STAT 443: Forecasting

Johnew Zhang

April 19, 2014

Contents

1 Introduction

Let T be an index set. The sequence of random variables $\{X_t, t \in T\}$ is a stochastic process if X_t is a random variable for all $t \in T$. If T is a set of time points, then $\{X_t\}$ is a time series. In this courses we will assume that T shows time points. If T is a discrete (continuous) set. then the time series $\{X_t\}$ is said to be discrete time (continuous time). The main focus of this course is to develop models for discrete time time series. For example, $x_5 = 10$ implies the value of X at time 5 is equal to 10.

There are couple of time series models: zero-mean models (i.i.d. noise, random walk and white noise) and models with trend & seasonality.

1.1 Time Series Models

Our interest lies in modeling and the analysis of data collected over time (time series). Ideally, given the random variables $X_1, X_2, X_3, \cdots$ one would like to specify all of joint distributions of the vectors $(X_1, X_2, \cdots, X_n)$ for all n, i.e.,

$$P(X_1 \leq x_1, \cdots, X_n \leq x_n), -\infty < x_1, \cdots, x_n < \infty, \forall n = 1, 2, 3, \cdots$$

In real world applications, this is usually not possible because enough information to fully specify the joint distributions is not available. The good news is that in most applications, most of the information about the joint distributions are provided in the first two moments and the covariances between pairs of random variables. In other words, $E(X_t), E(X_t^2)$ and $E(X_t X_{t^*}), \forall t, t^*$ summarizes most of the information content about the process.

If the joint distribution is multi-variate normal, then the three expectation $E(X_t), E(X_t^2)$, $E(X_t, X_{t^*}), \forall t, \forall t^*$ fully specify the joint distribution. Recall that the multivariate normal distribution is written as $N_p(\tilde{\mu}, \Sigma)$ where p is the dimension (number of X_i's) and μ is the mean vector $(p \times 1)$ and Σ is the variance-covariance matrix $(p \times p)$

$$\tilde{\mu} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{nn} & \cdots & \cdots & \sigma_n^2 \end{pmatrix}$$

where $\sigma_i^2 = Var(X_i), \sigma_{ij} = Cov(X_i, X_j)$. Therefore, given $E(X_t), E(X_t^2), E(X_t, X_{t^*})$, the joint multivariate normal distribution is fully specified. We can see that $E(X_t), E(X_t^2)$, $E(X_t, X_{t^*})$ contains a fair amount of information. Therefore, rather than working with joint distributions, we will work with time series models which mostly employ these 3 quantities.

Definition. *A time series model for observe data $\{x_t\}$ is a specification of the joint distributions (or possibly only the means, variances and covariances) of a sequence of random variables $\{X_t\}$ of which $\{x_t\}$ is postulated to be a realization. $X \rightarrow$ random variate and $x \rightarrow$ realization.*

For example, $y = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon$ and $E(Y|t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$. Hence $X = (t, t^2, t^3)$.

1.2 Some zero-mean models

1. i.i.d noise: If $X_1, X_2, X_3, \cdots, X_n$ are i.i.d. (independent identically distributed) random variables, then $Pr(X_1 \leq x_1, \cdots, X_n \leq x_n) =_{indep.} Pr(X_1 \leq x_1)Pr(X_2 \leq x_2) \cdots Pr(X_n \leq x_n) =_{identical} Pr(X_1 \leq x_1) \cdots Pr(X_1 \leq x_n) = \prod_{i=1}^{n} Pr(X_1 \leq x_i)$ so the joint distribution is defined by one marginal distribution.

   Observation: using independence assumption, we see that

   $$Pr(X_{n+h} \leq x | X_1 \leq x_1, \cdots, X_n \leq x_n) = \frac{Pr(X_{n+h} \leq x, X_r \leq x_1, \cdots, X_n \leq x_n)}{Pr(X_1 \leq x_1, \cdots, X_n \leq x_n)} = Pr(X_{n+h} \leq x)$$

   Therefore, independence implies that the history $(X_1, \cdots, X_n)$ has no value of predicting the future (X_{n+h}). If the sequence $X_1, \cdots, X_n$ above has the property of $E(X_i) = 0, \forall i$ then $X_1, \cdots, X_n$ are called i.i.d. noise.

2. Random walk: let $S_t = X_1 + X_2 + X_3 + \cdots + X_t$ where $X_t(t = 1, 2, \cdots)$ is i.i.d. noise. Then $\{S_t, t = 0, 1, 2, \cdots\}$ starting at $S_0 = 0$ is called a random walk. Notice that $E[S_t] = E[\sum_{i=1}^{t} X_i] = 0$.

3. White noise (a.k.a. zero-mean white noise): Another class of zero-mean time series models is which is a sequence of uncorrelated random variables (not necessarily independent), each with mean 0 and variance σ^2, we show such sequence as $\{X_t\} \sim WN(0, \sigma^2)$. Independence: $Pr(X_1 \leq x_1, \cdots, X_n \leq x_n) = Pr(X_1 \leq x_1) \cdots Pr(X_n \leq x_n)$. Uncorrelated: $E(X_i X_j) = E(X_i)E(X_j)$ and $Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) = 0$.

1.3 Models with Trend

Consider the models $X_t = m_t + Y_t$ where m_t is a slowly changing function, called the trend, and Y_t has zero mean i.e $E(Y_t) = 0, \forall t$. We have

$$E(X_t) = E(m_t) + E(Y_t) = m_t \forall t$$

Notice that m_t is a non-random function of time (t).

Examples: $m(t) = \alpha_0 + \alpha_1 t \implies$ linear trend, $m(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 \implies$ quadratic trend.

Example 1: Consider $X_t = m_t + Y_t$ where $m_t = 2 + t$ and $Y_t \sim N(0, 1)$. Then the time series will be exactly a graph of $y = m_t + Y_t$ and then we can compare this with linear regression. Use this example to motivate the use of regression in time series (trend estimation).

1.4 Models with Seasonal Component

In a similar setup to the previous case (models with trend) we can write $X_t = S_t + Y_t$ where $E(Y_t) = 0, \forall t$, and S_t is a periodic function with period d i.e., $S_t = S_{t+d}, \forall t$. In a sense, S_t is a particular type of trend. Same examples: $S_t = \alpha_0 + \alpha_1 \cos(\alpha_2 t) \implies$ used in signal processing.

$$S_t = \begin{cases} 1 & \text{if month is January} \\ 0 & \text{o.w.} \end{cases}$$

Used for monthly collected data (e.g. Danish Birth Data).

In both models $X_t = m_t + Y_t$ and $X_t = S_t + Y_t$, the parameters $\alpha_0, \alpha_1, \alpha_2, \cdots$ are usually estimated by maximum likelihood or least squares methods.

Note that in a general case, any may look at the model $X_t = m_t + S_t + Y_t$ (model with both trend and seasonal component). This is called classical decomposition (trend, seasonality, noise) and will be frequently referred to in this course, we can use regression models to estimate m_t and S_t. (talk about binary variates (indicator a seasonality modeling in regression).

1.5 Indicator Variables and Modeling Seasonal Behavior

Example: average seasonal temperature over many years

$$Z_1, Z_2, Z_3, \cdots \implies \{X_t : t = 1, 2, \cdots, 20\}$$

Suppose we want to fit a model of the form:

$$X_t = m_t + S_t + Y_t$$

where $m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p$

	X_1	X_2	X_3
spring	1	0	0
summer	0	1	0
fall	0	0	1
winter	0	0	0

$$X_1 = \begin{cases} 1 & \text{if season is spring} \\ 0 & \text{O.W.} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if season is summer} \\ 0 & \text{O.W.} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if season is fall} \\ 0 & \text{O.W.} \end{cases}$$

$$Z_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X - 3 + Y_t$$

where Y_t is just the random component and $S_t = \begin{cases} \alpha_1 & \text{spring} \\ \alpha_2 & \text{summer} \\ \alpha_3 & \text{fall} \\ 0 & \text{winter} \end{cases}$

Rule:

If a periodic trend with period d is being modeled through regression analysis, $d-1$ binary variates (indicator variables) should be introduced to the model. Suppose $p = 1 \implies z_t = \beta_0 + \beta_1 t + S_t + Y_t$.

1.5.1 Air Passengers Data

Let Y_t shows the total number of international passengers at time t.

Then $\log|X_t| = m_t + S_t + Y_t \implies {}_{iid}N(0, \sigma^2)$.

1. Left: $\log(Y_t) = \beta_0 + \beta_1 t + R_t$
2. Right: $\log(Y_t) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{11} X_{11} + R_t$

1.5.2 Model Checking

Residuals have no trend and points randomly scattered about 0.

Example 2: sees slides and R code online. In this example, we fitted the following model

$$\log(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 x^1 + \beta_3 x_2 + \cdots + \beta_{13} x_{11} + R_t$$

where the first three terms are trend m_t and following terms, seasonal component, is S_t and R_t is random component. Therefore, if interest lies in forecasting, this model fails. To be able to check the independence of residuals, as well as to introduce a new class of time series models, the concept of stationarity should be introduced.

1.6 Stationary Models and the Autocorrelation Function

Definition. *The time series $\{X_t : t \in T\}$ is called strictly (strongly) stationary if the joint distribution of $X_{t_1}, X_{t_2}, \cdots, X_{t_n}$ is the same as that of $X_{t_1-k}, X_{t_2-k}, \cdots, X_{t_n-k}$ for all $n, t_1, \cdots, t_n, k$. In other words, $\{X_t\}$ is strictly stationary of all its statistical properties remain the same under the shift.*

In practice, strict stationarity is too limiting of an assumption and rarely holds true. We mentioned earlier that a lot of the information about the joint distributions are provided on the moment $E[X_t], E[X_t^2]$ and $E[X_t X_{t^*}], t_1, t^*$.

This motivates introducing a type of stationarity based on these lower order moments (will be called weak stationarity).

To introduce weak stationarity we need some definitions first.

Definition. *Let $\{X_t\}$ be a time series with $E[X_t^2] < \infty$. This means function of $\{X_t\}$ is $\mu_x(t) = \mu_t = E[X_t]$ as a function of t.*

The covariance function $\{X_t\}$ is $\gamma_X(r,s), = Cov(X_r, X_s)$ and $E[(X_r - \mu_X(r))(X_S - \mu_X(s))], \forall r, s \implies$ function of r and s (function of time)

Definition. *The time series $\{X - t\}$ with $E[X_t^2] < \infty$ is said to be weakly stationary if*

1. *$\mu_X(t) = E[X_t]$ is indepedent of t*
2. *$\gamma_X(t, t+h) = Cov(X_t, X_{t+h})$ is independent of t for all $h \implies$ covariance only deadens on the distance between X_t and X_{t+h}.*

Notice that $E[X_t^2] < \infty$ is one of the conditions for weak stationarity (A total of three conditions should hold true)

Exercise: If $E[X_t^2] < \infty$, show that strict stationarity implies weak stationarity $E[X_t^2] < \infty$ strictly stationary.

Convention: whenever we refer to a stationary time series, we mean weakly stationary unless otherwise.

In view of condition ii of the weak stationarity definition, whenever we use the term "covariance function" with reference to a stationary time series $\{X_t\}$ we shall mean the function γ_X of one variable, defined by

$$\gamma_X(h) := \gamma_X(h, 0) = \gamma_X(t, t+h) = \gamma_X(t+h, t)$$

Last equality is because covariance is symmetric.

The function $\gamma_X(\cdot)$ will be referred to as the auto covariance function and $\gamma_X(h)$ as its value at $lagh$.

Definition. *Let $\{X_t\}$ be a stationary time series. The auto covariance function ($ACVf$) of $\{X_t\}$ at lagh is $\gamma_X(h) = Cov(X_{t+h}, X_t)$ The autocorrelation function (ACF) of $\{X_t\}$ at lagh is*

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = Corr(X_{t+h}, X_t)$$

where

$$Corr(X_t, X_{t+h}) = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t)Var(X_{t+h})}} = \frac{\gamma_X(h)}{\sqrt{Cov(X_t, X_t)Cov(X_{t+h}, X_{t+h})}} = \frac{\gamma_X(h)}{\gamma_X(0)}$$

Example 3

inverstigae the stationarity of white noise. Let $\{X_t\}$ be white noise $\{X_t\} \sim WN(0, \sigma^2)$. $\sigma^2 < \infty \implies Var(X) < \infty$ (this is equivalent to $E[X_t^2] < \infty$)

$E[X_t] = 0$ does not depend on t.

$$Cov(X_t, X_{t+h}) = \begin{cases} \text{Variance of } X_t & \text{if } h = 0 \text{ independent of t for all h} \\ 0 & \text{O. W.} \end{cases}$$

White noise is stationary.

Example 4

Investigate the stationarity of random walk. Let $\{X_t\}$ be a sequence of i.i.d. noise.

Define $S_t = X_1 + \cdots + X_t$, Then $\{S_t, t \geq 0\}$ is random walk in which $S_0 = 0$

$$Cov(S_t, S_t)\gamma_S(t, t_{t_0}) = Var(S_t) = Var(\sum_{i=1}^{t} X_i) = \sum_{i=1}^{t} Var(X_i) + 2\sum_{i=1}^{t} \sum_{j=1, j\neq i}^{t} Cov(X_i, X_j)$$

Therefore, $Var(S_t) = \sum_{i=1}^{t} \sigma^2 = t\sigma^2$. This depends on t, therefore random walk is not stationary.

Example 5

Consider the process $X_t = Z_t + \theta Z_{t-1}$ where $t = 0, \pm 1, \pm 2, \cdots$ where $\{z_t\} \sim WN(0, \sigma^2)$. This process is called the first order moving average $[MA(1)]$. Show that $\{X_t\}$ is stationary.

$$Var(X_t) = Var(Z_t + \theta Z_{t-1}) = Var(Z_t) + \theta^2 Var(Z_{t-1}) + 0 = \sigma^2(1 + \theta^2) < \infty$$

$$E(X_t) = E(Z_t + \theta Z_{t-1}) = E(Z_t) + \theta E(Z_{t-1}) = 0 \text{ implies independent of time}$$

$$\begin{aligned} \gamma(h) = Cov(X_t, X_{t+h}) &= Cov(Z_t + \theta Z_{t-1}, Z_{t+h} + \theta Z_{t+h-1}) \\ &= Cov(Z_t, Z_{t+h}) + \theta Cov(Z_t, Z_{t+h-1}) + \theta Cov(Z_{t+1}, Z_{t+h}) + \theta^2 Cov(Z_{t-1}, Z_{t+h-1}) \\ &= \begin{cases} \text{if } h = 0 & \implies \sigma^2 + 0 + 0 + \theta^2\sigma^2 \\ \text{if } |h| = 1 & \implies \theta\sigma^2 \text{independent of t } \forall h \\ \text{if } |h| > 1 & \implies 0 \end{cases} \end{aligned}$$

This implies X_t is a stationary process. We can derive the auto-correlation function (ACf) of X_t.

$$f(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} \frac{\sigma^2(1+\theta^2)}{\sigma^2(1+\theta^2)} & h = 0 \\ \frac{\theta\sigma^2}{\sigma^2(1+\theta^2)} & |h| = 1 \\ 0 & |h| > 1 \end{cases} = \begin{cases} 1 & h = 0 \\ \frac{\theta}{1+\theta^2} & |h| = 1 \\ 0 & |h| > 1 \end{cases}$$

Scratch: Covariance is symmetric so

$$Cov(X_t, X_{t+h}) = Cov(X_{t+h}, X_t)$$

$$\gamma(h) = \gamma(-h)$$

Therefore, $\gamma(h)$ (hence $\rho(h)$) are even functions of h $\implies$ symmetric about y axis.

Example 6

Let $\{X_t\}$ be a stationary time series, satisfying the equation $X_t = \phi X_{t-1} + Z_t, t = 0, \pm 1, \pm 2, \cdots$, where $|\phi| < 1$ and $\{Z_t\} \sim WN(0, \sigma^2)$, Also, let Z_t and X_s be uncorrelated for each $s < t$. The time series $\{x_t\}$ is called autoregressive process of order 1 $[AR(1)]$. Derive autocorrelation function (acf) of X_t.

$$E(X_t) = E(\phi X_{t-1} + Z - t) = \phi E(X_{t-1})$$

where $\{X_t\}$ is stationary $\implies$ $E(X_t) = \mu$ for all t. Therefore, $\mu = \phi\mu \implies {}_{\phi \neq 0}\mu = 0 \implies E(X_t) = 0, \forall t$. Now, let us derive the auto covariance function of X_t. If $h = 0$, then $\gamma(h) = Var(X_t)$ and

$$\begin{aligned} \gamma(0) &= Cov(X_t, X_t) = Var(X_t) = Var(\phi X_{t-1} + Z_t) \\ &= \phi^2 Var(X_{t-1}) + Var(Z_t) + 2\phi Cov(X_{t-1}, Z_t) \\ &= \phi^2 Var(X_{t-1}) + Var(Z_t) \\ &= \phi^2 \gamma(0) + \sigma^2 \end{aligned}$$

Hence $\gamma(0) = \frac{\sigma^2}{1-\phi^2}$

If $h > 0$, multiply both sides of $X_t = \phi X_{t-1} + Z_t$ by X_{t-h} and take expectation:

$$E[X_t X_{t-h}] = \phi E[X_{t-h} X_{t-1}] + E[X_{t-h} Z_t]$$

$$Cov(X_t, X_{t-h}) = \phi Cov(X_{t-h}, X_{t-1}) + 0$$

Hence $\gamma(h) = \phi\gamma(h-1), h = 1, 2, 3, \cdots$.

$$\gamma(1) = \phi\gamma(0) = \phi\frac{\sigma^2}{1-\phi^2}$$

$$\gamma(2) = \phi\gamma(1) = \phi^2\gamma(0) = \phi^2 \frac{\sigma^2}{1-\phi^2}$$

$$\vdots$$

$$\text{(induction)} \implies \gamma(h) = \phi^h \frac{\sigma^2}{1-\phi^2}, h = 1, 2, 3, \cdots$$

For $h < 0$, do the trick above by multiplying both sides of $X_t = \phi X_{t-1} + Z_t$ by X_{t+h} and take expectation. Doing so, you will get

$$\gamma(h) = \phi^{-h} \frac{\sigma^2}{1-\phi^2}, h = -1, -2, -3, \cdots$$

This implies

$$\gamma(h) = \phi^{|h|} \frac{\sigma^2}{1-\phi^2}, h = 0, \pm 1, \pm 2, \pm 3, \cdots$$

Therefore, the acf is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^{|h|}, h = 0, \pm 1, \pm 2, \pm 3, \cdots$$

Example 7

$$\text{Sample auto-correlation function} \begin{cases} \text{Point estimation} \\ \text{confidence interval} \\ \text{Usual trends in the acf plot} \end{cases} \qquad \text{Linear regression} \implies$$

if time allowed.

Show that $X_t = 5 + 2t + Z_t$ where $\{Z_t\} \sim WN(0, \sigma^2)$ is not stationary. $E[X_t] = 5 + 2t$ which depends on t . This implies X_t is not stationary.

1.7 The Sample Autocorrelation function

What we have see so far on ACF, is based on given models (theoretical). In practice, based on the observed data $\{X_1, X_2, \cdots, X_n\}$ we use the sample ACF to assess the degree of dependence in the data sample ACF is the estimate of the theoretical ACF (under stationarity).

Definition. *Let $x_1, x_2, \cdots, x_n$ be observations of a time series. The sample mean of $x_1, \cdots, x_n$ is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The sample auto covariance function is*

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

The sample autocorrelation function is $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, -n < h < n$.

Convention: we will use $\hat{\gamma}(h)$ to show the estimate (value). Notice that in this estimate, x_t and $\bar{x}$ (lower case) are sued. The corresponding estimator (random variable) to $\hat{\gamma}(h)$ is shown by $\tilde{r}(h)$.

$$\tilde{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} (X_{t+|h|} - \bar{X})(X_t - \bar{X})$$

Throughout the course, $\hat{}$ is estimate and $\tilde{}$ is estimator.

- $\gamma(h) \implies$ Theoretical, fixed but unknown
- $\tilde{\gamma}(h) \implies$ The estimator, random variable
- $\hat{\gamma}(h) \implies$ realization of $\tilde{\gamma}(h)$ based on a sample

The sample ACF measures the correlation in the data under stationarity. Therefore, it can be used to check the uncorrelatedness of the residuals of a regression model

$$\text{residual} \sim_{iid} N(0, \sigma^2)$$

$$\text{independent} \implies \text{uncorrelated}$$

$$\text{Not uncorrelated} \implies \text{Not independent}$$

It can be shown for iid noise with finite variance,

$$\tilde{\phi}(h) \sim N(0, \frac{1}{n})$$

for large values of n, Therefore, for data from such process (iid noise) we expect that 95% of the sample ACFs fall between $51.96/\sqrt{n}$

$$\tilde{\rho}(h) \sim N(0, \frac{1}{n}) \implies Pr(\frac{-1.96}{\sqrt{n}} < \tilde{\rho}(h) < \frac{1.96}{\sqrt{n}}) = 0.95$$

Based on the trends in the plot of sample ACF ($\hat{\rho}(h)$ vs h), we will decide on different models for the data (to be discussed later).

Remark: for the observed data $\{x_1, x_2, \cdots, x_n\}$ if the data contains a trend (non-constant mean), $|\hat{\rho}(h)|$ will exhibit a slow decay (linear decay) as h increases. If the data contains a substantial deterministic periodic term, $\hat{\rho}(h)$ will exhibit similar behavior with the same period.

2 Forecasting and Regression

2.1 Review of simple and multiple linear regression and prediction interval

1. Use linear and non-linear regression to estimate trend component
2. Review the basics of simple and multiple linear regression
3. Look at model selection strategies and regression diagnostics
4. Look at forecasting and prediction using regression

Simple Linear Regression $y = \beta_0 + \beta_1 x + R$

Multiple Linear regression $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + R$

2.2 Simple Regression

$$Y_i | \{X_i = x\} = \alpha + \beta x + R_i$$

where

- $R_i \sim N(0, \sigma^2)$ and
- R_i are iid random variables
- Can use least squares or maximum likelihood to estimate α, β and σ

2.3 Confidence Interval

Refer this portion of notes to STAT 331 notes.

$Pr(a < \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} < b) = 0.95$.

For χ^2 distribution,

$$\left(\frac{(n-p-1)\hat{\sigma}^2}{b\chi^2_{n-p-1,0.975}}, \frac{(n-p-1)\hat{\sigma}^2}{a\chi^2_{n-p-1,0.025}}\right)$$

$$SSY = SSR + SSE$$

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1}$$

Therefore

$$\frac{SSR/P}{SSE/n-p-1} = \frac{SSR/P}{\hat{\sigma^2}} \sim F_{(p,n-p-1)}$$

2.4 Prediction

$$\begin{aligned} Y_i &= \alpha' + \beta x_i + R_i, R_i \sim N(0, \sigma^2) \\ &= \alpha' + \beta\bar{x} - \beta\bar{x} + \beta x_i + R_i \\ &= \alpha' + \beta\bar{x} + \beta(x_i - \bar{x}) + R_i \\ &= \alpha + \beta(x_i - \bar{x}) + R_i \end{aligned}$$

Therefore, $\begin{cases} \tilde{\alpha} \sim G(\alpha, \frac{\sigma}{\sqrt{n}}) \\ \tilde{\beta} \sim G(\beta, \frac{\sigma}{\sqrt{S_{xx}}} \end{cases}$
Therefore, $Y = \alpha + \beta(x_{new} - \bar{x}) + R_{new}$

2.5 Bias-variance Trade-off

- The linear model has a small prediction error at price = 700
- but at that point we see the model does not fit well
- The lack of flexibility of the linear model causes a bias in that region of the graph
- When predicting need to think about both bias and variance in prediction

2.6 Adjusted R^2

$$\tilde{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

n is sample size, p number of regressors (number of explanatory variables $x_1, \cdots, x_p$)

2.7 Akaike's Information Criterion AIC

$$\frac{SSR}{SST} = 1 - \frac{SSE}{SSY}$$

AIC is defined as

$$-2l(\hat{\theta}) + 2N_p$$

where l is log-likelihood, $\hat{\theta}$ the MLE and N_p the number of parameters in model. The smaller, the better.

2.8 Other Criteria

AICC Correction:

$$AICC = AIC + \frac{N_p(N_p+1)}{n-N_p-1}$$

Bayesian Information Criterion :

$$BIC = -2l(\hat{\theta}) + N_p \log(n)$$

2.9 Interpolation vs Extrapolation

1. When functional form is estimated mostly from observed data not all predictions will be reliable

2. Important question is to determine range of validity

3. If the explanatory variates are in range of validity this is called interpolation, otherwise extrapolation

4. Let $h_{max} = \max(H_{ij})$ where $H = X(X^T X)^{-1} X^T$. If the point x satisfies $x^T(X^T X)^{-1} x \leq h_{max}$, then estimating y for x is an interpolation problem, otherwise extrapolation.

2.9.1 Shapiro-Wilk test of Normality

1. QQ plot is a graphical method in testing Normality

2. A more formal non-parametric test is Shapiro-Wilk

3. $H_0 : Y_1, \cdots, Y_n$ come from a Gaussian distribution

4. Reject H_0 if the p-value of this test is small

5. In R: if the data is stored in the vector y, use the command shapio.test(y)

2.9.2 Difference Sign Test

1. Count the number S of values such that $y_i - y_{i-1} > 0$

2. For large iid sequence

$$\mu_S = E(S) = \frac{n-1}{2}, \sigma_S^2 = \frac{n+1}{12}$$

3. For large n, S is approximately $N(\mu_S, \sigma_S^2)$, therefore

$$\frac{S-\mu_S}{\sqrt{\sigma_S^2}} \sim N(0,1)$$

4. Large positive (negative) value of $S - \mu_S$ indicates the presence of increasing)decreasing) trend

5. Reject H_0: data is random if $|\frac{S-\mu_S}{\sqrt{\sigma_S^2}}| > Z_{1-\alpha/2}$

6. This may not work for data with strong seasonal component

2.9.3 Runs Test for Randomness

1. Estimate the median and call it $m \rightarrow$ in R: median(y) where y is the vector of data

2. n_1: number of observations $\geq M$.

3. n_2: number of observations $< m$

4. Count R the number of consecutive observations which are all smaller (larger) than m

5. For large iid sequence

$$\mu_R = E(R) = 1 + \frac{2n_1 n_2}{n_1 + n_2}, \sigma_R^2 = \frac{(\mu_R - 1)(\mu_R - 2)}{n_1 + n_2 - 1}$$

6. For large number of observations

$$\frac{R - \mu_R}{\sigma_R} \sim N(0, 1)$$

2.10 Smoothing Methods

2.10.1 Models with trend and seasonality

1. Recall the classical decomposition

$$X_t = m_t + s_t + Y_t$$

   where Y_t is stationary random noise component

2. m_t is the slowly changing function (trend component)

3. s_t is the periodic term with period d (seasonal component)

4. For identification need $\sum_{t=1}^{d} s_t = 0$ and $E(Y_t) = 0$

5. The assumption of linearity is strong amy or may not hold true.

2.10.2 Trend Estimation

Non seasonal model with trend

$$X_t = m_t + Y_t, t = 1, 2, \cdots, n$$

where $E(Y_t) = 0$. If $E(Y_t) \neq 0$, write

$$X_t = (m_t + E(Y_t)) + (Y_t - E(Y_t)), t = 1, 2, \cdots, n$$

2.10.3 Finite moving average filter

Let q be a nonnegative integer and consider the two-sided moving average of the series $X_t = m_t + Y_t$

$$\begin{aligned} W_t &= \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t-j} \\ &= \frac{1}{2q+1} \sum_{j=-q}^{q} [m_{t-j} + Y_{t-j}] \\ &= \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t-j} + \frac{1}{2q+1} \sum_{j=-q}^{q} Y_{t-j} \approx m_t \end{aligned}$$

2.10.4 Exponential Smoothing

$$\hat{m}_t = \alpha X_t + (1-\alpha)\hat{m}_{t-1}, 0 \leq \alpha \leq 1$$

$$\begin{aligned} \hat{m}_t &= \alpha X_t + (1-\alpha)[\alpha X_{t-1} + (1-\alpha)\hat{m}_{t-2}] \\ &= \alpha X_t + \alpha(1-\alpha)X_{t-1} + (1-\alpha)^2\hat{m}_{t-2} \\ &= \alpha X_t + \alpha(1-\alpha)X_{t-1} + (1-\alpha)^2[\alpha X_{t-2} + (1-\alpha)\hat{m}_{t-3}] \\ &= \alpha X_t + \alpha(1-\alpha)X_{t-1} + (1-\alpha)^2\alpha X_{t-2} + (1-\alpha)\hat{m}_{t-3} \end{aligned}$$

2.10.5 Trend elimination by difference

Example: $X_t = \alpha + \beta t + Y_t$ where α and β are constants ($\neq 0$) and $Y_t \sim_{iid} N(0, \sigma^2)$.

1. Is X_t stationary?

   $E[X_t] = \alpha + \beta t + E(Y_t) = \alpha + \beta t$

2. Is ∇X_t stationary?

   $\nabla X_t = (1-B)X_t = X_t - X_{t-1} = (\alpha + \beta t + Y_t) - (\alpha + \beta(t-1) + Y_{t-1}) = \beta + Y_t - Y_{t-1}$ where $Y_t^* \sim_{iid} N(\beta, 2\sigma^2)$

2.10.6 Estimate seasonality and trend

1. Estimate the trend m_t by applying a moving average filter specially chosen to eliminate the seasonal component and dampen the noise

2. Estimate the seasonal component by averaging over the seasons

3. Eliminate the seasonal component s_t from the data

2.10.7 Estimate seasonal component

1. For each $k = 1, \cdots, d$ estimate seasonal component

2. Compute the average w_k of

$$\{x_{k+jd} - \hat{m}_{k+jd} | q < k + jd \leq n - q\}$$

   For monthly data, this is averaging each month across the whole data

3. Normalise to get

$$\hat{s}_k = w_k - \frac{\sum_1^d w_j}{d}$$

   so that $\sum_{j=1}^d s_j = 0$

4. Notice that $\hat{s}_k = \hat{s}_{k-d}$ for $k > d$.

2.11 Holt-Winters Algorithm

1. This generalizes exponential smoothing to the case where there is a trend and seasonlity

2. Following Chatfield and Yar define trend as long-term change in the mean level per unit time

3. Have local linear trend where mean level at time t is

$$\mu_t = L_t + T_t t$$

   where L_t and T_t vary slowly through time.

4. L_t: the level, T_t: the slope of the trend at time t.

5. Holt's idea:

2.11.1 Holt-Winters method: Additive case

$$L_t = \alpha(X_t/l_{t-p}) + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$
$$l_t = \gamma(X_t/l_t) + (1-\gamma)l_{t-p}$$

The forecast for h periods ahead is then

$$L_t + hT_t + l_{t-p+h}$$

2.11.2 Holt-Winters method

1. Need to provide starting values for L_t, T_t, l_t at the beginning of the series
2. provide values for α, β and γ
3. choose between additive and multiplicative models

2.11.3 Special Cases

1. $\beta = \gamma = 0$: this is the case with no trend and no seasonal updates in the H-W algorithm
2. $L_t = \alpha + (1-\alpha)L_{t-1}$
3. This is the exponential smoothing with trend playing the role of the "history"
4. $\gamma = 0$: this is the case with no seasonal updates in the H-W algorithm
5. There are two corresponding H-W equations for updating the level L_t and the trend T_t.
6. H-W under $\gamma = 0$ is called double exponential smoothing.

$$\begin{cases} L_t = \alpha X_t + (1-\alpha)(L_{t-1} - T_{t-1}) \\ T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \end{cases}$$

2.11.4 Exponential Smoothing

$$m_t = \alpha Y_t + (1-\alpha)m_{t-1}$$

where

$$\begin{aligned} \hat{Y}_{t+1} &= \alpha Y_t + (1-\alpha)m_{t-1} \\ &= m_{t-1} + \alpha(Y_t - m_{t-1}) \\ &= \hat{Y}_t + \alpha(Y_t - \hat{Y}_t) \end{aligned}$$

where $\hat{Y}_t$ is predicted at time t and $Y_t - \hat{Y}_t$ is predicted error at time t.

3 Stationary & Linear processes

To perform any form of forecasting, there must be an assumption that "somethings" are the "same" in future as in the past. The idea of "being constant over time" is central to stationary processes, therefore, we will use stationary processes as the main framework to develop forecasting models.

In this chapter/module, we will talk about moving average ($MA(q)$), autoregressive ($AR(p)$) processes, and will look at the connection between the two, and will develop forecasting methods within stationary process.

The $MA(q)$ process: $\{X_t, t \in T\}$ is called a moving average process of order q if $X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$ where $\{Z_t\} \sim WN(0, \sigma^2)$ and $\theta_1, \cdots, \theta_q$ are constants. Sometimes Z_t is referred to "innovation". Notice that these innovations are uncorrelated ($Cov(Z_t, Z_s) = 0, t \neq s$. Constant variance ($Var(Z_t) = \sigma^2, \forall t$) and zero mean ($E[Z_t] = 0, \forall t$).

Deriving the mean and auto covariance function of $MA(q)$, it is easy to see that this process is stationary.

Definition. *The process $\{X_t\}$ is called q-dependent if X_t and Xs are dependent whenever $|t - s| > q$ if X_t and $X\ s$ are within q steps of each other, they are dependent.*

Clearly, an iid sequence of r.v.s is zero-dependent. Similarly, we say that a stationary time series is q-correlated if $\gamma(h) = 0$ whenever $|h| > 0$. Clearly, white noise is 0-correlated.

Example: Show that $MA(1)$ process is 1-correlated. Use

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma^2 & \text{if } h = 0, \gamma(h) = 0 \forall |h| > 1 \\ \theta\sigma^2 & \text{if } |h| = 1, \implies MA(1) \text{ is 1-correlated} \\ 0 & \text{if } |h| > 1 \end{cases}$$

It is easy to show that the $MA(q)$ process is q-correlated. The inverse of this statement is also true.

If $\{X_t : t \in T\}$ is stationary q-correlated time series, with mean 0. Then it can be represented as the $MA(q)$ process.

3.1 Autoregressive process AR(1)

Consider the process $\{X_t : t \in T\}$ defined by $X_t = \phi X_{t-1} + Z_t, t = 0, \pm 1, \cdots,$ where $\{Z_t\} \sim WN(0, \sigma^2)$ this process is called the first order autoregressive process we can also show this process by $(1 - \phi B)X_t = 2t$. Notice that if $|\phi| = 1$, then $\{X_t\}$ forms a random walk which we showed that it is not stationary, therefore depending on the value of ϕ, $\{X_t\}$ may or may not be stationary connection between $AR(1)$ and MA process. Consider the

$AR(1)$ process with the condition $|\phi| < 1$, then we have

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t = \phi(\phi X_{t-2} + Z_{t-1}) + Z_t = \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t \\ &= \phi^3 X_{t-3} + \phi^2 Z_{t-2} + \phi Z_{t-1} + Z_t \cdots = Z_t + \sum_{i=1} \phi^i Z_{t-i} = \sum_{i=0}^{n} \phi^i Z_{t-i} \end{aligned}$$

Define $\theta_i = \phi^i$, we have written X_t as an $MA(\infty)$.

Autoregressive of order P: $AR(P)$

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, \{Z_t\} \sim WN(0, \sigma^2)$$

Definition. *$\{X_t : t \in T\}$ is called a Gaussian time series if all its joint definitions are multivariate normal, i.e., for any set $i_1, i_2, \cdots, i_n (n \in \mathbb{N})$ the random vector $(X_{i1}, \cdots, X_{in})$ has a multivariate normal distribution.*

Example: Consider the stationary gaussian time series $\{X_t : t \in T\}$. Suppose X_n has been observed and we want to forecast X_{n+h} using $m(X_n)$ function of X_n. Let's measure the quality of a forecast by

$$MSE = E([X_{n+h} - m(X_n)]^2 | X_n)$$

It can be shown that the function $m(\cdot)$ which minimize MSE in a general case (not necessarily Gaussian) is

$$m(X_n) = E[X_{n+h} | X_n]$$

Stationarity implies that $E(X_{n+h}) = E(X_n) = \mu$. Also $Cov(X_{n+h}, X_n) = Cov(X_r, X_n) = \gamma(\sigma) = Var(X_{n+h}) = Var(X_n)$.

$$Corr(X_{n+h}, X_n) = \frac{Cov(X_{n+h}, X_n)}{\sqrt{Var(X_{n+h}) Var(X_n)}} = \frac{\gamma(h)}{\gamma(0)} = \rho(h)$$

Hence

$$(X_{n+h}, X_n) \sim MVN([\mu\mu]^T, \begin{pmatrix} \sigma^2 & \rho(h)\sigma^2 \\ 0 & \sigma^2 \end{pmatrix}$$

$$X_{n+h} | X_n = X \sim N(\mu + \sqrt{\frac{\gamma(0)}{\gamma(0)}} \rho(h)(X - \mu), \sigma^2(1 - \rho(h)))$$

$$\begin{aligned} m(X_n) &= m(X_{n+h} | X_n) \\ &= \mu + \rho(h)(X_n - \mu) \implies \rho(h) X_n + (1 + \rho(h))\mu = aX_n + b \end{aligned}$$

$$MSE = E[(X_{n+h} - E(X_{n+h} | X_n))^2 | X_n] = Var(X_{n+h} | X_n) = \sigma^2(1 - \rho(h))$$

In general, looking at prediction of the form $aX_n + b$ which is a linear function of history, is of interest. In previous example, knowing mean and correlation function result in this linear predictor. Even if the normality assumption does not hold true. We can still look at predictor $aX_n + b$ where a, b are computed in the form of

$$min_{a,b}E[(X_{n+h} - aX_n - b)^2|X_n]$$

MA(q) process with mean 0, it is a q-correlated process which is why ACF is 0 after lag q. AR(1) is an exponential decay and asymptotically approaches 0. Therefore, $MA(\infty) = AR(1)$.

In a gaussian process, $m(X_n) = \mu + \rho(h)(X_n - \mu)$. If we go with gaussian assumption, then select the best linear process, not necessarily the best assumption.

$$aX_n + b$$

Even if the normality assumption does not hold true, we still can look at the predictor $m(X_n) = aX_n + b$ when a and b are computed from

$$\min E[[X_{n+h} - aX_n - b]^2]$$

3.2 Linear Prediction

We can consider the problem of predicting $X_{n+h}, h > 0$ for a stationary time series with known mean μ and know ACVF $\gamma(\cdot)$, based on previous values $\{X_n, X_{n-1}, \cdots, X_1\}$

Showing the linear predictor of X_{n+h} by P_nX_{n+h}, we are interested in

$$P_nX_{n+h} = a_0 + a_1X_n + a_2X_{n-1} + \cdots + a_nX_1$$

which minimizes

$$S(a_0, \cdots, a_n) = E[(X_{n+h} - P_nX_{n+h})^2]$$

To get $a_0, \cdots, a_n$, we need to solve the system $\frac{dS}{da_j} = 0, j = 1, 2, \cdots, n$. Doing so, we get $a_0 = \mu(1 - \sum_{i=1}^{r} a_i), \Gamma_n a_n = \gamma_n(h)$, wehere $a_n = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$.

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & \ddots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \gamma(0) \end{pmatrix}, \gamma_n(h) = \begin{pmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+n) \end{pmatrix}$$

$$\implies P_nX_{n+h} = a_0 + \sum_{i=1}^{n} a_iX_{n-i+1} = \mu(1 - \sum_{i=1}^{n} a_i) + \sum_{i}^{n} a_iX_{n-i+1} = \mu + \sum_{i=1}^{n} a_i(X_{n+1-i} - \mu)$$

Some properties

1. $P_n X_{n+h}$ is defined by $\mu, \gamma(h)$

2. It can be shown that
$$E[(X_{n+h} - P_n X_{n+h})^2] = \gamma(0)$$

3. $E[X_{n+h} - P_n X_{n+h}] = 0$ (prediction error on average is 0

4. $E[(X_{n+h} - P_n X_{n+h})X_j] = 0, j = 1, 2, 3, \cdots, n.$

In a more general setup, suppose that Y and $W_1, \cdots, W_n$ are any random variables with finite second moments and means $\mu_Y = E[Y], \mu_i = E[W_i]$ and $Cov(Y, Y)$, $Cov(Y, W_i)$, $i = 1, \cdots, n$. $Cov(W_i, W_j)$ are all known. Define $W = (W_n, \cdots, W_1)$ and $\mu(W) = (\mu_n, \cdots, \mu_1)^T$. and $\gamma = Cov(Y, W) = (Cov(Y, W_n), \cdots, Cov(Y, W))^T$ and $\Gamma = Cov(W, W) = [Cov(W_{n+1-i}, W_{n+1-j})]_{i,j=1}^n$

Now, by the some argument used in the derivation of $P_n X_{n+h}$, the "best" linear predictor of Y in terms of $\{W_n, W_{n-1}, \cdots, W_1\}$ is $P_W Y = P(Y|W) = \mu_Y + a_n^T(W -_W)$ where $a_n = (a_1, \cdots, a_n)^T$ is the solution of $\Gamma a = \gamma$.

Example 10: Derive the one-step prediction for AR(1) model.

Suppose $X_t = \phi X_{t-1} + Z_t$, where $|\phi| < 1$ and $\{Z_t\} \sim WN(0, \sigma^2)$. In example 6, we showed that
$$\gamma(h) = \phi^{|h|}\gamma(0), h = 0, 1, 2, \cdots$$

Also $E[X_t] = \mu = 0$. To find the linear predictor, we need to solve:

$$\Gamma_n a_n = \gamma_n(h)$$
$$\begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & \ddots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+n) \end{pmatrix}$$

We divide both side by $\gamma(0)$.

$$\begin{pmatrix} 1 & \phi & \cdots & \phi^{n-1} \\ \phi & 1 & \cdots & \phi^{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \cdots & \cdots & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^n \end{pmatrix}$$

An obvious solution to this system is $a_n = \begin{pmatrix} \phi \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

$$P_nX_{n+1} = Y + \sum_{i=1}^{n} a_i(X_{n+1-i} - Y) = \sum_{i=1}^{n} a_iX_{n+1-i} = a_iX_n + 0$$

Therefore

$$P_nX_{n+1} = \phi X_n$$

.

$$MSE = E[[X_{n+1} - P_nX_{n+1}]^2] = E[[X_{n+1} - \phi X_n]^2] = E[Z_{n+1}^2] = Var(Z_{n+1}^2) = \sigma^2$$

You can also use the formula for MSE to calculate it.

$$MSE = \gamma(0) - a_n^T\gamma_n(h) = \gamma(0) - \begin{pmatrix}\phi & 0 & 0 & \cdots & 0\end{pmatrix}\begin{pmatrix}\gamma(1)\\ \gamma(2)\\ \vdots\\ \gamma(n)\end{pmatrix}$$
$$= \gamma(0) - \phi\gamma(1) = \gamma(0) - \phi^2\gamma(t) = (1-\phi^2)\gamma(0) = \sigma^2$$

3.2.1 Properties of linear predictor

Suppose $E[W^2] < \infty, E[V^2] < \infty, \Gamma = Cov(W, W), B, \alpha_1, \cdots, \alpha_n$

1. $P(V|W) = E[V] + a_n^T(W - \mu_w)$ where $Pa_n = \gamma$.
2. $E[U - P(U|W)W] = 0$ and $E[U - P(U|W)] = 0$
3. $E[(U - P(U|W))^2] = Var(V) - a_n^TCov(U, W)$
4. $P_w(a_1U + a_2V + B) = a_1P_w(U) + a_2P_w(V) + B$
5. $P(\sum_{i=1}^n a_iW_i + B|W) = \sum_{i=1}^n a_iw_i + B$
6. $P(U|W) = E[U]$ if $Cov(V, W) = 0$

3.3 Linear Processes

We have discussed linear prediction in which future values are predicted by linear combination of historical values. This section focuses on a class of linear time series which provides a general framework for studying stationary processes.

Definition. *The time series $\{X_t\}$ is a linear process if*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

for all t, where $\{Z_t\} \sim WN(0, \sigma^2)$ and ψ_j is a sequence of constants such that $\sum_{j=-\infty}^{\infty} \psi_j < \infty$.

Example 11: show that $AR(1)$ with $|\phi| < 1$ is a linear process. We know that $X_t = \phi X_{t-1} + Z_t$. We showed before that

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$$

Since $|\phi| < 1 \implies \sum_{j=0}^{\infty} |\phi|^j < \infty \implies \sum_{j=0}^{\infty} |\phi^j| < \infty$. Therefore all assumptions in the definition above are satisfied and AR(1) is a linear process.

For prediction purposes we may not want to have dependence on the future innovations (Z_ts). However, the general form $\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ involves future innovations.

Definition. *A linear process $\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ is called causal or future independent if $\psi_j = 0, \forall j < 0$.*

Examples: $AR(1) \implies X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$. Then

$$MA(q) : X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}$$

3.4 Box-Jenkins Models

The Box-Jenkins Methodology uses ARMA and ARIMA models for forecasting. The class of ARMA models tries to balance goodness of fit with a limited number of parameters. Whenever the series is not stationary, ARIMA models (ARMA with differecing) are used. When seasonal effect is present, the more general SARIMA model will be used all theres models use two key functions: ACF and PACF.

Definition. $\{X_t, t \in T\}$ *is an ARMA(p, q) process if*

1. $\{X_t, t \in T\}$ *is stationary.*
2. $X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$ *where* $\{Z_t\} \sim WN(0, \sigma^2)$.
3. *Polynomials* $(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p)$ *and* $(1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q)$ *have no common factors (no common root).*

$\{X_t, t \in T\}$ is an ARMA process with mean μ if $\{X_t - \mu\}$ is an ARMA(p, q) process. Recall the backward shift operator $BX_t = X_{t-1}$. By iteration we have $B^j X_t = X_{t-j}$. Therefore, we can write ARMA(p, q) process as $(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)X_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)Z_t$ and that is $\theta(B)X_t = \theta(B)Z_t$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$.

The general model in above equation has a unique stationary solution for X_t if $\phi(z) = 1 - \phi_1 z - \cdots - \theta_p z^p \neq 0$ for all complex z such that $|z| = 1$. Recall that a complex number z is $z = a + bi$ where $i = \sqrt{-1}$ and $|z| = \sqrt{a^2 + b^2}, a, b \in \mathbb{R}$.

If for all z such that $|z| = 1$ we have $\phi(z) \neq 0$, then there exists $\delta > 0$ such that

$$\frac{1}{\phi(z)} = \sum_{j=-\infty}^{\infty} x_j z^j, 1 - \delta < |z| < 1 + \delta$$

$$\sum_{j=-\infty}^{\infty} |x_j| < \infty$$

Under this condition

$$\frac{1}{\phi(B)} = \sum_{j=-\infty}^{\infty} x_j B^j$$

is a linear filter. Hence

$$\phi(B)X_t = \theta(B)Z_t \implies X_t = \frac{1}{\theta(B)}\theta(B)Z_t$$

since $\frac{1}{\theta(B)}$ is a polynomial and $\theta(B)$ is a polynomial. Thus $\frac{1}{\phi(B)}\theta(B) = \psi(B)$ is a polynomial. Therefore

$$X_t = \frac{1}{\phi(B)}\theta(B)Z_t = \psi(B)Z_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

where $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

3.5 Causality

An ARMA(p, q) process $\phi(B)X_t = \theta(B)Z_t$ where $\{Z_t\} \sim WN(0, \sigma^2)$ is causal if there exists constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = \psi(B)Z_t, \forall t$. This condition is equivalent to

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \neq 0, \forall z \text{such that } |z| \leq 1$$

Causal $\iff$ roots of $\phi(z)$ are outside the unit circle. If the condition above holds true, then $\frac{\theta(B)}{\phi(z)} = \psi(z) \implies \theta(z) = \phi(z)\psi(z), \forall z$. This implies

$$1 + \theta_1 z + \cdots + \theta_q z^q = (-\phi_1 z - \cdots - \phi_p z^p)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots)$$

where $1 = \psi_0, \theta_1 = \psi_1 - \phi_1\psi_1, \cdots$

Note:

1. If $\phi(z) = 1 \implies \phi(B)X_t = \theta(B)Z_t$ reduces to

$$X_t = \theta(B)Z_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

   Therefore, this is a MA(q)

2. If $\theta(B) = 1$ we have $\theta(B)X_t = Z_t \implies X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$. This is AR(p).

   Notice that AR(p) and MA(q) are special cases of ARMA(p, q) processes.

$$AR(p) = ARMA(p, 0)$$

$$MA(q) = ARMA(0, q)$$

3.6 Invertibility

An ARMA(p, q) process $\{X_t\}$ is invertible if there exist constants $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and $Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = \pi(B)X_t, \forall t$. Invertibility is equivalent to the condition

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q \neq 0, \forall z \text{ such that } |z| \leq 1$$

We have that

$$\frac{\phi(z)}{\theta(z)} = \pi(z) \implies \phi(z) = \theta(z)\pi(z)$$

$$(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = (1 + \theta_1 z + \cdots + \theta_q z^q)(\pi_0 + \pi_1 z + \pi_2 z^2 + \cdots)$$

where $\pi_0 = 1, -\phi_1 = \pi_0\theta + \pi_1$

Example 12: consider $\{X_t, t \in T\}$ satisfying $X_t - 0.5X_{t-1} = Z_t + 0.4Z_{t-1}$ where $\{Z_t\} \sim WN(0, \sigma^2)$. Investigate the causality and invertibility of X_t. If the series is causal (invertible) provide the causal (invertible) solution. (there are also called MA(∞) and AR(∞) representations)

Causality $\phi(z) = 1 - 0.5z$ and $\phi(z) = 0$. This implies $1 - 0.5z = 0 \implies z = 2, |z| = 2 > 1$. The root is outside the unit circle so X_t is causal. $\theta(z) = \phi(z)\psi(z)$. $1 + 0.4z = (1 - 0.5z)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots)$. $\psi_0 = 1, \psi_1 - 0.5\psi_0 = 0.4 \to \psi_1 = 0.9$ and $\psi_2 - 0.5\psi_1 = 0 \to \psi_2 = 0.5 \times 0.9$ and $\psi_3 - 0.5\psi_2 = 0 \to \psi_3 = 0.5^2 \times 0.9, \cdots$

Therefore, $\begin{cases} \psi_0 = 1 \\ \psi_j = 0.5^{j-1} \times 0.9, j = 1, 2, 3, \cdots \end{cases}$

The causal solution is $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = Z_t + 0.9 \sum_{j=1}^{\infty} 0.5^{j-1} Z_{t-j}$

Invertibility : $\theta(z) = 0 \to 1 + 0.4z = 0 \implies z = -\frac{1}{0.4} = -10/4$ and then $|z| = \frac{10}{4} > 1$.

The set of $\theta(z)$ is outside the unit circle. Then X_t is invertible.

$$\phi(z) = \theta(z)\pi(z)$$

$$(1 - 0.5z) = (1 + 04z)(\pi_0 + \pi_1 z + \pi_2 z^2 + \cdots)$$

$$\pi_0 = 1$$

$$\pi_1 + 0.4\pi_0 = -0.5 \to \pi_1 = 0.9$$

$$\pi_2 + 0.4\pi_1 = 0 \to \pi_2 = (-0.4)(-0.9)$$

$$\vdots$$

$$\pi_j = -0.9 \times (-0.4)^{j-1}, j = 1, 2, 3, \cdots$$

Therefore, the AR(∞) representation of X_t is

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = X_t - 0.9 \sum_{j=1}^{\infty} (-0.4)^{j-1} X_{t-j}$$

$$Z_t = X_t - 0.9 \sum_{j=1}^{\infty} (-0.4)^{j-1} X_{t-j}$$

3.7 ACVF of ARMA(p, q)

Consider a causal, stationary ARMA process $\phi(B)X_t = \theta(B)Z_t$, $\{Z_t\} \sim WN(0, \sigma^2)$. The MA($\infty$) representation of X_t is

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

where $E[X_t] = 0, \forall t$. We have,

$$\gamma(h) = Cov(X_t, X_{t+h}) = E(X_t, X_{t+h}) - E(X_t)E(X_{t+h}) = E[\sum_{j=0}^{\infty} \psi_j Z_{t-j} \sum_{j=0}^{\infty} \psi_j Z_{t+h-j}]$$

Notice that $E(Z_t Z_s) = 0, \forall t \neq s$. Then $(Cov(Z_t, Z_s) = 0, \forall t \neq s)$.

If $h \geq 0$:

$$\gamma(h) = \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \sigma^2$$

If $h < 0$,

$$\gamma(h) = \sum_{j=0}^{\infty} \psi_j \psi_{j-h} \sigma^2$$

Then

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$$

Example 15: Derive the ACVF for the following ARMA(1, 1) process

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ and $|\phi| < 1$. If $|\phi| < 1$ we show that X_t is casual. $\phi(z) = 0 \implies 1 - \phi z = 0 \to z = \frac{1}{\phi}$. $|\frac{1}{\phi}| > 1 = 0$ casual.

$$\theta(z) = \theta(z)\psi(z)$$

$$(1 + \theta z) = (1 - \phi z)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots)$$

$$\psi_0 = 1, \psi_1 - \phi\psi_0 = \theta \to \psi = \phi + \theta$$

$$\psi_2 - \psi_1 \phi = 0 \to \psi_2 = \phi(\phi + \theta)$$

$$\cdots, \psi_j = \phi j - 1(\phi + \theta), j = 1, 2, 3, \cdots$$

If $h = 0$, then

$$\begin{aligned}
\gamma(0) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 &= \sigma[1 + \sum_{j=1}^{\infty} \psi_j^2] \\
&= \sigma^2[1 + (\theta + \phi)^2 \sum_{j=1}^{\infty} \phi^{2(j-1)}] = \sigma^2[1 + (\theta + \phi)^2 \sum_{i=0}^{\infty} \phi^{2i}] \\
&= \sigma^2[1 + (\theta + \phi)^2 \frac{1}{1 - \phi^2}]
\end{aligned}$$

where $i = j - 1$

If $h \neq 0$,

$$\begin{aligned}
\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|} &= \sigma^2[\psi_0 \psi_{|h|} + \sum_{j=1}^{\infty} \psi_j \psi_{j+|h|}] \\
&= \sigma^2[\phi^{|h|-1}(\theta + \phi) + (\theta + \phi)^2 \sum_{j=1}^{\infty} \phi^{j-1} \phi^{j+|h|-1}] \\
&= \sigma^2[\phi^{|h|-1}(\theta + \phi) + \phi^{|h|}(\theta + \phi)^2 \sum_{j=1}^{\infty} \phi^{2(j-1)}] \\
&= \sigma^2[\phi^{|h|-1}(\theta + \phi) + \phi^{|h|} \frac{(\theta + \phi)^2}{1 - \phi^2}]
\end{aligned}$$

with $i = j - 1$.

Example 16: Derive the ACVF of an AR(1) process ($|\phi| < 1$) using the general format $\sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$ of ARMA processes.

Example 17: Derive the ACVF of an MA(q) process.

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

$$\theta(z) = \theta(z)\psi(z)$$

$$(1 + \theta_1 z + \cdots + \theta_q z^q) = |x(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots)$$

Therefore, $\psi_0 = 1, \psi_1 = \theta_1, \psi_2 = \theta_2, \cdots \psi_q = \theta_q, \psi_j = 0, \forall j > q$. Therefore,

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|} = \begin{cases} \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} & |h| \le q \\ 0 & |h| > q \end{cases}$$

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} / \sigma^2 \sum_{j=0}^{q} \theta_j^2 & |h| > q \\ 0 & O.W. \end{cases} = \begin{cases} \frac{\sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|}}{\sum_{j=0}^{q} \theta_j^2} & |h| > q \\ 0 & O.W. \end{cases}$$

We can see that $\gamma(h) = 0$ after q lags, confirming that the process is an MA(q) [q-corelatedness]. However, there are models with infinite number of non-zero values of $\gamma(h)$ (e.g. AR(p)). Therefore, it is useful to introduce another tool to help us identify time series models.

3.8 Partial Autocorrelation Function (PACF)

ACF measures the correlation between X_n and X_{n+h}. This correlation can be due to direct connection, or through the intermediate steps $X_{n+1}, X_{n+2}, \cdots, X_{n+h-1}$. PACF looks at the correlation between X_n and X_{n+h} once the effect of the intermediate steps are removed.

We remove the effect of the intermediate steps by deriving the linear predictor

$P(X_{n+h}|X_{n+1}, \cdots, X_{n+h-1})$ and $P(X_n|X_{n+1}, \cdots, X_{n+h-1})$. The partial auto-correlation function (PACF) is shown by $\alpha(h)$ and is defined to be

$$\alpha(h) = \begin{cases} 1 & \text{if } h = 0 \\ Corr(X_n, X_{n+1}) = \rho(1) & \text{if } h = 1 \\ Corr[X_n - P(X_n|X_{n+1}, \cdots, X_{n+h-1}), X_{n+h} - P(X_{n+h}|X_{n+1}, \cdots, X_{n+h-1})] \end{cases}$$

Example 18: Derive the PACF for an AR(1) process ($|\phi| < 1$). We saw in example 10 that $P(X_{n+1}|X_n) = \phi X_n$ where $X_t = \phi X_{t-1} + Z_t$ is an AR(1) process. $h = 0 \implies \alpha(0) = 1, h = 1 \implies \alpha(1) = Corr(X_t, X_{t+1}) = Corr(X_t, X_{t+1}) = \rho(1) = \phi$.

Then $h = 2$:

$$\begin{aligned}\alpha(2) &= Corr[X_t - P(X_t|X_{t+1}), X_{t+2} - P(X_{t+2}|X_{t+1}) \\ &= Corr[X_t - P(X_t|X_{t+1}), X_{t+2} - \phi X_{t+1}], x_{t+1} \\ &= Corr[\text{linear function of } X_{t+1}, Z_{t+2}] = 0\end{aligned}$$

Similarly, $\alpha(h) = 0, \forall h > 0$. Therefore,

$$\alpha(h) = \begin{cases} 1 & h = 0 \\ \phi & h = 1 \\ 0 & h \geq 2 \end{cases}$$

Notice that similar to ACF, the PACF is symmetric in h so $h < 0$ is omitted from derivations above.

Theorem. $\{X_t, t \in T\}$ is a causal AR(p) process if and only if its PACF has the following properties

$$\alpha(p) \neq 0, \alpha(h) = 0, \forall h > p$$

Furthermore, $\alpha(p) = \phi_p$.

This theorem shows that PACF is a powerful tool for identifying AR(p) processes. In fact, ACF to MA(q) is like the PACF to AR(q) from the visual point of view (trend). In summary

	ACF	PACF
MA(q)	zero after lag q	decays exponentially
AR(p)	decays exponentially	zero after lag p

In the general case of ARMA processes, the PACF is defined as $\alpha(0) = 1$ and $\alpha(h) = \Phi_{hh}, h \geq 1$ where Φ_{hh} is the last component of the vector $\Phi_h = \Gamma_h^{-1}\gamma_h$ in which

$$\gamma_h = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(h-1) \\ 0 & \gamma(0) & \cdots & \gamma(h-2) \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \gamma(0) \end{pmatrix}$$

Based on observations (data) $\{x_1, \cdots, x_n\}$ with $x_i \neq x_j$ for $i, j = 1, \cdots, n, i \neq j$, the sample PACF $\hat{\alpha}(h)$ is given by

$$\hat{\alpha}(0) = 1, \hat{\alpha}(h) = \hat{\Phi}_{hh}, h \geq 1$$

where $\hat{\Phi}_{hh}$ is the last component of $\hat{\Phi}_h = \hat{\Gamma}_h^{-1}\hat{\gamma}_h$

3.9 ARMA(p, d, q) process

Definition. *Let d be a non-negative integer $\{X_t, t \in T\}$ is an ARIMA(p, d, q) process if $Y_t := (1-B)^d X_t$ is a causal ARMA(p, q) process.*

The definition above means that $\{X_t\}$ satisfies an equation of the form

$$\phi^*(B)X_t = \phi(B)(1-B)^d X_t = \theta(B)Z_t, \{Z_t\} \sim WN(0, \sigma^2)$$

Notice that if $d \neq 0 \implies \phi^*(1) = 0 \implies X_t$ is not stationary. Therefore, $\{X_t\}$ is stationary if and only if $d = 0$, in which case it is reduced to an ARMA(p, q) process.

If $\{X_t\}$ exhibits a trend which is well-approximated by a polynomial $m(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_d t^d$, then $(1-B)^d X_t$ will not have a time-dependent trend. Therefore, ARIMA models are appropriate when the non-stationary is due to the existence of a trend.

Example 21:

Consider the process $X_t = 0.8Z_{t-1} + 2t + Z_t$ where $\{Z_t\} \sim WN(0, \sigma^2)$

Write this process as ARIMA(p, d, q) process.

$$X_t - 0.8X_{t-1} = Z_t + 2t \rightarrow (1 - 0.8B)X_t = 2t + Z_t$$

2t is a linear trend, so let us look at $(1-B)X_t$.

$$\begin{aligned}
\nabla X_t &= (1-B)X_t \\
&= X_t - X_{t-1} \\
&= 0.8X_{t-1} + Z_t + 2t - 0.8X_t - 2Z_{t-1} - 2(t-1) \\
&\implies X_t - X_{t-1} = 0.8(X_{t-1} - X_{t-2}) + Z_t - Z_{t-1} + 2 \\
&\implies Y_t - 0.8Y_{t-1} = Z_t - Z_{t-1} + 2 \\
&\implies (Y_t - 10) - 0.8(Y_{t-1} - 10) = Z_t + Z_{t-1}
\end{aligned}$$

Since Y_t is an ARIMA(1,1) with mean 10, then X_t is an ARIMA(1,1, 1) process.

We have see how differencing can be bused to remover a trend. Seasonality is a particular type of trend which can be removed by a particular type of differencing . This is discussed under SARIMA (seasonal ARIMA) model.

3.10 SARIMA(p, d, q) × (P, D, Q) process

Recall the operator B, where $BX_t = X_{t-1}$ and $B^k X_t = X_{t-k}$. Examples: $B^2 X_t = X_{t-2}, B^2 X_t = X_{t-12}$. Hence

$$(1-B)^2 X_t = (1 - 2B + B^2)X_t = X_t - 2X_{t-1} + X_{t-2} \implies \text{2 times of differencing}$$

$$(1-B)^2 X_t = X_t - X_{t-2} \implies \text{Differencing in lag 2}$$

Therefore $(1-B^k)$ and $(1-B)^k$ are different fitness. The latter is performing k times of differencing, but the former is differencing one time in lag k.

In R we have:

```
diff(x, difference= k)
diff(x, lag=k)
```

As an example, consider the process $\{X_t\}$ where t represents the month.

if there is a seasonal effect for month, i.e. $S(t) = S(t-12)$ then the effect of the seasonal trend for X_t and X_{t-12} should be the same as they are exactly 12 steps apart (period = 12). Therefore, one may hope that $Y_t = X_t - X_{t-12}$ does not exhibit any seasonal trend.

We have seen how differencing can be used to remove a trend . We talked about "seasonal differencing" to remove the effect of a periodic trend. If we apply differencing at lag s

$$((1 - B^s)X_t = X_t - X_{t-s})$$

where s represents the season we can (in theory) remove the effect of the seasonal trend. Therefore, fitting and ARMA(p, q) model to the differenced series $Y_t = (1 - B^s)X_t$ is the same as fitting the model

$$\phi(B)(1 - B^s)X_t = \theta(B)Z_t, \{Z_t\} \sim WN(0, \sigma^2)$$

This is a special case of SARIMA models.

Definition. *If d, D are non-negative integers, then $\{X_t, t \in T\}$ is a seasonal ARIMA(p, d, q)$\times(P, D, Q)_s$ process with a period s if the differenced series*

$$Y_t = \nabla^d \nabla_s^D X_t = (1 - B)^d (1 - B^s)^D X_t$$

is a causal ARMA process defined by

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t, \{Z_t\} \sim WN(0, \sigma^2)$$

where

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

$$\Phi(z)i = 1 - \Phi_1 z - \cdots - \Phi_p z^p$$

$$\theta(z)i = 1 + \theta_1 z + \cdots + \theta_p z^p$$

$$\Theta(z)i = 1 + \Theta_1 z + \cdots + \Theta_p z^p$$

Remark 1: Notice that the process $\{X_t, t \in T\}$ is causal if and only if $\phi(z) = 0$ and $\Phi(z) \neq 0, \forall z, |z| \leq 1$.

Remark 2: In practice D is rarely more than 1 and P, Q are typically less than 3.

Example 22: Write down the equation form of $ARMA(1, 1)_{12}$ process.

$$ARMA(1, 1)_{12} = SARIMA(0, 0, 0) \times (1, 0, 1)_{12}$$

$$\phi(B)\Phi(B^s)\nabla^d \nabla_s^D X_t = \theta(B)\Theta(B^s)Z_t$$

$$1 \times (1 - \Phi_1 B^{12})(1 - B)0(1 - B^{12})^0 X_t = 1 \times (1 + \Theta_1 B^{12}) Z_t$$
$$\implies (1 - \Phi_1 B^{12}) X_t = (1 - \Theta_1 B^{12}) Z_t, \{Z_t\} \sim WN(0, \sigma^2)$$

If $d \neq 0$ or $D \neq 0$, then SARIMA models are not stationary. This model ($ARMA(1,1)_{12}$) looks like ARMA(1, 1): $(1 - \phi B)X_t = (1 + \theta B)Z_t$. In fact, this model is an ARMA(1, 1) sitting on the season.($s = 12$).

Example 23: Derive the ACF of $SARIMA(0,0,1)_{12} = SARIMA(0,0,0) \times (0,0,1)_{12}$

$$\begin{aligned} d = D = 0 &\implies \phi(B)\Phi(B^s)X_t = \theta(B)\Theta(B^s)Z_t \\ &\implies 1 \times 1 \times X_t = 1 \times (1 + \Theta_1 B^{12})Z_t \\ &\implies X_t = Z_t + \Theta_1 Z_{t-12}, \{Z_t\} \sim WN(0, \sigma^2) \end{aligned}$$
$$\gamma(h) = Cov(X_t, X_{t+h}) = \begin{cases} (1 + \Theta_1^2)\sigma^2 & h = 0 \\ \Theta_1 \sigma^2 & h = 12 \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} 1 & h = 0 \\ \frac{\Theta_1}{1+\Theta_1^2} & h = 12 \\ 0 & \text{otherwise} \end{cases}$$

Aside, for MA(1),

$$\rho(h) = \begin{cases} 1 & h = 0 \\ \frac{\theta}{1+\theta^2} & h = 1 \\ 0 & \text{otherwise} \end{cases}$$

Example 24: Write down the $ARIMA(0,1,1) \times (0,1,1)_{12}$ is the equation format.

To use Box-Jenkins methodology,

1. check for seasonal and non-seasonal trends (stationarity)
2. use differencing to make the process stationary.
3. Identify p, q, P, Q: visually (from ACF and PACF) and/or with formal model selection methods
4. forecast the future with the appropriate model.

4 Parameter Estimation in ARMA processes

This section concentrates on estimation of the parameters $\phi_i, i = 1, 2, \cdots, p$ and $\theta_j, j = 1, \cdots, q$ and σ^2 (the variance of W.N.) in the ARMA(p, q) process $\phi(B)X_t = \theta(B)Z_t, \{Z_t\} \sim WN(0, \sigma^2)$. We assume that p and q have been correctly specified. If the mean of the series is not zero, we will use the model $\phi(B)(X_t - \mu) = \theta(B)Z_t$ where $\mu = E[X_t], \forall t$. Also $\tilde{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The common parameter estimation methods are maximum likelihood, least squares, Yule-Walker, Innovations algorithm, Durbin-Levinson method.

4.1 Yule-Walker estimation in AR(p)

Consider a causal AR(p) model $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$ with causal solution $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ where $\{Z_t\} \sim WN(0, \sigma^2)$. Multiplying both sideby $X_{t-j}, j = 0, 1, 2, \cdots, p$ and taking expectation we have

$$E[X_t X_{t-j}] - \phi_1 E[X_t X_{t-j}] - \cdots \phi_p E[X_{t-p} X_{t-j}] = E[Z_t X_{t-j}], i = 0, 1, 2, \cdots, p$$

Since $E[X_t] = 0, \forall t$ is simplified to

$$\gamma(j) - \phi_1\gamma(j-1) - \cdots - \phi_p\gamma(j-p) = E[Z_t X_{t-j}]$$
$$j = 0 \implies E[Z_t X_{t-j}] = E[Z_t X_t] = E[Z_t \sum_{j=0}^{\infty} \psi_j Z_{t-j}] = E[Z_t^2] = \sigma^2$$
$$j > 0 \implies E[Z_t X_{t-j}] = 0$$

We have for $j = 0$, $\gamma(0) - \phi_1\gamma(1) - \cdots - \phi_p\gamma(p) = \sigma^2$
for $j = 1$, $\gamma(1) - \phi_1\gamma(0) - \cdots - \phi_p\gamma(p-1) = 0$

$$\vdots$$

for $j = p$, $\gamma(p) - \phi_1\gamma(p-1) - \cdots \phi_p\gamma(0) = 0$.
Rearranging the terms:

$$\sigma^2 = \gamma(0) - \phi_1\gamma(1) \cdots - \phi_p\gamma(p)$$

$$\phi_1\gamma(0) + \cdots + \phi_p\gamma(p-1) = \gamma(1)$$
$$\phi_1\gamma(1) + \cdots + \phi_p\gamma(p-2) = \gamma(2)$$
$$\vdots$$
$$\phi_1\gamma(p-1) + \cdots + \phi_p\gamma(0) = \gamma(p)$$

Above is Yule-Walker equations.

This system of $p+1$ equations are called Yule-Walker equations. System A can be written in matrix form as

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \cdots & \gamma(p-1) \\ & \gamma(0) & \gamma(1) & \cdots & \gamma(p-2) \\ & \cdots & \ddots & \ddots & \vdots \\ & \cdots & \cdots & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix}$$

Based on a sample $\{X_1, X_2, \cdots, X_n\}$ the parameters ϕ and σ^2 can be estimated by

$$\begin{cases} \hat{\phi} = \hat{\Gamma}_p^{-1}\hat{\gamma}_p \\ \hat{\sigma}^2 = \hat{\gamma}^2(0) - \hat{\phi}^T\hat{\gamma}_p \end{cases} \quad \text{where } \hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \end{pmatrix}, \hat{\gamma}_p = \begin{pmatrix} \hat{\gamma}(1) \\ \hat{\gamma}(2) \\ \vdots \\ \hat{\gamma}(p) \end{pmatrix}$$

The system is called the sample yule-walker equations. we can write Yule-Walker equations in terms of ACF too.

Yule-Walker equations can be written in terms of acf (rather than acvf). Dividing both sides of Yile-Wlake equations defined above and simplifying the resulting equations, we get

$$(\star)\begin{cases} \hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ \vdots \\ \hat{\phi}_p \end{pmatrix} = \hat{R}_p^{-1}\hat{\rho}_p \\ \hat{\sigma}^2 = \hat{\gamma}(0)[1 - \hat{\phi}^T\hat{\rho}_p] \end{cases}$$

where $\hat{\rho}_p = \begin{pmatrix} \hat{\rho}(1) \\ \vdots \\ \hat{\rho}(p) \end{pmatrix}$ and $\hat{R}_p = \frac{\hat{\Gamma}_p}{\gamma(0)} = [\hat{\rho}(i-j)]_{i,j=1}^p$

Notice that $\hat{\gamma}(0)$ is the sample variance of $\{x_1, \cdots, x_n\}$. Based on a sample $\{x_1, \cdots, x_n\}$, $(\star)$ will provide the parameter estimates.

Asside: in R, $a \leftarrow acf(x)$ Using advanced probability theory, it can be shown that

$$\tilde{\phi} = \begin{pmatrix} \tilde{\phi}_1 \\ \vdots \\ \tilde{\phi}_p \end{pmatrix} \sim MVN(\phi, \frac{\sigma^2}{n}\Gamma_p^{-1})$$

If we replace σ^2 and Γ_p by their sample estimates $\hat{\sigma}^2$ and $\hat{\Gamma}_p$, we can use this result for large-sample confidence intervals for the parameters $\phi_1, \cdots, \phi_p$, we show this in an example.

Example 24: Based on the following sample act and pact, an AR(2) model has been proposed for the data. Provide the Yule-Walker estimates of the parameters as well as

95% confidence interval for the parameters in $\phi = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}$. The data was collected over a window of 200 points with sample variance 3.96.

h	0	1	2	3	4	5	6	7
$\hat{\rho}(h)$	1	0.821	0.764	0.644	0.586	0.49	0.411	0.354
$\hat{\gamma}(h)$	1	0.821	0.277	-0.121	0.052	-0.06	-0.072	

$X_t = \phi_1 + X_{t-1} + \phi_2 X_{t-2} + Z_t, \{Z_t\} \sim WN(0, \sigma^2)$
We need to estimate ϕ_1 and ϕ_2.

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix} = \hat{R}_2^{-1} \hat{\rho}_2 = \begin{pmatrix} 0.594 \\ 0.276 \end{pmatrix}$$

We have

$$\hat{\sigma}^2 = \hat{\gamma}(0)[1 - \begin{pmatrix} \hat{\phi}_1 & \hat{\phi}_2 \end{pmatrix} \begin{pmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \end{pmatrix}] = 1.112$$

Therefore, the estimated model is

$$X_t = 0.594 X_{t-1} + 0.276 X_{t-2} + Z_t, \{Z_t\} \sim WN(0, 1.112)$$

$$\begin{pmatrix} \tilde{\phi}_1 \\ \tilde{\phi}_2 \end{pmatrix} \sim N(\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \frac{\sigma^2}{n} \Gamma_2^{-1})$$

$$\hat{\Gamma}_2 = \hat{\gamma}(0) \hat{R}_2 = 3.96 \begin{pmatrix} 1 & 0.821 \\ 0.821 & 1 \end{pmatrix}$$

$$\hat{\Gamma}_2^{-1} = \begin{pmatrix} 0.831 & -0.683 \\ -0.683 & 0.831 \end{pmatrix}$$

Therefore, $\frac{\hat{\sigma}^2}{n} \hat{\Gamma}_2^{-1} = \begin{pmatrix} 0.005 & -0.004 \\ -0.004 & 0.005 \end{pmatrix}$

Therefore, a 95% confidence interval for ϕ_1 is $\hat{\phi}_1 \pm 1.96 \sqrt{v\hat{a}r(\hat{\phi}_1)} \to (0.455, 0.733)$. a 95% confidence interval for ϕ_2 is $\hat{\phi}_2 \pm 1.96 \sqrt{v\hat{a}r(\hat{\phi}_2)} \to (0.137, 0.415)$.

4.2 Likelihood Methods

To use likelihood methods, we have to have some distributional assumptions. Consider $\{X_t, t \in T\}$ to be a Gaussian process. Therefore ,Z_t in $\phi(B) X_t = \theta(B) Z_t$ is i.i.d. $G(0, \sigma)$.

Based on the observations $x_1, x_2, \cdots, x_n$ at times $1, 2, \cdots, n$ the likelihood function of the parameters ϕ, θ and σ^2 is

$$L(\theta, \phi, \sigma^2) = \frac{1}{(2\pi)^{n/2}|\Gamma_n|^{1/2}} e^{-1/2 x^T \Gamma_n^{-1} x}$$

where $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ and

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

Notice that it is assumed that $E[X_t] = 0, \forall t$.

To estimate $\phi, \theta \& \sigma^2$, we maximize the likelihood function. Usually, it is easier to maximize the log of the likelihood function, which is called the log-likelihood. In this likelihood function, $\gamma(h)$ (hence Γ_h) depends on the parameters $\theta, \phi, \& \sigma^2$ in a non-linear way. Furthermore, as the dataset gets larger (n increases), the inversion Γ_n^{-1} can be computationally challenging. Therefore, efficient computational methods are needed for likelihood estimation.

4.3 Forecasting ARMA models

Based on the history of the process up to including time $n(x_1, x_2, \cdots, x_n)$, we are interested in deriving the predictor for $x_{n+h}, h >$ shown by $P(x_{n+h}|x_1, \cdots, x_n) = \hat{x}_{n+h}$ which minimizes the MSE. We know that $\hat{x}_{n+h}$ is of the form

$$\hat{x}_{n+h} = E[x_{n+h}|x_1, \cdots, x_n]$$

Therefore, in different cases of ARMA processes we will derive this conditional expectation. We will see that in the case of ARMA processes (linear). This expectation is in fact, the best linear predictor, $P_n X_{n+h}$.

4.4 Forecasting AR(p) process

Let $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + z_j, \{z_t\} \sim WN(0, \sigma^2)$, be a causal AR(p) process we have.

$$\begin{aligned} \hat{x}_{n+h} &= E[x_{n+h}|x_1, \cdots, x_n] \\ &= E[\sum_{j=1}^{p} \phi_j x_{n+h-j}|x_1, \cdots, x_n] + E[Z_{n+h}|x_1, \cdots, x_n] \\ &= E[\sum_{j=1}^{h-1} \phi_j x_{n+h-j} + \sum_{j=h}^{p} \phi_j x_{n+h-j}|x_1, \cdots, x_n] \end{aligned}$$

If $h = 1$, then above is just $\sum_{j=1}^{p} \phi_j x_{n+h-j}$ then $\sum_{j=1}^{p} \phi_j x_{n+1-j}$.

If $h = 2, 3, \cdots, p$, $j < h \implies n + h - j > n \implies$ first summation and $j \geq h \implies n + h - j \leq n \implies$ second summation.

Therefore,

$$\begin{aligned}
\sum_{j=h}^{p} \phi_j x_{n+h-j} + E[\sum_{j=1}^{h-1} \phi_j x_{n+h-j} | x_1, \cdots, x_n] \\
&= \sum_{j=h}^{p} \phi_j x_{x+h-j} - \sum_{j=1}^{h-1} \phi_j E[x_{n+h-j} | x_1, \cdots, x_n] \\
&= \sum_{j=1}^{h-1} \phi_j \hat{x}_{n+h-j} + \sum_{j=h}^{p} \phi_j x_{n+h-j}
\end{aligned}$$

If $h > p$: $n + h - j > n$. This implies

$$\begin{aligned}
E[\sum_{j=1}^{p} \phi_j x_{n+h-j} | x_1, \cdots, x_n] &= \sum_{j=1}^{p} \phi_j E[x_{n+h-j} | x_1, \cdots, x_n] \\
&= \sum_{j=1}^{p} \phi_j \hat{x}_{n+h-j}
\end{aligned}$$

In summary, for a causal AR(p) process, the h-step predictor is

$$\hat{x}_{n+h} = \begin{cases} \sum_{j=1}^{p} \phi_j x_{n+h-j} & h = 1 \\ \sum_{j=1}^{h-1} \phi_j \hat{x}_{n+h-j} + \sum_{j=h}^{p} \phi_j x_{n+h-j} & h = 2, 3, \cdots, p \\ \sum_{j=1}^{p} \phi_j \hat{x}_{n+h-j} & h > p \end{cases}$$

Note: in AR(p), the h-step prediction is a linear combination of the previous steps. We either have the previous p steps in $x_1, \cdots, x_n$ so we substitute the values (like the $h = 1$ case), or we don't have all or some of them, which we recursively predict.

Given a dataset, ϕ_j can be estimated ($\hat{\phi}_j$) and $\hat{x}_{n+h}$ will be computed.

Example 25: Based on the annual sales data of a chain store, an AR(2) model with parameters $\hat{\phi}_1 = 1$ and $\hat{\phi}_2 = -0.21$ has been fitted. If the total sales of the last 3 years have been 9, 11 and 10 million dollars. Forecast this year total sales (2013) as well as that of 2015.

$$X_t = \hat{\phi}_1 X_{t-1} + \hat{\phi}_2 X_{t-2} + Z_t$$

$$X_t = X_{t-1} - 0.21 X_{t-2} + Z_t, \{Z_t\} \sim WN(0, \sigma^2)$$

$$\hat{X}_{2013} = \hat{\phi}_1 X_{2012} + \hat{\phi}_2 X_{2011}$$
$$= X_{2012} - 0.21X_{2011} = 9 - 0.21 \times 11 = 6.69$$

Hence by doing the similar prediction, $\hat{X}_{2015} = 3.4$.

4.5 Forecasting in MA(q) process

MA processes are linear combination of white noise. To do forecasting in

$$MA(q) : X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

We need to estimate $\theta_1, \cdots, \theta_q$ as well as "approximate" the innovations $Z_t, Z_{t+1}, \cdots$. First consider the very simple case of MA(1):

$$X_t = Z_t + \theta Z_{t-1}, \{Z_t\} \sim WN(0, \sigma^2)$$

$$\hat{X}_{n+h} = E[X_{n+h}|X_1, \cdots, X_n] = E[Z_{n+h}|X_1, \cdots, X_n] + \theta E[Z_{n+h-1}|X_1, \cdots, X_n]$$

If $h = 1$, $= E[Z_{n+1}|X_1, \cdots, X_n] + \theta E[Z_n|X_1, \cdots, X_n] = E[Z_{n+1}] + \theta E[Z_n|X_1, \cdots, X_n] = \theta E[Z_n|X_1, \cdots, X_n] = \theta Z_n$

If $h > 1$, $= E[Z_{n+h}] + \theta E[Z_{n+h-1}] = 0$

Now, we need to plugin a value for Z_n. We "approximate" Z_i's by U_is as follows:

$$U_0, X_t = Z_t + \theta Z_{t-1} \implies Z_t = X_t - \theta Z_{t-1} \implies U_t = X_t - \theta U_{t-1}, U_0 = 0$$

$$U_0 = 0$$
$$U_1 = X_1 - \theta U_0 = X_1$$
$$U_2 = X_2 - \theta U_1 = X_2$$
$$\vdots$$

Notice that as $i \to \infty$, U will need a convergence condition where $|\theta| < 1$ is sufficient. This was the invertibility condition for MA(1).

We see that U_t's are recursively calculable. This implies for an invertible MA(1) process we have

$$\hat{X}_{n+h} = \begin{cases} \theta U_t & h = 1 \\ 0 & h > 0 \end{cases}$$

Where $U_t = X_t - \theta U_t, U_0 = 0$.

Now consider MA(q) process $X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$. $\hat{X}_{n+h} = E[X_{n+h}|X_1, \cdots, X_n] = E[Z_{n+h}|X_1, \cdots, X_n] + \theta_1 E[Z_{n+h-1}|X_1, \cdots, X_n] + \cdots + \theta_q E[Z_{n+h-q}|X_1, \cdots, X_n]$

Clearly, if $h > q \to n + h - q > n$, hence $\hat{X}_{n+h} = 0$. If $0 < h \leq q$, then at least some of the terms are non-zero. Then $= 0 + \sum_{j=1}^{q} \theta_j E[Z_{n+h-j}|X_1, \cdots, X_n] = 0 + \sum_{j=h}^{q} \theta_j E[Z_{n+h-j}|X_1, \cdots, X_n] = \sum_{j=h}^{q} \theta_j Z_{n+h-j}$ for $j = h, h+1, h+2, \cdots, q$,

$$E[Z_{n+h-j}|X_1, \cdots, X_n] = Z_{n+h-j}$$

Similar to MA(1), we approximate Z_i's by U_i's provided the MA(q) process is invertible, i.e., $\theta(z) = 1+\theta_1 z+\cdots+\theta_q z^q \neq 0, \forall z : |z| \leq 1$. Therefore, assuming $U_0 = U_{-1} = U_{-2} = \cdots = 0$, then $U_n = X_n - \sum_{j=1}^{q} \theta_j U_{n-j}$.

Therefore, $U_0 = 0, U_1 = X_1 - \sum_{j=1}^{q} \theta_j U_{1-j} = X_1$, $U_2 = X_2 - \theta_1 U_1 = X_2 - \theta_1 X_1, \cdots$.

in summary, for an invertible MA(q) process we have:

$$\hat{X}_{n+h} = \begin{cases} \sum_{j=h}^{q} \theta_j U_{n+h-j} & 1 \leq h \leq q \\ 0 & h > q \end{cases}$$

where, $U_0 = U_i = 0$ and $U_n = X_n - \sum_{j=1}^{q} \theta_j U_{n-j}, n = 1, 2, 3, \cdots$