $Y = X^2$ and gaussian X (say, standard normally distributed).

$$Cov(X, Y) = E[X^3] - E[X]E[X^2] = 0$$

But note that X and Y are <u>not</u> independent.

1.4 Stationary Time Series

[Section 1.5]

- What does it mean that X_t is stationary? Stochastically, its joint distribution does not change
- For $x_1, \ldots, x_t, \ldots$ their joint distribution does not change in the following sense:

- **Strictly stationary** time series [Definition 1.6]
  For $F_{t_1,t_2,\ldots,t_n}(x_1,\ldots,x_n) = P(X_{t_1} \leq x_1, X_{t_2} \leq x_2, \ldots, X_{t_n} \leq x_n)$,
  $F_{t_1+h,t_2+h,\ldots,t_n+h}(x_1,\ldots,x_n) = F_{t_1,t_2,\ldots,t_n}(x_1,\ldots,x_n)$ for all $n = 1, 2, \ldots,$
  for all time points and all time shifts h.

In words, a strictly stationary time series is one for which the probabilistic behavior of every collection of values $\{x_{t_1}, x_{t_2}, \ldots, x_{t_n}\}$ is identical to that of the time shifted set $\{x_{t_{1+h}}, x_{t_{2+h}}, \ldots, x_{t_{n+h}}\}$.

Example: For $n = 1, F_t(x) = P(X_t \leq x) = F_{t+h}(x)$.

Example: For $n = 2$, the joint distribution of X_t and X_s is the same as that of X_{t+h} and X_{s+h}.
$\rightarrow Cov(X_t, X_s) = Cov(X_{t+h}, X_{s+h})$

- All said and done, however, for practical purposes it is impossible to estimate the joint distribution of more than 2 variables.

1.5 Weakly Stationary Time Series

[Definition 1.7]

- X_t is weakly stationary if
  (1) $E[X_t]$ is constant
  (2) $Cov(X_{t+h}, X_{s+h}) = Cov(X_t, X_s)$ for any t, s
  (3) If $t = s$ then we have $Var(X_{t+h}) = Var(X_t)$, which is some constant σ_X^2.

- Weakly stationary X_t has

  $$\gamma_X(h) = Cov(x_t, x_{t+h}) \text{ that is only dependent on } h \text{ (and not } t)$$

  and $\gamma_X(h)$ is known as the autocovariance function of X (where h is the lag).

- Is white noise w_t stationary? [Example 1.19]
  - ⋄ $E[w_t] = 0$
  - ⋄ $Var[w_t] = \sigma_w^2$
  - ⋄ $Cov(w_t, w_s) = 0, t \neq s$
  - ⋄ Overall we have

$$\gamma_w(h) = cov(w_{t+h}, w) = \begin{cases} \sigma_w^2 & h = 0, \\ 0 & h \neq 0. \end{cases}$$

1.6 Autocorrelation Function

[Definition 1.9]

- The autocorrelation function (ACF) or $\rho_X(h)$ of a stationary time series X_t is given by

$$\rho_X(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}}$$

- In particular, if X_t is (weakly) stationary,

$$\rho_X(h) = \frac{\gamma(h)}{\gamma(0)}$$

- By the C-Cauchy-Schwartz inequality, $-1 \leq \rho_X(h) \leq 1$.
- Similarly, the autocovariance function γ_X satisfies (see 1.25 in text)
  - ⋄ $|\gamma(h)| \leq \gamma(0)$
  - ⋄ $\gamma(h) = \gamma(-h)$

2 Linear Models

[Definition 1.12]

2.1 Moving Average

[Definition 3.3]

- Moving average of order (q), $MA(q)$, is defined as:

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q} + \mu$$

  the coefficents θ's are not random but constants that are unknown; so our job is to estimate them from the data.

- For $q = 1$ or $MA(1)$ we have

$$x_t = w_t + \theta_1 w_{t-1} + \mu$$

  - ⋄ Is the $MA(1)$ stationary?
    $E[X_t] = E[w_t] + \theta_1 E[w_{t-1}] + \mu = \mu$
  - ⋄ What is its autocovariance function γ_X?

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma_w^2 & h = 0, \\ \theta\sigma_w^2 & h = 1, \\ 0 & h > 1. \end{cases}$$

- Overall $\gamma(h)$ does not depend on t, and $E[X_t]$ is constant. Hence the process is stationary.

- What about the autocorrelation $\rho(h)$?

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} 1 & h = 0, \\ \frac{\theta}{1+\theta^2} & h = 1, \\ 0 & h > 1. \end{cases}$$

  Example 3.5: An $MA(1)$ process with $w_t \sim N(0,1)$ and $\theta = 5$ gives:

$$\gamma(h) = \begin{cases} 26 & h = 0, \\ 5 & h = 1, \\ 0 & h > 1. \end{cases}$$

- Interestingly an $MA(1)$ process with $w_t \sim N(0,1)$ and $\theta = 1/5$ gives the same results.
  We will give preference to the latter (as this is related to the idea of invertiblity).

2.2 Infinite Moving Average Process

- $X_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \mu$
- $\sum_{j=1}^{\infty} |\theta_j| < \infty$
  Note that the term μ can be removed by demeaning.

Then X_t is stationary.

- The autocovariance function is given by

$$\gamma(h) = \sigma_w^2 \sum_{j=0}^{\infty} \theta_j \theta_{j-h}.$$

  Note that θ_w^2 will have to be estimated form the data. Or for simplicity, we can assume it to be 1.

- We have so far looked at a specific model, $MA(q)$.
- The next question then is to see how, given a series, say $x_1, \ldots, x_{1000}$, we might be able to fit the series into a certain model after computing sample autoovariance/autocorrelation functions, etc.

2.3 General Approach to Modelling Time Series

- A first step usually is to plot the given series to examine its main features:
  - ◇ a trend
  - ◇ a seasonal component
  - ◇ any apparent sharp changes in behaviour
  - ◇ any outlier observations
- We then remove the trend and seasonal component(s), if any, to get stationary series or residuals.
  - ◇ If requires, a transformation is used, e.g. $\ln x_t$ can be used
  - ◇ Or the series may be differenced, e.g. $\Delta X_t = X_t - X_{t-1}$, etc.
  - ◇ Or even a combination, log-difference, may be employed, and so on.

- Next we choose a model to fit the series or residuals, making use of various statistics.
- We would then often forecast future observations or make predictions, etc.

In sum, we have the classical decomposition model:

$$x_t = m_t + s_t + y_t$$

where m_t is a trend component, s_t is a seasonal component, and y_t is a random noise component. Usually, y_t is modeled by a stationary model, e.g., $MA(q)$.

2.4 Trend Estimation

There are a number of statistical methods to estimate trends.

- Smoothing with a finite moving average filter [Example 2.10]
  Let's take a nonseasonal model as follows:

  $$x_t = m_t + y_t$$

  Then for a non-negative integer q, define

  $$v_t = \frac{1}{2q+1} \sum_{j=-q}^{q} x_{t-j}$$

  Then

  $$v_t = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t-j} + \frac{1}{2q+1} \sum_{j=-q}^{q} y_{t-j}$$

  of which the term will come close of some trend m_t if in a linear trend while the second term will vanish to zero (i.e. $\hat{m}_t = v_t$)

- Exponential smoothing

  $$\begin{cases} \hat{m}_t = \alpha x_t + (1-\alpha)\hat{m}_{t-1} & t = 2, \ldots \\ \hat{m}_1 = x_1 \end{cases}$$

- Smoothing splines [Example 2.14]
  Typically piecewise polynomials of order 3, which is called 'cubic splines'.

Piecewise polynomials f_t is given by minimizing the fit and degree of smoothness

$$\sum_{t=1}^{n}[x_t - f_t]^2 + \lambda \int (f_t'')^2 dt,$$

$\lambda > 0$ determines the degree of smoothness. But how do we determine this?

- Kernel Smoothing [Example 2.12]:
  In moving average, for example,

$$\hat{m}_t = \frac{1}{3}x_{t-2} + \frac{1}{3}x_{t-1} + \frac{1}{3}x_t$$

  Perhaps this is too simple. Why equal weights?

$$\hat{m}_t = \sum_{i=1}^{n} w_i(t)x_i,$$

  where the weight $w_i(t) = \dfrac{K(\frac{t-i}{b})}{\sum_j K(\frac{t-j}{b})}$.
  There are a number of kernel functions we could use, for example, the standard normal density function $K(z) = \dfrac{1}{\sqrt{2\pi}}\exp\left(-z^2/2\right)$.
  The parameter b is the bandwidth.
  The weight $w_i(t)$ assigns the i-th observation's contribution to x_t.

- Trend estimation by differencing [Example 2.4]
  For $x_t = m_t + y_t$, by ordinary least squares (OLS), say, we had $\hat{m}_t = -11.2 + 0.006t$.
  Then the detrended signal is given by $\hat{y}_t = x_t + 11.2 - 0.006t$.
  Compare the above difference with the difference $x_t - x_{t-1} = m_t - m_{t-1} + y_t - y_{t-1}$.
  If the trend is locally linear then $m_t - m_{t-1}$ is constant.
  Therefore in general let us consider $z_t = \Delta y_t = y_t - y_{t-1}$ for stationary y_t.
  The covariance function $\gamma_z(h)$ is then

$$\gamma_z(h) = cov(z_{t+h}, z_t) = 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1),$$

  which implies z is also stationary.
  More often than not, z_t allows easier modeling than the OLS approach (see figure 2.5 of the ACFs of the detrended and of the differenced series.)

2.5 Backshift or Lag Operator

[Definition 2.4]

- The backshift operator is

$$
\begin{aligned}
BX_t &= X_{t-1} \\
\Delta X_t &= (1-B)X_t
\end{aligned}
$$

  Then, $\Delta(\Delta X_t) = X_t - 2BX_t + B^2X_t = (1-B)^2X_t = \Delta^2 X_t$.
  That is the difference of order d is given by

$$\Delta^d = (1-B)^d.$$

2.6 Seasonality Estimation

- Can be modeled by regression analysis, S-ARIMA models, FFT analysis, etc.
- For the FFT (fast Fourier Transform) approach, we will look at the peridogram, which allows us to detect a certain dominant frequency [Example 2.9].
- We will cover ARIMA models first.
- For the sake of model building, a general model is

$$x_t = m_t + s_t + y_t$$

  where m_t is the trend, y_t is seasonality, and y_t is residuals. We will examine in depth the stationary model for y_t.

2.7 Autoregressive Model

[Definition 3.1]

- First, the AR(1) or autoregressive model of order 1:

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

  The constant term μ can be assumed to be zero by demeaning.

$$
\begin{aligned}
E[x_t] &= \phi E[x_{t-1}] + E[w_t] + \mu \\
Var[x_t] &= \phi^2 Var[x_{t-1}] + \sigma_w^2 + 2\phi\ Cov[x_{t-1}, w_t]
\end{aligned}
$$

The term $Cov[x_{t-1}, w_t]$ is zero.
By recursively applying the model, we have

$$\begin{aligned} x_t &= w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \dots \\ &\quad + \mu + \phi\mu + \phi^2\mu + \dots + \phi^m X_{t-m} \end{aligned}$$

and the last term goes to zero assuming that $\phi \leq 1$.

- It follows that $E[x_t] = \dfrac{\mu}{1-\phi}$ and $Var[x_t] = \dfrac{\sigma_w^2}{1-\phi^2}$.
  We observe that $Cov[x_t, w_{t+1}] = 0$, which means that x_t and w_{t+1} are not correlated and we have a causal system.

Definition 3.7 Causal ARMA model:

$$\begin{aligned} x_t = \Theta_0 w_t + \Theta_1 w_{t-1} + \Theta_2 w_{t-2} + \dots \\ \sum_{j=0}^{\infty} |\Theta_j| < \infty \end{aligned}$$

Example: $x_t = 2x_{t-1} - w_t$ is not causal.

- Again, we assume in causal models that x_t and w_{t+1} are not correlated.
  What about $AR(2)$?

- Recall the backshift operator:

$$\begin{aligned} BX_t &= X_{t-1} \\ B^2 X_t &= X_{t-2} \\ B^m X_t &= X_{t-m} \\ x_t &= \phi x_{t-1} + w_t \\ x_t &= \phi B x_t + w_t \\ (1-\phi B) x_t &= w_t \end{aligned}$$

  Then $x_t = (1-\phi B)^{-1} w_t$ provided $|\phi| < 1$.
  Then

$$\begin{aligned} x_t &= (1-\phi B)^{-1} w_t \\ &= (1 + \phi B + \phi^2 B^2 + \dots) w_t \\ &= w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \dots \end{aligned}$$

- AR(2) is

$$\begin{aligned} x_t &= \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t + \mu \\ x_t &= (1 - \phi_1 B - \phi_2 B^2)^{-1} (w_t + \mu), \end{aligned}$$

but under what conditions is this invertable? We need to factorize the polynomial as

$$(1-\phi_1 B-\phi_2 B^2)=(1-\alpha_1 B)(1-\alpha_2 B)$$

which is called the charateristic equation.
And the condition for invertibility is

$$|\alpha_1|<1, |\alpha_2|<1.$$

We can show further that the above conditions are equivalent to:

$$\phi_1+\phi_2<1, \phi_2-\phi_1<1, |\phi_2|<1.$$

- The same steps are applicable to AR(3)

$$\begin{aligned} x_t &= \phi_1 x_{t-1}+\phi_2 x_{t-2}+\phi_3 x_{t-3}+w_t+\mu \\ x_t &= (1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)^{-1}(w_t+\mu), \end{aligned}$$

  We need to factorise the polynomial:

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)=(1-\alpha_1 B)(1-\alpha_2 B)(1-\alpha_3 B)$$

  The condition for stationarity or invertibility is

$$|\alpha_1|<1, |\alpha_2|<1, |\alpha_3|<1.$$

  The closed form in terms of ϕ_i are unavailable and hence numerical methods are employed.

- Let's take an example of a AR(2) model:

$$\begin{aligned} x_t &= x_{t-1}-0.89x_{t-2}+w_t \\ y^2-y-0.89=0 \leftrightarrow y=0.5\pm 0.8i \quad &\text{for } |y_1|<1, |y_2|<1 \end{aligned}$$

  i.e. causal and stationary!

$$\begin{aligned} (1-B+0.89B^2)x_t &= w_t \\ x_t &= (1-B+0.89B^2)^{-1}w_t \\ x_t &= w_t+\psi_1 w_{t-1}+\psi_2 w_{t-2}+\dots \\ &= w_t+(\phi_1-1)w_{t-1}+(0.89-\phi_1+\phi_2)w_{t-1}+\dots \end{aligned}$$

  which leads to $\psi_1=1, \psi_2=0.11$, and so on.
  From the above example, $E[X_t]=\dfrac{1}{1-1+0.89}$

- To identify an appropriate model from a given time series, we need to know the autocovariance $\gamma_X(h)$ and autocorrelation $\rho_X(h)$.

- Let's look for the autocovariance $\gamma_X(h)$ and autocorrelation $\rho_X(h)$ of AR models.

  AR(2): $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$
  Multiply throughout by x_{t-h} to get

$$x_t x_{t-h} - \phi_1 x_{t-1} x_{t-h} - \phi_2 x_{t-2} x_{t-h} = w_t x_{t-h}$$

  Then take expectations.
  For $h = 0$, we have

$$\gamma_X(0) - \phi_1 \gamma_X(1) - \phi_2 \gamma_X(2) = \sigma_w^2$$

  Similarly, for $h = 1$, we have

$$\gamma_X(1) - \phi_1 \gamma_X(0) - \phi_2 \gamma_X(1) = 0$$

- Hence, we have the so-called Yule-Walker equations [Definition 3.10]

$$\gamma_X(h) - \phi_1 \gamma_X(h-1) - \phi_2 \gamma_X(h-2) = \begin{cases} 0 & h > 0, \\ \sigma_w^2 & h = 0 \end{cases}$$

  Dividing by $\gamma_X(0)$ gives the expression in terms of $\rho_X(h)$, the autocorrelation functions

$$\rho_X(h) - \phi_1 \rho(h-1) - \phi_2 \rho(h-2) = 0, \quad \text{for } h = 0, 1, 2, \ldots$$

  Notice that $\rho_X(0) = 1$ always.
  We can apply the above to AR(p) model.

- For AR(1), $x_t = \phi_1 x_{t-1} + w_t$, and $|\phi_1| < 1$.

$$\rho_X(h) = \phi \rho(h-1), \quad \text{for } h = 0, 1, 2, \ldots$$
$$\rho_X(h) = \phi^h \to \text{exponential decay}$$

- Let's investigate the behavior of $\rho(h)$ for $AR(p)$. Consider the AR(2), e.g.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

  Y-K: $\rho(h) = \phi_1 \rho(h-1) + \phi_2 \rho(h-2)$.
  Then $\rho(h)$ can be expressed as

$$\rho(h) = c_1 y_1^h + c_2 y_2^h \quad \text{for some } \phi_1, \phi_2$$

  where y_1 and y_2 are the roots of $y^2 - \phi_1 y - \phi_2 = 0$.
  How do we compute c_1 and c_2?
  Use $\rho(0) = 1$ which leads to $c_1 + c_2 = 1$, and $\rho(1) = \dfrac{\phi_1}{1 - \phi_2}$ which leads to $c_1 y_1 + c_2 y_2$.

Example 3.12 If y_1 and y_2 are real numbers, under stationary conditions, they are linear combinations of two exponentially decaying series.
- If they are complex numbers, the series behaves like sinuous decay.
- If $y_1 = y_2$,

$$\rho(h) = \left(1 + \frac{1+\phi_2}{1-\phi_2}h\right)(\phi_1/2)^h$$

Example $x_t = x_{t-1} - 0.89x_{t-2} + w_t$,
Y-K: $\rho(h) = \rho(h-1) - 0.89\rho(h-2)$ and

$\rho(1) = 1/(1+0.89), \rho(2) = \rho(1) - 0.89$.
How do we get $\gamma_X(0)$?

$$\begin{aligned}Var(x_t) &= Var(x_{t-1} - 0.89x_{t-2} + w_t) \\ &= Var(x_{t-1}) + 0.89^2 Var(x_{t-2}) + \sigma_w^2 - 2 \times 0.89\gamma(0)\rho(1).\end{aligned}$$

Example 3.26 Given $x_1, x_2, \ldots, x_{144}$ observations, we would like to fit an AR(2) model to the data, i.e. $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$. How can we estimate ϕ_1 and ϕ_2?

(1) Compute $\hat{\gamma}(0) = 8.903$, and then $\hat{\rho}(1) = 0.849, \hat{\rho}(2) = 0.519$.
Recall $\rho(h) = Cov(x_t, x_{t+h}) = E[(x_t - m)(x_{t+h} - m)]$,

$$\hat{\gamma}(h) = \frac{1}{N-h}\sum_{t=1}^{N-h}(x_t - \bar{x})(x_{t+h} - \bar{x}),$$

$$\hat{\rho} = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

(2) Use $\rho(h) = \phi_1\rho(h-1) + \phi_2\rho(h-2)$ for $h = 1, 2, 0$ with $\hat{\rho}(h)$ in place of $\rho(h)$.
Hence we have three unknowns, i.e. ϕ_1, ϕ_2 and σ_w^2 and three equations.

- To sum up, AR(p) models are represented by

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + w_t + \mu$$

The Yule-Walker equations are

$$\gamma_X(h) - \phi_1\gamma_X(h-1) - \ldots - \phi_p\gamma_X(h-p) = \begin{cases} 0 & h > 0, \\ \sigma_w^2 & h = 0 \end{cases}$$

Dividing by $\gamma_X(0)$, the Y-K equations with the autocorrelation $\rho_X(h)$.

2.8 ARMA(p, q)

Let us look at ARMA(1,1) as an example:

- $x_t = \phi x_{t-1} + w_t + \theta w_{t-1}$
  Then $(1-\theta B)x_t = (1-\theta B)w_t$ and $(1-\theta B)^{-1}(1-\theta B)w_t$, if $|\phi| < 1$.
  From the above, we can express x_t as $c_t = w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \cdots$.
  Then we have $(1-\theta B)(1-\psi_1 B + \psi_2 B^2 + \cdots)w_t = (1-\theta B)w_t$.
  We can compute ψ_1 and ψ_2 in terms of ϕ and θ, that is:

  $$E[x_t] = \frac{\mu}{1-\phi}$$

  and

  $$Var[x_t] = \phi^2 Var[x_{t-1}] + \sigma_w^2 + \theta^2\sigma_w^2 + 2\phi\theta\sigma_w^2$$

  .

- We can generalize to ARMA(p, q) as

  $$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q} + \mu$$

  Similarly as in AR(p), the ARMA(1,1) we multiple $x_t = \phi_1 x_{t-1} + w_t + \theta w_{t-1}$ by x_{t-h} to get $\gamma(h) = \phi_1\gamma(h-1) + w_t x_{t-h} - \theta w_{t-1}x_{t-h}$. Note that

  $$w_t x_{t-h} = \begin{cases} \sigma_w^2, & h = 0, \\ 0, & h = 1, 2, \cdots \end{cases}$$

- Notice also that

  $$\theta w_{t-1} x_{t-h} = \begin{cases} \psi_1\sigma_w^2, & h = 0, \\ \sigma_w^2, & h = 1, \\ 0, & h = 2, 3, \cdots \end{cases}$$

- Dividing the above by $\rho(0)$ gives $\rho(h) = \phi_1\rho(h-1)$ for $h = 2, \cdots$, which is exponential decaying.

- We can do the same for ARMA(2,1), i.e. multiply $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t + \theta w_{t-1}$ by x_{t-h}, then take expectations and evaluate for $h = 1, 2, 3, \cdots$.

- Autocorrelation functions alone however do not discriminate AR(p) with ARMA(p, q).

2.9 Partial Correlation

- Why is the number of churches highly correlated with the number of crimes? Answer: Population
- For random variables X, Y the correlation coefficient $\rho_{X,Y} = Corr(X, Y)$ captures the degree of linear dependence between the two variables
- Partial correlation excluding X, Y excluding Z is

$$\rho_{XY.Z} = Corr(X, Y|Z)$$

and can be computed by regressing X on Z and Y on Z to remove the influence of Z on X and Y respectively. That is, we have

$$\begin{aligned} X &= \alpha Z + \text{error}_X \\ Y &= \beta Z + \text{error}_Y \end{aligned}$$

Then the partial correlation is the correlation between the errors.

- We can show that this is

$$\rho_{XY.Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}$$

- Let's think a little about regresisons: Regress Y on X aftering centering [Example 3.14]

$$\begin{aligned} &\min_\alpha E[(Y = \alpha X)^2] \\ &= \min_\alpha Var(Y) - 2\alpha Cov(X, Y) + \alpha^2 Var(x). \end{aligned}$$

Taking the F.O.C gives

$$\hat{\alpha} = \frac{Cov(X, Y)}{Var(X)} \simeq corr(X, Y)$$

Note that the approximation was made under $Var(X) = Var(Y)$.

- Let's apply this in this time series context. Say for x_h for $h = 1, 2, \ldots$, the partial correlation can be thought of as

$$corr(x_t, x_{t-h}|x_{t-h-1}, x_{t-h+2}, \ldots, x_{t-1}) := \phi_{hh}$$

That is

$$\begin{aligned} \phi_{hh} = corr(&x_t - \alpha_1 x_{t-1} - \alpha_2 x_{t-2} - \ldots - \alpha_{h-1}x_{t-h+1}, \\ &x_{t-h} - \beta_1 x_{t-1} - \beta_2 x_{t-2} - \ldots - \beta_{h-1}x_{t-h+1}) \end{aligned}$$

- Let's look at the case for $AR(1)$. What is its partial correlation?

$$\phi_{11} = corr(x_t, x_{t-1}) = \rho(1) = \phi$$

$$\begin{aligned}\phi_{22} &= corr(x_t, x_{t-2}|x_{t-1}) \\ &= corr(x_t - \alpha x_{t-1}, x_{t-2} - \beta x_{t-1}) \\ &= corr(w_t, w_{t-1}) = 0.\end{aligned}$$

$$\begin{aligned}\phi_{33} &= corr(x_t, x_{t-3}|x_{t-1}, x_{t-2}) \\ &= corr(x_t - \alpha_1 x_{t-1} - \alpha_2 x_{t-2}, x_{t-2} - \beta_1 x_{t-1} - \beta_2 x_{t-2}) \\ &= corr(w_t, w_{t-2}) = 0.\end{aligned}$$

Notice that for $AR(1)$ we only get $\phi_{11} = \phi$ of $AR(1)$.

- What about for $AR(2)$? The partial correlations can be found the in the same manner.

$$\phi_{11} = \frac{\phi_1}{1 - \phi_{22}},\ \phi_{22} = \ldots$$

$$\begin{aligned}\phi_{33} = corr(x_t, x_{t-3}|x_{t-1}, x_{t-2}) \\ = corr(x_t - \alpha_1 x_{t-1} - \alpha_2 x_{t-2}, x_{t-2} - \beta_1 x_{t-1} - \beta_2 x_{t-2}) = 0, \\ \phi_{44} = 0, \ldots\end{aligned}$$

- In general, the PACF (Partial Autocorrelation Function) is computed via

$$\begin{aligned}\rho_1 &= \phi_{h1} + \phi_{h2}\rho_1 + \ldots + \phi_{hh}\rho_{h-1} \\ \rho_2 &= \phi_{h1}\rho_1 + \phi_{h2} + \ldots + \phi_{hh}\rho_{h-2} \\ &\cdots \\ \rho_h &= \phi_{h1}\rho_{h-1} + \phi_{h2}\rho_{h-2} + \ldots + \phi_{hh}\end{aligned}$$

And ϕ_{hh} gives us the PACF.

- Let's take an exmple of $AR(2)$

$$\begin{bmatrix}\rho(1) \\ \rho(2)\end{bmatrix} = \begin{bmatrix}\rho(0) & \rho(1) \\ \rho(1) & \rho(0)\end{bmatrix}\begin{bmatrix}\phi_{21} \\ \phi_{22}\end{bmatrix}$$

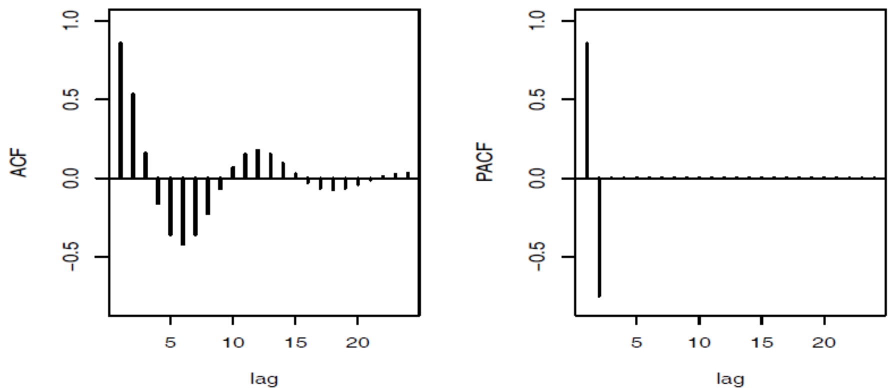
Note $\rho(0) = 1$. We can use the Y-W equations to find $\rho(1)$ and $\rho(2)$ then solve the linear system for ϕ_{21} and ϕ_{22} (the last term being the PACF).

2.10 ACF and PACF

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

We can determine the appropriate ARMA model after observing the ACF and PACF [See example 3.17].

Figure 1: Behaviour of ACF and PACF in Example 3.17



- Looks like AR(2) may be a good model.

2.11 Model Building

- Given $x_1, x_2, \dots, x_n$ we will try fitting an ARMA(p, q) model.
  1. Model identification (decide on values for p and q)
  2. Model estimation (estimate unknown parameters)
  3. Diagnostic checking (verify that we have a reasonable model)
  4. Prediction/Forecast
- Often the model identification and diagnostics are inseparable (they are considered together)
- (1) - (3) involve
  1. ACF and PACF
  2. Asymptotic (large-n) tests
     - Box-Ljung test, Sign test, Rank test, Q-Q plot, etc.
  3. AIC, BIC, FPE, ...

2.12 Testing whether ACF follows a WN

- We often test whether $\hat{\rho}(h)$:

$$\begin{cases} H_0: & \hat{\rho}(h) \text{ is same as WN,} \\ H_a: & \text{not } H_0 \end{cases}$$

We state Property 1.1

$$\hat{\rho}(h) \sim N(0, 1/\sqrt{n}), \quad \text{for large } n$$

To see why, use $\hat{\rho}(h) = \dfrac{1}{n-h} \displaystyle\sum_{t=h+1}^{n} w_t w_{t-h}$
And $E[\hat{\rho}] = 0, Var[\hat{\rho}] = \dfrac{1}{n-h} \simeq \dfrac{1}{n}$ when $n \gg h$

- Usually confidence interval of $\hat{\rho}$ is then $2/\sqrt{n}$.

2.13 Checking Residuals

- After a model is fit, the residual $\hat{w}_t$ should behave like a white noise
- One way to test whether $\hat{w}_t$ behaves like a white noise is with the Ljung-Box-Pierce Q-statistic, which is a χ^2-statistic
  - ◇ For autocorrelation functions $\hat{\rho}_h$ of residuals after fitting and ARMA(p,q) model, we have $\hat{\rho}_h \sim N(0, 1/n)$. Therefore, by definition, $\sqrt{n}\hat{\rho}_h^2 \sim \chi_1^2$
  - ◇ Thus the Box-Ljung statistic for ARMA(p,q) is

$$Q = n \sum_{h=1}^{k} \hat{\rho}_h^2 \sim \chi^2_{k-p-q}$$

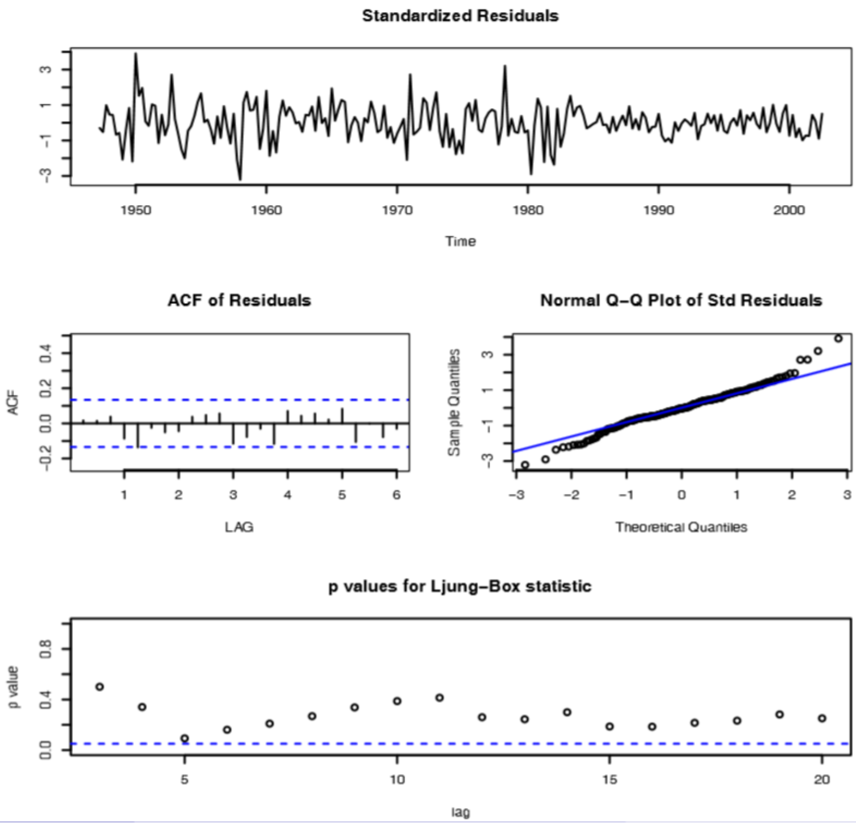
  - ◇ A large Q (or small p-value) suggest that the property of WN is not satisfied (and a better model is adviced)
  - ◇ Sometimes a modified version is used

$$Q = n(n+2) \sum_{h=1}^{k} \hat{\rho}_h^2 (n-k) \sim \chi^2_{k-p-q}$$

  - ◇ In practice, k is chosen to be around 20
- The (SSR) sum of squared residuals informs us how well the model fits

- The Box-Ljung Q statistic tells us how residuals as a group and their interrelations behave like white noise
- It is also instructive to draw a normal-probability plot (or q-q plot) to check whether residuals follow a normal distribution

Figure 2: Using sarima in R: Example 3.37



2.14 Model Selection by AIC and BIC

An important philosophy of time series analysis is parsimony. AIC and/or BIC help in model selection.

- While adding more parameters reduces the residuals, it worsens predictive power
- The AIC (Akaike Informtion Criteia) can be used as an indicator of theoretical prediciton performance

$$AIC = -2\log(\hat{L}) + \frac{2(p+q+1)n}{n-p-q-2}$$

where $\hat{L}$ is the likelihood value after fitting some appropriate ARMA(p, q). The second term is come penalty factor added for large p and/or q. The idea is we want a measure that compromises between model fitting and the number of parameters

- The idea is to find a model with the smallest AIC
- Another often used criterion is the BIC or Bayesian Information Criteria
$$BIC = -2\log(\hat{L}) + 2(p + q + 1)\log\, n$$
- See example 2.2

2.15 Model Estimation