STAT 443 (Winter 2014 - 1141)

Forecasting

Prof. R. Ramezan
University of Waterloo

LaTeXer: W. KONG
http://wwkong.github.io
Last Revision: September 3, 2014

Table of Contents

1 Introduction **1**

2 Time Series Models **1**
2.1 Zero Mean Models . . . 1
2.2 Models with Trend . . . 1
2.3 Models with Seasonal Component . . . 2

3 Stationary Models **2**
3.1 Moving Averages . . . 3
3.2 The Sample Autocorrelation Function . . . 4
3.3 Linear Regression . . . 4

4 Prediction **4**
4.1 Model Selection . . . 5
4.2 Interpolation vs. Extrapolation . . . 6
4.3 Tests on Residuals . . . 6

5 Smoothing Methods **6**
5.1 Trend Estimation . . . 7
5.2 Estimating Seasonality . . . 7
5.3 Modeling Residuals . . . 8

6 Stationary and Linear Processes **9**
6.1 Linear Prediction . . . 11
6.2 Linear Processes . . . 13
6.3 Box-Jenkins Models . . . 13
6.4 Invertibility . . . 14
6.5 Partial Autocorrelation Function (PACF) . . . 16
6.6 $ARIMA(p,d,q)$ Processes . . . 18
6.7 $SARIMA(p,d,q)\times(P,D,Q)_S$ Processes . . . 18

7 Box-Jenkins Methodology **19**

8 Parameter Estimation in ARMA Processes **19**
8.1 Yule-Walker Methods . . . 20

9 Likelihood Models **21**

10 Forecasting **21**
10.1 Forecasting AR(p) . . . 22
10.2 Forecasting MA(q) . . . 23
10.3 Forecasting ARMA(p,q) Processes . . . 25

These notes are currently a work in progress, and as such may be incomplete or contain errors.

ACKNOWLEDGMENTS:

Special thanks to *Michael Baker* and his LaTeX formatted notes. They were the inspiration for the structure of these notes.

Abstract

The purpose of these notes is to provide the reader with a secondary reference to the material covered in STAT 443. The formal prerequisite to this course is STAT 331 but this author believes that the overlap between the two courses is less than 10%. Readers should have a good background in linear algebra, basic statistics, and calculus before enrolling in this course.

1 Introduction

A **time series** is a sub-class of stochastic processes which are indexed by time and can be represented by $\{X_t : t \in T\}$. 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 course, 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 (continuous) time. The main focus of this course is to develop models for discrete time series.

Example 1.1. For example, when we say $x_5 = 10$, we mean the value of x at time 5 is equal to 10.

2 Time Series Models

Our interest lies in modeling and the analysis of data collected over time (time series). Ideally, given a discrete stochastic process $\{X_t\}$, we want the joint distribution of $X_1, X_2, ..., X_n$ for all $n \in \mathbb{N}$. In real world applications, this is generally not possible because we don't have enough information to fully specify the joint distribution. The good news is that in most information about the joint distribution is 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 $E(X_t X_{t*})$ for any $t, t^* \in \mathbb{N}$ summarizes most of the information content about the process.

If the joint distribution is multivariate normal, then the three expectations above fully specify the joint distribution. Recall that the multivariate normal distribution is written as $N_p(\mu, \Sigma)$ where p is the dimension, μ is mean vector, and Σ is the variance-covariance matrix which is $p \times p$. It is easy to see that μ and Σ are parametrized by the three expectations above, which we now call (*).

Since (*) contains a fair amount of information, instead of working with the joint, we will work with time series models which employ (*).

Definition 2.1. A **time series model** for observed data $\{x_t\}$ is a specification of the joint distributions (or possibly only (*)) of a sequence of random variables $\{X_t\}$ of which $\{x_t\}$ is postulated to be realization.

2.1 Zero Mean Models

(1) iid noise: If $\{X_1, ..., X_k\}$ are iid random variables, then

$$P(X_1 \leq x_1, ..., X_k \leq x_k) = \prod_{n=1}^{k} P(X_n \leq x_n) = \prod_{n=1}^{k} P(X_1 \leq x_n)$$

and the joint is defined by one marginal distribution with zero mean. Observe that using the independence assumption we see that

$$P(X_{n+h} \leq x | X_1 \leq x_1, ..., X_n \leq x_n) = P(X_{n+h} \leq x)$$

(2) Random walk: $\{S_t, t \in \mathbb{N}\}$, starting at $S_0 = 0$ is a random walk if $S_t = \sum_{k=1}^{t} X_k$ where X_k are white noise random variables.

(3) White noise (zero-mean): A white noise process is a sequence of uncorrelated random variables $\{X_t\}$ each with constant mean 0 and constant variance σ^2. We denote this by $\{X_t\} \sim WN(0, \sigma^2)$.

2.2 Models with Trend

Consider the model $X_t = m_t + Y_t$ where m_t is a slowly changing function, called the **trend**, and Y_t has zero mean. We have $E(X_t) = m_t \; \forall t$. Notice that m_t is a *non-random* function of time t. This trend component can be linear, quadratic or any kind of arbitrary function.

Example 2.1. Consider $X_t = m_t + Y_t$ where $m_t = 2 + t$ and $Y_t \sim N(0, 1)$. This is a linear trend with a perturbation being modeled as a standardized normal random variable.

2.3 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 (the **seasonal component**) with period d ($S_t = S_{t+d}\ \forall t$). In a sense, s_t is a particular kind of trend. An example would be the seasonal component $s_t = \alpha_0 + \alpha_1 \cos(\alpha_2 t)$. We can also use indicator functions (on certain months) for the seasonality component as well.

In both models $X_t = m_t + Y_t$ and $X_t = s_t + Y_t$ the parameters $\alpha_0, \alpha_1, \alpha_2, ...$ are usually estimated by maximum likelihood or least squares methods. Notice that in the general case, we can use the model $X_t = m_t + s_t + Y_t$ which contains both the seasonal and trend component. This is called the **classical decomposition** and will be frequently referred to in this course. We an use regression models to estimate m_t and s_t.

Example 2.2. Consider the average seasonal temperature over many years where $Z_1 =$ Spring of 2004, $Z_2 =$ Summer of 2004, ..., Z_{20}. Suppose that we want to fit a model of the form $X_t = m_t + s_t + Y_t$ where m_t is polynomial in t. We use the dummy coding contrasts matrix in the form $\begin{bmatrix} I & 0 \end{bmatrix}^T$ in the order of Spring, Summer, Fall, and Winter. Suppose that X_1, X_2, X_3, X_4 represent the categories of the seasons. Then, we use the general model

$$Z_t = \underbrace{Y_t}_{\text{Error}} + \underbrace{\sum_{i=0}^{p} \beta_i t^i}_{m_t} + \underbrace{\sum_{j=1}^{3} \alpha_j X_j}_{s_t}$$

We should use the rule that says that if a periodic trend with period d is being modeled through regression analysis, $d-1$ binary variates should be introduced to the model.

Now consider Example 2 (generated from R code on the UW Learn page). In this example, we fitted the model

$$\ln(Y_t) = \sum_{i=1}^{3} \beta_i t^i + \sum_{j=3}^{11} \beta_j x_{j-2} + R_t$$

with R_t being the random component which is i.i.d. $N(0, \sigma^2)$. Although the model is good in terms of fit, it does not satisfy the fundamental assumption of independent residuals. 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.

3 Stationary Models

Definition 3.1. The time series $\{X_t : t \in T\}$ is called **strictly (strong) stationary** if the joint distribution of $X_{t_1}, X_{t_2}, ..., X_{t_n}$ is the same as that of $X_{t_1-k}, X_{t_2-k}, ..., X_{t_n-k}$ for all $n, t_1, ..., t_n, k \in \mathbb{N}$. In other words, $\{X_t\}$ is strictly stationary if *all* of its statistical properties remain the same under time shifts.

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 in the moments $E[X_t], E[X_t^2]$ and $E[X_t X_{t^*}]$ for all $t, t*$. This motivates introducing a a type of stationarity based on these lower order moments, which we will call **weak stationarity**. To introduce weak stationarity, we need some more definitions first.

Definition 3.2. Let $\{X_t\}$ be a time series with $E[X_t^2] < \infty$. The **mean function** of $\{X_t\}$ is $\mu_X(t) = \mu_t = E[X_t]$ and the **covariance function** of $\{X_t\}$ is

$$\gamma_X(r, s) = Cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

Definition 3.3. 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 independent of t
2. $\gamma_X(t, t+h) = Cov(X_t, X_{t+h})$ is independent of t for all h; the covariance only depends on the distance h instead of t

3. $E[X_t^2] < \infty$ is also one of the conditions for weak stationarity.

Also, in view of the latter condition above, we use the term "covariance function" with reference to a stationary time series $\{X_t\}$ we shall mean the function of one variable defined by

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

Exercise 3.1. If $E(X_t^2) < \infty$, show that strict stationarity implies weak stationarity.

Definition 3.4. Let $\{X_t\}$ be a stationary time series. The **autocovariance function** (ACVF) of $\{X_t\}$ at lag h is $\gamma_X(h) = Cov(X_{t+h}, X_t)$. The **autocorrelation function** (ACF) of $\{X_t\}$ at lag h is $\gamma_X(h) = \gamma_X(h)/\gamma_X(0) = Corr(X_{t+h}, X_t)$.

Example 3.1. Investigate the stationarity of white noise. Let $\{X_t\}$ be white noise with $\{X_t\} \sim WN(0, \sigma^2)$. Automatically, $\sigma^2 < \infty \implies Var(X_t) < \infty$, $E[X_t] = 0$ does not depend on t, and $Cov(X_t, X_{t+h}) = \sigma\delta_{th}$ where δ_{th} is the Dirac delta function. Hence, white noise is weakly stationary.

Example 3.2. A random walk $\{S_t\}$ is not stationary because $Var(S_t) = t\sigma^2$.

Notation 1. Whenever we refer to a stationary time series from now on (since Jan. 16, 2014), we mean weakly stationary unless otherwise specified.

3.1 Moving Averages

Example 3.3. Consider the process $X_t = Z_t + \theta Z_{t-1}$ where $t \in \mathbb{Z}$ and $Z_t \sim WN(0, \sigma^2)$. This process is called the **first-order moving average** [$MA(1)$]. Show that $\{X_t\}$ is stationary.

It can be shown that

$$\begin{aligned} Var(X_t) &= \sigma^2(1+\theta^2) < \infty \\ E(X_t) &= 0 \\ Cov(X_t, X_{t+h}) &= \begin{cases} \sigma^2(1+\theta^2) & h = 0 \\ \theta\sigma^2 & |h| = 1 \\ 0 & |h| > 1 \end{cases} \end{aligned}$$

and hence because all functions are independent of t and the variance is finite, then X_t is stationary. We can also derive the ACF as

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

Note that this illustrates that $\gamma(h)$ is an **even function**.

Example 3.4. Let $\{X_t\}$ be a stationary time series satisfying the equations $X_t = \phi X_{t-1} + Z_t$ for $t \in \mathbb{Z}$ where $|\phi| < 1$ and $\{Z_t\} \sim WN(0, \sigma^2)$. Also let Z_t and X_s be uncorrelated for each $s < t$. Then time series $\{X_t\}$ is called an **autoregressive process** of order 1 [$AR(1)$].

Note that because $\{X_t\}$ is stationary, we have $E[X_t] = \mu = \phi\mu \implies \mu = 0$ for any t. We also have that $\gamma(0) = \phi^2\gamma(0) + \sigma^2 \implies \gamma(0) = \frac{\sigma^2}{1-\phi^2}$. Now if $h > 0$, multiply both sides of the expression for X_t by X_{t-h} and take expectations to get

$$\begin{aligned} E[X_t X_{t-h}] = \phi E[X_{t-h}X_{t-1}] + E[X_{t-1}Z_t] \quad &\implies \quad \gamma(h) = \phi\gamma(h-1) \\ &\implies \quad \gamma(k) = \phi^k\gamma(0) = \frac{\phi^k\sigma^2}{1-\phi^2} \end{aligned}$$

You can repeat the same trick for $h < 0$ to get $\gamma(k) = \frac{\phi^k\sigma^2}{1-\phi^2}$ and hence the ACF is $\rho(h) = \phi^{|h|}$.

3.2 The Sample Autocorrelation Function

What we have seen so far on ACF is based on given models (theoretical). In practice, based on the observed data $\{x_1, x_2, ..., x_n\}$ we use the **sample ACF** to assess the degree of dependence in data. Sample ACF is the estimate of the theoretical ACF (under stationarity).

Definition 3.5. Let $x_1, ..., x_n$ be observations of a time series. The **sample mean** of $x_1, ..., x_n$ is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The **sample autocovariance function** is

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

The **sample autocorrelation function** is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, h \in (-n, n)$$

The convention that will use is the hat notation for estimates and the tilde notation (e.g. $\tilde{\gamma}(h)$) for estimators. Variables without a hat or a tilde are theoretical fixed values. In summary,

$$\begin{aligned}
\gamma(h) &\mapsto \text{Theoretical; fixed but unknown} \\
\tilde{\gamma}(h) &\mapsto \text{The estimator; random variable} \\
\hat{\gamma}(h) &\mapsto \text{Realization of } \tilde{\gamma}(h) \text{based on a sample}
\end{aligned}$$

The sample ACF measures the correlation in the data (under stationarity). Therefore, it can be used to the "uncorrelatedness" of the residuals of a regression model. Note that if we have Gaussian residuals, Independent $\iff$ Not Correlated. It can also be show that for i.i.d. noise with finite variance,

$$\tilde{\rho}(h) \sim N\left(0, \frac{1}{n}\right)$$

where n is the sample size for large values of n. Therefore, for data from such processes (i.i.d. noise) we expect than 95% of the sample ACFs fall between $\pm 1.96/\sqrt{n}$. That is,

$$P\left(\frac{-1.96}{\sqrt{n}} < \tilde{\rho}(h) < \frac{1.96}{\sqrt{n}}\right) = 0.95$$

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

Remark 3.1. For the observed data $\{x_1, ..., x_n\}$:

- 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 behaviour with the "same period"

3.3 Linear Regression

This was just a review of STAT 331 and STAT 371 done in two lectures. Nothing to see here. Carry on.

4 Prediction

Suppose that we have a model

$$Y_i = \alpha' + \beta x_i + R_i, R_i \sim N(0, \sigma^2)\ i.i.d.$$

We want to predict Y_{new} for a new value for $x = x_{new}$. If $\alpha = \alpha' + \beta\bar{x}$, we can rewrite our model as

$$Y_i = \alpha + \beta(x_i - \bar{x}) + R_i, R_i \sim N(0, \sigma^2)\ i.i.d.$$

It can be shown that

$$\begin{cases} \tilde{\alpha} \sim G\left(\alpha, \frac{\sigma}{\sqrt{n}}\right) \\ \tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{XX}}}\right) \end{cases}$$

which are independent, and the estimator $\tilde{\mu}(x_{new}) = E[Y|X = x_{new}]$ is

$$\tilde{\mu}(x_{new}) = \tilde{\alpha} + \tilde{\beta}(x_{new} - \bar{x}) \sim N\left(\alpha + \beta(x_{new} - \bar{x}), \sigma^2\left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{XX}}\right)\right)$$

where

$$\begin{aligned} E[\tilde{\mu}(x_{new})] &= \alpha + \beta(x_{new} - \bar{x}) + \underbrace{E[R_{new}]}_{=0} = \alpha + \beta(x_{new} - \bar{x}) \\ Var[\tilde{\mu}(x_{new})] &= Var(\tilde{\alpha}) + (x_{new} - \bar{x})^2 Var(\tilde{\beta}) = \sigma^2\left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{XX}}\right) \end{aligned}$$

Since $Y_{new} \sim N\left(\alpha + \beta(x_{new} - \bar{x}), \sigma^2\right)$ then $Y_{new} - \tilde{\mu}_{new} \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{XX}}\right)\right)$ and

$$\frac{Y_{new} - \tilde{\mu}_{new}}{\tilde{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{XX}}}} \sim t_{n-2}$$

If $c = F^{-1}_{t_{n-2}}(0.975)$, then the 95% prediction interval is

$$\hat{\mu}(x_{new}) \pm c\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{XX}}}$$

and the prediction interval for multiple linear regression is

$$\hat{\mu}(x_{new}) \pm c\hat{\sigma}\sqrt{1 + x_{new}^T (X^T X)^{-1} x_{new}}$$

where c is from a t_{n-p-1} distribution.

Note 1. In prediction, we must always consider the **bias-variance tradeoff**. That is, as the model becomes more flexible (less bias and more variance), the less prediction power it has because it does not understand the pattern as well.

4.1 Model Selection

We just give the formulas for the various information criterions here:

$$\begin{aligned} AIC &= -2l(\hat{\theta}) + 2N_p \\ AICc &= AIC + \frac{2N_p(N_p + 1)}{n - N_p - 1} \\ BIC &= -2l(\hat{\theta}) + N_p \log(n) \end{aligned}$$

where $\hat{l}(\theta)$ is the log-likelihood function. Also, here is the formula for the PRESS statistic

$$\text{PRESS} = \sum_{y \in \text{validation set}} (y - \hat{y})^2$$

Here are some strategies involving these statistics:

- Use a stepwise strategy
- Build up by adding one variable at a time
- Build down by subtracting one variable at a time

- Use a mixed strategy

4.2 Interpolation vs. Extrapolation

If we let $h_{max} = \max(H_{ij})$ where $H = X(X^TX)^{-1}X^T$ then if the point x satisfies $x^T(X^TX)^{-1}x \leq h_{\max}$, then estimating y for x is an interpolation problem, otherwise extrapolation. (cf. Montgomery, E.A Peck)

4.3 Tests on Residuals

The **Shapiro-Wilk Test** is as follows:

- $H_0 : Y_1, ..., Y_n$ come from a Gaussian distribution
- Reject H_0 if the p-value of this test is small
- In R, if the data is stored in the vector y, then use the command *shapiro.test(y)*.

The **Difference Sign Test** is as follows:

- Count the number S of values such that $y_i - y_{i-1} > 0$
- For large i.i.d. sequences
$$\mu_S = E[S] = \frac{n-1}{2}, \sigma_S^2 = \frac{n+1}{12}$$
- For large n, S is approximately $N(\mu_S, \sigma_S^2)$, therefore,
$$W = \frac{S - \mu_S}{\sqrt{\sigma_S^2}} \sim N(0,1)$$
- A large positive value of $S - \mu_S$ indicates the presence of increasing (decreasing) trend
- We reject (H_0 : data is random) if $|W| > z_{1-\alpha/2}$ but this may not work for seasonal data

The **Runs Test** is as follows:

- Estimate the median and call it m
- Let n_1 be the number of observations $> m$ and n_2 be the number of observations $< m$
- Let R be the number of consecutive observations which are all smaller (larger) than m
- For large i.i.d. sequences
$$\mu_R = E[R] = 1 + \frac{2n_1n_2}{n_1+n_2}, \sigma_R^2 = \frac{(\mu_R - 1)(\mu_R - 2)}{n_1 + n_2 - 1}$$
- For large number of observations,
$$\frac{R - \mu_R}{\sigma_R} \sim N(0,1)$$

5 Smoothing Methods

Recall the **classical decomposition**
$$X_t = m_t + s_t + Y_t$$
with period d, noise Y_t and trend m_t. For identification, we need $\sum_{t=1}^{d} s_t = 0$ and $E[Y_t] = 0$. Here, the assumption of linearity is strong, in the sense that it may or may not hold. Out goal is to estimate and extract m_t and s_{t^-} and hope that the random component Y_t will turn out to be stationary time series.

5.1 Trend Estimation

Consider a non-seasonal model with a trend and stochastic component. If $E[Y_t] \neq 0$ then we define $m_t' = m_t + E[Y_t]$ and $Y_t' = Y_t - E[Y_t]$ to create new variables which follow our classical assumptions. There are many ways to estimate trend in this model which we list below:

1. (Finite Moving Average Filter) Let q be a non-negative integer and consider the two-sided moving average of the series X_t. We have
$$m_t \approx \frac{1}{2q+1}\sum_{j=-q}^{q} X_{t-j} = \frac{1}{2q+1}\sum_{j=-q}^{q} m_{t-j} + \underbrace{\frac{1}{2q+1}\sum_{j=-q}^{q} Y_{t-j}}_{\approx 0}$$

2. (Exponential Smoothing) For fixed $\alpha \in [0,1]$ define the recursion
$$\hat{m}_t = \alpha X_t + (1-\alpha)\hat{m}_{t-1}$$
with initial condition $\hat{m}_1 = X_1$. This gives an exponentially decreasing weighted moving average where in the general $t \geq 2$ case,
$$\hat{m}_t = \sum_{j=0}^{t-2} \alpha(1-\alpha)^j X_{t-j} + (1-\alpha)^{t-1} X_1$$
Note that a smaller α creates a smoother plot compared to a larger α.

3. (Polynomial Regression) This is just developing a parametric polynomial form of m_t in the form
$$m_t = \sum_{i=0}^{k} \beta_i t^i$$
where k is chosen arbitrarily.

4. We can also eliminate the trend through **differencing** where
$$\nabla X_t = X_t - X_{t-1} = (1-B)X_t$$
and ∇, B are known to be the differencing and backshift operators respectively. Exponentiating these operators is equivalent to function composition. In this case, we are applying differencing to get a stationary process (by eliminating the trend).

Example 5.1. Consider $X_t = \alpha + \beta t + Y_t$. Since $E[X_t]$ depends on time, this series is non-stationary. Under differencing,
$$\nabla X_t = \beta + \underbrace{Y_t - Y_{t-1}}_{=Y_t^*}$$
where Y_t^* is stationary. Similarly, if $X_t' = \alpha + \beta t + \gamma t^2 + Y_t$, then
$$\nabla X_t' = \beta + 2\gamma t - \gamma + Y_t - Y_{t-1}$$
but it can be shown that $\nabla^2 X_t'$ IS stationary.

5.2 Estimating Seasonality

Suppose that $\hat{m}_t$ is a moving average filter for the trend of the data. For each $k = 1, ..., d$, estimate w_k as
$$w_k = \frac{\sum_{q<k+jd\leq n-q}(x_{k+jd} - \hat{m}_{k+jd})}{|\{x_{k+jd} - \hat{m}_{k+jd} | q < k+jd \leq n-q\}|}$$

which is the average of $\{x_{k+jd} - \hat{m}_{k+jd} | q < k + jd \leq n - q\}$. Normalize to get

$$\hat{s}_k = w_k - \frac{\sum_{i=1}^{d} w_i}{d}$$

so that $\sum_{j=1}^{d} s_j = 0$. Note that $\hat{s}_k = \hat{s}_{k-d}$ for $k > d$.

5.3 Modeling Residuals

Holt-Winters

- This generalizes exponential smoothing to the case where there is a trend and seasonality
- Following Chatfield and Yar (1988), we define trend as long-term change in the mean level per unit time
- 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 across time.
- L_t is the **level** and T_t is the **slope** of the trend at time t
- Holt's idea is that $\hat{Y}_{t+h} | Y_1, ..., Y_t = L_t + h \times T_t$
- There are two forms of seasonality to be added to Holt's model: additive and multiplicative

Holt-Winters (Additive Case)

- Define Level, Trend, and Seasonal Index at time t by L_t, T_t, I_t where seasonal effect is of period p
- The update rules are

  $$\begin{aligned} L_t &= \alpha(X_t - I_{t-p}) + (1-\alpha)(L_t + T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ I_t &= \gamma(X_t - L_t) + (1-\gamma)I_{t-p} \end{aligned}$$

- The forecast for h periods is $L_t + hT_t + I_{t-p+h}$

Holt-Winters (Multiplicative Case)

- Define Level, Trend, and Seasonal Index at time t by L_t, T_t, I_t where seasonal effect is of period p
- The update rules are

  $$\begin{aligned} L_t &= \alpha(X_t / I_{t-p}) + (1-\alpha)(L_t + T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ I_t &= \gamma(X_t / L_t) + (1-\gamma)I_{t-p} \end{aligned}$$

- The forecast for h periods is $(L_t + hT_t)\, I_{t-p+h}$

Holt-Winters (Algorithm)

- This is a recursive algorithm so you will need to provide initial values for L_t, T_t, I_t at the beginning of the series
- Values will need to be provided for α, β, γ (R minimizes the squared one-step prediction error to estimate these parameters).

- You then will need to choose between additive and multiplicative models (In R, this is the only step that you execute)

Holt-Winters (Special Cases)

- In the case that $\beta = \gamma = 0$ we have no trend or seasonal updates in the H-W algorithm
- Here, we have $L_t = \alpha X_t + (1-\alpha)L_{t-1}$ which is **exactly (simple) exponential smoothing under** α
- In the case that $\gamma = 0$ we have no seasonal component and there are two H-W equations for updating L_t and T_t
- We call the above case **double exponential smoothing**

Example 5.2. (Inventory Prediction) This example has a trend but no seasonality. Use the H-W algorithm with $\gamma = 0$ (double exponential smoothing). In the case of simple exponential smoothing,

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

where the forecaster is $\hat{Y}_{t+1} = m_t$. This is why the exponential smoother looks like it is one step "behind". We can rewrite this equation as

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

which is a trend plus a weighted forecasting error. This case is $\beta = \gamma = 0$ in the H-W algorithm.

Remark 5.1. Double exponential smoothing captures more of the underlying trend than simple exponential smoothing.

[CHECK OUT THE LECTURE SLIDES FOR MORE EXAMPLES (for the midterm) !]

6 Stationary and Linear Processes

To perform any form of forecasting, there must be an assumption that some things are the same in the future as in the past. The idea of *being constant over time* is central to stationary processes. Therefore, we'll use stationary processes as the main framework to develop forecasting models.

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

Definition 6.1. A process $\{X_t\}$ is called a **moving average process of order** q if

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

where $\{Z_t\} \sim WN(0, \sigma^2)$ and $\theta_1, ..., \theta_q$ are constants. Sometimes Z_t is referred to as the **innovation**. Notice that these innovations are uncorrelated, have constant variance and zero mean. Deriving the mean and autocovariance function of $MA(q)$, it is easy to see that this process is stationary.

Definition 6.2. We say that a process $\{X_t\}$ is $q-$**dependent** if X_t and are X_s are independent if $|t-s| > q$. That is, they are dependent if they are within q steps of each other. Similarly, we saay that that stationary time series is $q-$**correlated** if $\gamma(h) = 0$ whenever $|h| > q$.

Example 6.1. It is easy to show that the $MA(q)$ process is $q-$correlated. The inverse of this statement is also true.

Proposition 6.1. *If $\{X_t\}$ is a stationary $q-$correlated time series with mean 0, then it can be represented as the $MA(q)$ process. (ON MIDTERM?)*

Definition 6.3. Consider the process $\{X_t\}$ denoted by $X_t = \phi X_{t-1} + Z_t$ for $t = 0, 1, 2, ...$ where $\{Z_t\} \sim WN(0, \sigma)$. This process is called the **first-order autoregressive process** or $AR(1)$. We can show this process by $(1 - \phi B)X_t = Z_t$. Notice that if $|\phi| = 1$, then $\{X_t\}$ forms a random walk that is not stationary. Therefore, depending on the value of ϕ, $\{X_t\}$ may or may not be stationary.

Remark 6.1. Consider the $AR(1)$ process with the condition $|\phi| \leq 1$. We have, by induction,

$$X_t = Z_t + \sum_{j=1}^{\infty} \phi^j Z_{t-j} = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$$

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

Definition 6.4. process $\{X_t\}$ is called a **autoregressive process of order** p if

$$X_t = X_t + \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + Z_t$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ and $\phi_1, ..., \phi_p$ are constants.

Definition 6.5. $\{X_t\}$ is called a **Gaussian time series** if all its joint distributions are multivariate normal. That is for any set $i_1, ..., i_m$ with each $n \in \mathbb{N}$, the random vector $(X_{i_1}, ..., X_{i_m})$ follows a multivariate normal distribution.

Note 2. If $(X_1, X_2) \sim MVN\left(\begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}^T, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right)$ then

$$\begin{aligned} X_1|X_2 = x_2 &\sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}\right) \\ &= N\left(\mu_1 + \frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2)\right), \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} \end{aligned}$$

Example 6.2. Consider the stationary Gaussian time series $\{X_t\}$. Suppose X_n has been observed and we want to forecast X_{t+h} using $m(X_n)$, a function of X_n. Let us measure the quality of the forecast by

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

It can be shown that $m(\cdot)$ which minimizes **MSE** in a general case is $m(X_n) = E(X_{n+h}|X_n)$. In this case, stationarity implies that $E[X_{n+h}] = E[X_n] = \mu$ and

$$Cov(X_{n+h}, X_{n+h}) = \gamma(0) = Var(X_n) = \sigma^2$$

and

$$Cor(X_{n+h}, X_n) = \frac{\gamma(h)}{\gamma(0)}$$

If this process was a Gaussian time series, then

$$(X_1, X_2) \sim MVN\left(\begin{bmatrix} \mu & \mu \end{bmatrix}^T, \begin{bmatrix} \sigma^2 & \sigma^2\rho(h) \\ \sigma^2\rho(h) & \sigma^2 \end{bmatrix}\right)$$

$$X_1|X_2 = x_2 \sim N\left(\mu + \rho(h)(x_2 - \mu_2), \sigma^2(1-\rho(h)^2)\right)$$

$$\underbrace{m(X_n) = E[X_{n+h}|X_n]}_{(1)} = \underbrace{\mu + \rho(h)(X_n - \mu)}_{(2)}$$

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

and even if the normality does not hold, we can still look at the predictor $m(X_n) = aX_n + b$ where a and b where there are derived from

$$\min_{a,b} E\left(\left[X_{n+h} - \underbrace{(aX_n + b)}_{m(X_n)}\right]^2\right)$$

Equation (1) will be shown later. Equation (2) is the best form of $m(x)$ such that MSE is minimized under the Gaussian

process assumption. Note that there is no conditional above because[1]

$$\begin{aligned} E\left[(X_{n+h}-m(X_n))^2\right] &= E\left\{E\left[(X_{n+h}-m(X_n))^2|X_n\right]\right\} \\ &\geq E\left\{E\left[(X_{n+h}-m^*(X_n))^2|X_n\right]\right\}, m^*(X_n)=E\left(X_{n+h}|X_n\right) \\ &= E\left[(X_{n+h}-m^*(X_n))^2\right] \end{aligned}$$

6.1 Linear Prediction

We now consider the problem of predicting $X_{n+h}, h>0$ for a stationary time series with known mean μ and ACVF $\gamma(\cdot)$ based on previous values $\{X_n,...,X_1\}$ showing the linear predictor of X_{n+h} by P_nX_{n+h}. We are interested in

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

which minimizes

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

To get $a_0,a_1,...,a_n$ we need to solve the system $\frac{\partial S}{\partial a_j}=0$ for $j=0,1,...,n$. Doing so, we get

$$a_0=\mu\left(1-\sum_{i=1}^{n}a_i\right), \Gamma_n a_n=\gamma_n(h)$$

where

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

which implies

$$\begin{aligned} P_nX_{n+h} &= a_0+\sum_{i=1}^{n}a_iX_{n-i+1} \\ &= \mu\left(1-\sum_{i=1}^{n}a_i\right)+\sum_{i=1}^{n}a_iX_{n-i+1} \\ &= \mu\sum_{i=1}^{n}a_i\left(X_{n-i+1}-\mu\right) \end{aligned}$$

Note 3. Here are some properties from the above:

- P_nX_{n+h} is defined by $\mu,\gamma(h)$
- It can be shown that $E\left[(X_{n+h}-P_nX_{n+h})^2\right]=\gamma(0)-a_n^T\gamma_n(h)$
- $E(X_{n+h}-P_nX_{n+h})=0$
- $E\left[(X_{n+h}-P_nX_{n+h})X_j\right]=0$ for $j=1,2,...,n$

In a more general set-up, suppose that Y and $W_1,...,W_n$ are any random variables with finite second moments and means $\mu_Y=E(Y),\mu_i=E(W_i)$ and $Cov(Y,Y),Cov(Y,W_i),Cov(W_i,W_j)$ are all known for $i=1,...,n$.

Define $\tilde{W}=(W_n,...,W_1)$ and $\mu_W=(\mu_n,...,\mu_1)^T$. Then

$$\begin{aligned} \gamma &= Cov(Y,\tilde{W})=(Cov(Y,W_n),...,Cov(Y,W_1))^T \\ \Gamma &= Cov(\tilde{W},\tilde{W})=\left[Cov(W_{n+1-i},W_{n+1-i})\right]_{i,j=1}^{n}\in\mathbb{R}^{n\times n} \end{aligned}$$

[1] Midterm content ends here.

Now by the same argument used in the derivation of P_nX_{n+h}, the **"best" linear predictor** of Y in terms of $\{W_n, ..., W_1\}$ is

$$P_{\tilde{W}}Y = P(Y|\tilde{W}) = \mu_Y + a_n^T(\tilde{W} - \mu_W)$$

where a_n is the solution of $\Gamma a = \gamma$. Also the MSE of this predictor is

$$E\left[(Y - P_{\tilde{W}}Y)^2\right] = Var(Y) - a_n^T\gamma$$

Note 4. In this case, we have the following properties

- Suppose that $E[U^2] < \infty, E[V^2] < \infty, \Gamma = Cov(\tilde{W}, \tilde{W})$ and $\beta, \alpha_1, ..., \alpha_n$ are constants. Then the following are true:
  1. $P_{\tilde{W}}U = E[U] + \tilde{a}_n^T(\tilde{W} - \mu_W)$ where $\Gamma\tilde{a}_n = \gamma$
  2. $E\left[(U - P_{\tilde{W}}U)\,\tilde{W}\right] = 0$ and $E\left[U - P_{\tilde{W}}U\right] = 0$
  3. $E\left[(U - P_{\tilde{W}}U)^2\right] = Var(U) - \tilde{a}_n^T Cov(U, \tilde{W})$
  4. $P_{\tilde{W}}\left[\alpha_1 U + \alpha_2 V + \beta\right] = \alpha_1 P_{\tilde{W}}U + \alpha_2 P_{\tilde{W}}V + \beta$
  5. $P_{\tilde{W}}\left[\sum_{i=1}^n \alpha_i W_i + \beta\right] = \sum_{i=1}^n \alpha_i w_i + \beta$
  6. $P_{\tilde{W}}U = E[U]$ if $Cov(U, \tilde{W}) = 0$

Exercise 6.1. What is the best linear predictor for X_{n+1} in an $AR(p)$ process, based on $X_1, ..., X_n$ for $n > p$?

Example 6.3. Derive the one-step prediction for the $AR(1)$ model. (Here, $h = 1$)

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

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

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

$$\begin{aligned}
\Gamma_n a_n = \gamma_n(h) \quad &\Longrightarrow \quad \frac{\Gamma_n a_n}{\gamma(0)} = \frac{\gamma_n(h)}{\gamma(0)} \\
&\Longrightarrow \quad \begin{pmatrix} 1 & \phi & \cdots & \phi^{n-1} \\ \phi & 1 & \cdots & \phi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^n \end{pmatrix}
\end{aligned}$$

An obvious solution is

$$\begin{aligned}
a_n = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \quad &\Longrightarrow \quad P_nX_{n+1} = \mu + \sum_{i=1}^n a_i(X_{n+1-i} - \mu) \\
&\Longrightarrow \quad P_nX_{n+1} = \sum_{i=1}^n a_i X_{n+1-i} = a_1X_n + 0 = \phi X_n
\end{aligned}$$

Note that

$$\begin{aligned}
MSE &= E\left[(X_{n+1} - P_nX_{n+1})^2\right] \\
&= E\left[(X_{n+1} - \phi X_n)^2\right] = E[Z_{n+1}^2] = \sigma^2
\end{aligned}$$

or you can use the formula of MSE to get

$$\begin{aligned} MSE &= \gamma(0) - a_n^T \gamma_n(h) \\ &= \gamma(0) - \phi\gamma(1) \\ &= \gamma(0) - \phi^2\gamma(0) \\ &= \gamma(0)[1-\phi^2] = \sigma^2 \end{aligned}$$

6.2 Linear Processes

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

Definition 6.6. 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 6.4. Show that $AR(1)$ with $|\phi| < 1$ is a linear process. We know that

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

and we showed before that $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$. Since $|\phi| < 1$ then if $\psi_j = \phi^j$ then $\sum_{j=-\infty}^{\infty} |\psi_j|$ and therefore all assumptions in the definition above are satisfied. So $AR(1)$ is a linear process.

For prediction purposes, we may not want to have dependence on the future innovations (Z_t's). However, the general form of a linear process involves future innovations.

Definition 6.7. A linear process $\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ is **causal** or **future independent** if $\psi_j = 0$ for any $j < 0$.

Example 6.5. Both $AR(1)$ and $MA(q)$ are causal where

$$X_t^{AR(1)} = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$$

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

6.3 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 differencing) are used. When seasonal effect is present, the more general SARIMA model will be used. All these models are two key functions ACF and PACF.

Definition 6.8. $\{X_t, t \in T\}$ is an $ARMA(p, q)$ process if

1) $\{X_t, t \in T\}$ is stationary

2) $X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}$ where $\{Z_t\} \sim WN(0, \sigma^2)$

3) Polynomials $(1 - \phi_1 z - ... - \phi_p z^p)$ and $(1 + \theta_1 z + ... + \theta_q z^q)$ have no common factors/roots (IMPORTANT FOR THE FINAL!)

We say that $\{X_t, t \in T\}$ is an **ARMA process** with mean μ if $\{X_t - \mu\}$ is an $ARMA(p, q)$ process. Recall the backward shift operator $BX_t = X_{t-1}$. With this operator, we can rewrite the ARMA process as

$$(1)\ \underbrace{(1 - \phi_i B - \phi_2 B^2 - ... - \phi_p B^p)}_{\phi(B)} X_t = \underbrace{(1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q)}_{\theta(B)} Z_t$$

The general model above has a unique stationary solution X_t iff $\phi(z) \neq 0$ for all complex $z \in \mathbb{C}$ such that $|z| = 1$. If $\forall z$ such that $|z| = 1$ we have $\phi(z) \neq 0$, then there exists $\delta > 0$ such that

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

Under this condition,

$$(2)\ \frac{1}{\phi(B)} = \sum_{j=-\infty}^{\infty} \chi_j B^j$$

is a linear filter and substituting (2) in (1), we get

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

Since $1/\phi(B)$ and $\theta(B)$ are polynomials, then so is $\psi(B) = \frac{1}{\phi(B)} \times \theta(B)$. We then have

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

where $\psi(B)$ is of degree ∞ and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. (Why?)

Definition 6.9. 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 $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ for any t. This condition is equivalent to

$$\phi(z) = 1 - \phi_1 z_1 - \phi_2 z^2 - ... - \phi_p z^p \neq 0$$

for any $z \in \mathbb{C}$ such that $|z| \leq 1$.

Remark 6.2. If the condition above holds true, then

$$\begin{aligned} \frac{\theta(z)}{\phi(z)} = \psi(z) \quad &\implies \quad \theta(z) = \phi(z) \cdot \psi(z) \\ &\implies \quad 1 + \theta_1 z + ... + \theta_q z^q = (1 - \phi_1 z - ... - \phi_p z^p)(\psi_0 + \psi_1 z + ...) \end{aligned}$$

and we have

$$\begin{aligned} 1 &= \psi_0 \\ \theta_1 &= \psi_1 - \phi_1 \psi_0 \\ &\vdots \end{aligned}$$

Note 5. We try a few special cases:

1) If $\phi(z) = 1$ then $\phi(B)X_t = \theta(B)Z_t$ reduces to $X_t = \theta(B)Z_t = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}$ which is an $MA(q)$ process.

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

From here, we can see that the $AR(p)$ and $MA(q)$ are special cases of $ARMA(p, q)$ processes where

$$\begin{aligned} AR(p) &= ARMA(p, 0) \\ MA(q) &= ARMA(0, q) \end{aligned}$$

6.4 Invertibility

An $ARMA(p, q)$ process $\{X_t\}$ is **invertible** if there exists constants $\{\Pi_j\}$ such that $\sum_{j=0}^{\infty} |\Pi_j| < \infty$ and $Z_t = \sum_{j=0}^{\infty} \Pi_j X_{t-j}$ for all t. Invertibility is equivalent to the condition

$$\theta(z) = 1 + \theta_1 z + ... + \theta_q z^q \neq 0$$

for any $z \in \mathbb{C}$ such that $|z| \leq 1$. Using the same methods above, one can get that

$$\begin{aligned} \Pi_0 &= 1 \\ -\phi_1 &= \Pi_0\theta_1 + \Pi_1 \\ &\vdots \end{aligned}$$

Example 6.6. 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) then provide the causal (invertible) solutions. These are called the $MA(\infty)$ and $AR(\infty)$ representations.

[Causality] We have $\phi(z) = 1 - 0.5z \implies z = 2 \implies |z| > 1$. Since this is outside the unit circle, X_t is causal. We then have

$$\begin{aligned} 1 + 0.4z = (1 - 0.5z)(\psi_0 + \psi_1 z + ...) &\implies \psi_0 = 1, \psi_1 - 0.5\psi_0 = 0.4, \psi_2 - 0.5\psi_1 = 0, ... \\ &\implies \psi_0 = 1, \psi_1 = 0.9, \psi_2 = 0.9(0.5), \psi_3 = 0.9(0.5)^2, ... \end{aligned}$$

We can kind of see the pattern (and prove using induction)

$$\psi_j = \begin{cases} \psi_j = 1 & j = 0 \\ \psi_j = 0.9(0.5)^{j-1} & j \neq 0 \end{cases} \implies X_t = Z_t + 0.9\sum_{j=1}^{\infty}(0.5)^{j-1}Z_{t-j}$$

[Invertibility] We have $\theta(z) = 1 + 0.4z = 0 \implies z = -10/4 \implies |z| > 1$. Since this is outside the unit circle, X_t is invertible. We then have, like above,

$$\begin{aligned} 1 - 0.5z = (1 + 0.4z)(\Pi_0 + \Pi_1 z + ...) &\implies \Pi_0 = 1, \Pi_1 + 0.4\Pi_0 = -0.5, \Pi_2 + 0.4\Pi_1 = 0, ... \\ &\implies \Pi_0 = 1, \Pi_0 = -0.9, \psi_2 = -0.9(-0.4), \psi_3 = -0.9(-0.4)^2, ... \end{aligned}$$

We can kind of see the pattern (and prove using induction)

$$\psi_j = \begin{cases} \psi_j = 1 & j = 0 \\ \psi_j = -0.9(-0.4)^{j-1} & j \neq 0 \end{cases} \implies X_t = Z_t - 0.9\sum_{j=1}^{\infty}(-0.4)^{j-1}Z_{t-j}$$

Remark 6.3. (ACVF of ARMA processes) Consider a causal, stationary process $\phi(B)X_t = \theta(B)Z_t$ with $Z_t \sim WN(0, \sigma^2)$. The $MA(\infty)$ representation of X_t is $X_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j}$ where $E[X_t] = 0$. We have

$$\begin{aligned} \gamma(h) &= E[X_t X_{t+h}] - \underbrace{E[X_t]E[X_{t+h}]}_{=0} \\ &= E\left[\left(\sum_{j=0}^{\infty}\psi_j Z_{t-j}\right)\left(\sum_{j=0}^{\infty}\psi_j Z_{t+h-j}\right)\right] \end{aligned}$$

Notice that $E[Z_t Z_s] = 0$ when $t \neq s$. We then have

$$\gamma(h) = \begin{cases} \sum_{j=0}^{\infty}\psi_j\psi_{j+h}E[Z_j^2] & h \geq 0 \\ \sum_{j=0}^{\infty}\psi_j\psi_{j-h}E[Z_j^2] & h < 0 \end{cases} = \sigma^2\sum_{j=0}^{\infty}\psi_j\psi_{j+|h|}$$

Example 6.7. 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 $|\phi| < 1$. Note that $\phi(z)$ is causal because $1 - \phi z = 0 \implies z = 1/\phi > 1$. It can be shown, with similar methods above, that

$$\psi_j = \begin{cases} \psi_j = \phi(\phi + \theta) & j = 0 \\ \psi_j = \phi^{j-1}(\phi + \theta) & j \neq 0 \end{cases}$$

Now if $h = 0$ then

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

If $h \neq 0$ then

$$\begin{aligned}\gamma(0) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|} &= \sigma^2 \left[\psi_0 \psi_{|h|} + \sum_{j=1}^{\infty} \psi_j \psi_{j+|h|}\right] \\ &= \sigma^2 \left[\phi^{|h|-1}(\theta+\phi) + (\theta+\phi)^2 \sum_{j=1}^{\infty} \phi^{j-1}\phi^{j+|h|}\right] \\ &= \sigma^2 \left[\phi^{|h|-1}(\theta+\phi) + (\theta+\phi)^2 \phi^{|h|-1} \sum_{j=1}^{\infty} \phi^{2j}\right] \\ &= \sigma^2 \left[\phi^{|h|-1}(\theta+\phi) + \frac{(\theta+\phi)^2\phi^{|h|+1}}{1-\phi^4}\right]\end{aligned}$$

6.5 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 $X_{n+1}, X_{n+2}, ..., X_{n+h-1}$. 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 predictors

$$P(X_n | X_{n+1}, ..., X_{n+h-1}) \text{ and } P(X_{n+h} | X_{n+1}, ..., X_{n+h-1})$$

Definition 6.10. The **partial autocorrelation function** (PACF) is shown by $\alpha(h)$ and is defined to be

$$\alpha(h) = \begin{cases} 1 & h = 0 \\ Cor(X_n, X_{n+1}) = \rho(1) & h = 1 \\ Cor\left[X_n - P(X_n|X_{n+1}, ..., X_{n+h-1}), X_{n+h} - P(X_{n+h}|X_{n+1}, ..., X_{n+h-1})\right] & o/w \end{cases}$$

Example 6.8. Derive the PACF for an $AR(1)$ process with $|\phi| < 1$. 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. We then have

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

For $h = 2$ we have

$$\begin{aligned} Cor\left[X_t - P(X_t|X_{t+1}), X_{t+2} - P(X_{t+2}|X_{t+1})\right] &= Cor\left[X_t - \underbrace{P(X_t|X_{t+1})}_{f(X_{t+1})}, \underbrace{X_{t+2} - \phi X_{t+1}}_{Z_{t+2}}\right] \\ &= 0 \end{aligned}$$

Similarly, $\alpha(h) = 0$ for any $h > 2$. This is fairly similar to the ACF of an $MA(1)$ process where the value cuts off after 1. Also notice that this is similar to ACF in the sense that the PACF is symmetric in h so $h < 0$ is omitted from deviations above.

Theorem 6.1. *$\{X_t, t \in T\}$ is a causal $AR(p)$ process if and only if its PACF has the following arguments:*

1) $\alpha(p) \neq 0$

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

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

This theorem show 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 $\alpha(h) = \Phi_{hh}$ for $h \geq 1$ where Φ_{hh} is the last component of the vector $\Phi_h = \Gamma_h^{-1}\gamma_h$ in which

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

Definition 6.11. Based on observations $\{x_1, ..., x_n\}$ with $x_i \neq x_j$ for $i, j = 1, ..., n$. 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

$$\hat{\Phi} = \hat{\Gamma}_h^{-1}\hat{\gamma}_h$$

where the terms on the right are sample estimates.

Example 6.9. Calculate $\alpha(2)$ for an $MA(1)$ process

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

We have shown before that

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

We have $\Phi = \Gamma_h^{-1}\gamma_h$. So $\alpha(h)$ is the last element of Φ_h and

$$h = 1 \implies \Phi_{11} = (\gamma(0))^{-1}\gamma(1) = \frac{\gamma(1)}{\gamma(0)} = \frac{\theta}{1+\theta^2}$$

$$h = 2 \implies \begin{pmatrix} (1+\theta^2)\sigma^2 & \theta\sigma^2 \\ \theta\sigma^2 & (1+\theta^2)\sigma^2 \end{pmatrix}^{-1} \begin{pmatrix} \theta\sigma^2 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\theta(1+\theta^2)\sigma^4}{(1+\theta^2)^2\sigma^4 - \theta^2\sigma^4} \\ \frac{-\theta\sigma^2}{(1+\theta^2)^2\sigma^4 - \theta^2\sigma^4} \end{pmatrix}$$

Where the last element of the case of $h = 2$, in reduced form, is

$$\alpha(2) = \Phi_{22} = \frac{-\theta^2}{1+\theta^2+\theta^4}$$

It can be shown, in general, that

$$\alpha(h) = \Phi_{hh} = \frac{-(-\theta)^h}{\sum_{i=0}^{h}\theta^{2h}}$$

6.6 $ARIMA(p,d,q)$ Processes

Definition 6.12. 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 $\{X_t, t \in T\}$ satisfies an equation of the form

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

Note that $\phi^*(1) = 0 \implies X_t$ is not stationary unless $d = 0$. Therefore, $\{X_t\}$ is stationary iff $d = 0$ in which case it is reduced to an $ARMA(p,q)$ process in the previous case.

Recall that if $\{X_t\}$ exhibits a polynomial trend of the form $m(t) = \alpha_0 + \alpha_1 t + ... + \alpha_d t^d$ then $(1-B)^d X_t$ will not have that trend any more. Therefore, ARIMA models (when $d \neq 0$) are appropriate when the trend in the data is well approximated by a polynomial degree d.

Example 6.10. Consider the process $X_t = 0.8X_{t-1} + 2t + Z_t$ where $\{Z_t\} \sim WN(0,\sigma^2)$. Write this process in $ARIMA(p,d,q)$ format. To get rid of the linear trend $2t$, we perform one time differencing. So

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

If $Y_t = (1-B)X_t$ then the above is written as

$$Y_t - 0.8Y_{t-1} = Z_t - Z_{t-1} + 2 \implies (Y_t - 10) - 0.8(Y_{t-1} - 10) = Z_t - Z_{t-1}$$

Therefore $\{Y_t - 10\}$ is an $ARMA(1,1)$ process. Hence Y_t i s an $ARMA(1,1)$ process with mean 10 and X_t is an $ARIMA(1,1,1)$ process.

We have seen how differencing can be used to remove a trend. Seasonality is a particular type of trend which can be removed by a particular type of differencing. This will be discussed under SARIMA (seasonal ARIMA models)

6.7 $SARIMA(p,d,q) \times (P,D,Q)_S$ Processes

Recall the operator B where $B^k X_t = X_{t-k}$. Clearly $(1-B^k)$ and $(1-B)^k$ are different filters. The latter is performing k times differencing, but the former is differencing once in lag k. In R, we will write

$$\begin{aligned} \text{diff(x,difference=k)} &\equiv (1-B)^k X_t \\ \text{diff(x,lag=k)} &\equiv (1-B^k)X_t \end{aligned}$$

As an example, consider the process $\{X_t\}$ where t represents the month. If there exists a seasonal effect, i.e. $S(t) = S(t+12)$, then the effect of seasonal trend for X_t and X_{t-12} should be the same. That is $Y_t = X_t - X_{t-12}$ should not exhibit any seasonal trends.

Therefore, if we apply differencing using the latter of the above equations, we can (in theory) remove the effect of the seasonal trend. Therefore, fitting an $ARMA(p,q)$ model to the differenced series $Y_t = (1-B^s)X_t$ is the same as fitting the model

$$\phi(B)(1-B^S)X_t = \theta(B)Z_t$$

where S represents the season. This is a special case of SARIMA models.

Definition 6.13. 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 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

$$\begin{aligned}\phi(z) &= 1-\phi_1 z-...-\phi_p z^p\\ \Phi(z) &= 1-\Phi_1 z-...-\Phi_P z^P\\ \theta(z) &= 1+\theta_1 z+...+\theta_q z^q\\ \Theta(z) &= 1+\Theta_1 z+...+\Theta_Q z^Q\end{aligned}$$

Remark 6.4. Notice that the process $\{X_t, t\in T\}$ is causal iff $\phi(z)\neq 0 \wedge \Phi(z)\neq 0$ for all $\forall z : |z|<1$.

Remark 6.5. In practice D is rarely more than 1 and typically P, Q are typically less than 3.

Example 6.11. Write down the equation form of the $ARMA(1,1)_{12}$ process. This is equivalent to

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

and the general form is

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

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

Example 6.12. Derive the ACF of $SARIMA(0,0,1)_{12} = SARIMA(0,0,0)\times(0,0,1)_{12}$. This gives us the general form

$$X_t = Z_t + \Theta_1 Z_{t-12}, Z_t \sim WN(0,\sigma^2)$$

Show, as an exercise, that

$$\gamma(h) = Cov(X_t, X_{t+h}) = \begin{cases}(1+\Theta_1^2)\sigma^2 & h=0\\ \Theta_1\sigma^2 & |h|=12\\ 0 & \text{otherwise}\end{cases}$$

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

7 Box-Jenkins Methodology

To use Box-Jenkins methodology you do the following:

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, PACF or with formal model selection methods
4. Forecast the future with the appropriate model

(See slides for more info)

8 Parameter Estimation in ARMA Processes

This section concentrates on estimation of the parameters $\phi_i, i=1,...,p$ and $\theta_j, j=1,...,q$ as well as σ^2, the variance of the WN, in the $ARMA(p,q)$ process $\phi(B)X_t = \theta(B)Z_t$ where $\{Z_t\}\sim WN(0,\sigma^2)$. We assume that p and q are correctly specified. If the mean of the series in 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 algorithms, and the Durbin-Levinson method. We will only focus on the first two for the remainder of this course.

8.1 Yule-Walker Methods

Consider a causal $AR(p)$ model

$$(1)\ X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t$$

with causal solution $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ where $\{Z_t\} \sim WN(0, \sigma^2)$. Multiply both sides of (1) by X_{t-j} with $j = 0, 1, 2, ..., p$ and taking expectations will give us

$$E[X_t X_{t-j}] - \phi_1 E[X_{t-1} X_{t-j}] - ... - \phi_p E[X_{t-p} X_{t-j}] = E[Z_t X_{t-j}]$$

$$\implies \gamma(j) - \phi_1 \gamma(j-1) - ... - \phi_p \gamma(j-p) = E[Z_t X_{t-j}]$$

We then have

$$\begin{cases} E[Z_t X_{t-j}] = E[Z_t X_t] = E\left[Z_t \sum_{j=0}^{\infty} \psi_j Z_{t-j}\right] = E[Z_t^2] = \sigma^2 & j = 0 \\ E[Z_t X_{t-j}] = 0 & j > 0 \end{cases}$$

So the original equation reduces to

$$\begin{cases} \gamma(0) - \phi_1 \gamma(1) - ... - \phi_p \gamma(p) = \sigma^2 & j = 0 \\ \gamma(j) - \phi_1 \gamma(|j-1|) - .... - \gamma(|j-p|) = 0 & j \neq 0 \end{cases}$$

These are called the **Yule-Walker equations**. This can be easily generalized to a matrix form $\Gamma_p \phi = \gamma_p$. Based on a sample $\{x_1, x_2, ..., x_n\}$ the parameters ϕ and σ^2 can be estimated by

$$\hat{\phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p$$

where the matrices are defined in a similar fashion as the best linear predictor section. The system above is called the **sample Yule-Walker equations**. We can write Yule-Walker equations in terms of ACF too.

Explicitly, if we divide $\hat{\gamma}_p$ by $\gamma(0)$ and multiply it in $\hat{\Gamma}_p$ then

$$\hat{\phi} = \hat{R}_p^{-1} \hat{\rho}_p$$

$$\hat{R}_p = \frac{\hat{\Gamma}_p}{\hat{\gamma}(0)} \implies \hat{R}_p^{-1} = \hat{\Gamma}_p^{-1} \cdot \hat{\gamma}(0)$$

$$\hat{\rho}_p = \hat{\gamma}_p / \hat{\gamma}(0)$$

where $\hat{\sigma}^2 = \hat{\gamma}(0)\left[1 - \hat{\phi} \cdot \hat{\rho}_p\right]$. Notice that $\hat{\gamma}(0)$ is the sample variance of $\{x_1, ..., x_n\}$. Based on a sample $\{x_1, ..., x_n\}$, the above equations will provide the parameter estimates. Using advanced probability theory, it can be shown that

$$\tilde{\phi} = \begin{bmatrix} \tilde{\phi}_1 \\ \vdots \\ \tilde{\phi}_p \end{bmatrix} \sim MVN\left(\phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_p \end{bmatrix}, \frac{\sigma^2}{n} \Gamma_p^{-1}\right)$$

for large n. 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 $\phi_1, ..., \phi_p$.

Example 8.1. Based on the following sample ACF and PACF, an $AR(2)$ has been proposed for the data. Provide the Yule-Walker estimates of the parameters as well as 95% confidence intervals for the parameters in ϕ. The data was collected over a window of 200 points with sample variance 3.69 with the following table:

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

We want to estimate ϕ_1 and ϕ_2 in

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t, \{Z_t\} \sim N(0, \sigma^2)$$

The system is

$$\hat{\phi} = \begin{bmatrix} 1 & 0.821 \\ 0.821 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.821 \\ 0.764 \end{bmatrix} = \begin{bmatrix} 0.594 \\ 0.276 \end{bmatrix}$$

Similarly,

$$\hat{\sigma}^2 = \underbrace{\hat{\gamma}(0)}_{3.69} \left[1 - \hat{\phi} \begin{bmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \end{bmatrix} \right] = 1.112$$

Therefore the estimated model is

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

Now

$$\begin{aligned} \tilde{\phi} \sim N\left(\phi, \frac{\sigma^2}{n}\Gamma_2^{-1}\right) &= N\left(\begin{bmatrix} 0.594 \\ 0.276 \end{bmatrix}, \frac{1.112}{200} \begin{bmatrix} 0.831 & -0.683 \\ -0.683 & 0.831 \end{bmatrix} \right) \\ &= N\left(\begin{bmatrix} 0.594 \\ 0.276 \end{bmatrix}, \begin{bmatrix} 0.005 & -0.004 \\ -0.004 & 0.005 \end{bmatrix} \right) \end{aligned}$$

So the 95% C.I.'s for ϕ_1, ϕ_2 are

$$\begin{aligned} \hat{\phi}_1 \pm 1.96\sqrt{\hat{Var}(\tilde{\phi})} &= 0.594 \pm 1.96\sqrt{0.005} = \quad (0.455, 0.733) \\ \hat{\phi}_2 \pm 1.96\sqrt{\hat{Var}(\tilde{\phi})} &= 0.276 \pm 1.96\sqrt{0.005} = \quad (0.137, 0.415) \end{aligned}$$

(Johnson & Wichard discuss ellipsoid C.I.'s in "Applied Multivariate Statistical Analysis")

9 Likelihood Models

To use **likelihood models**, we have to make some distributional assumptions. Consider $\{X_t, t \in T\}$ to be a Gaussian process. We have that Z_t in $\phi(B)X_t = \theta(B)Z_t$ is i.i.d. $G(0, \sigma)$. Based on the observations $x_1, ..., x_n$ at times $1, 2, ..., n$, the likelihood function of the parameters ϕ, θ and σ^2 is

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

Notice that it is assumed that $E[X_t] = 0$ for all t. To estimate ϕ, θ and σ^2, we maximize the likelihood function above. Usually, it is easier to maximize the log of $L(\theta, \phi, \sigma^2)$ which is called the log-likelihood.

In this likelihood function, $\gamma(h)$ depends on the parameters θ, ϕ and σ^2 in a 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.

10 Forecasting

We discuss how **forecasting** works under our studied processes.

10.1 Forecasting AR(p)

Let $X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + Z_t, Z_t \sim WN\{0, \sigma^2\}$ be a causal $AR(p)$ process. We have

$$\begin{aligned}
\hat{X}_{n+h} &= E[X_{n+h}|X_1, ..., X_n], h > 0 \\
&= E\left[\sum_{j=1}^{h-1} \phi_j X_{n+h-j} + \sum_{j=h}^{p} \phi_j X_{n+h-j} | X_1, ..., X_n\right] + \underbrace{E\left[Z_{n+h}|X_1, ..., X_n\right]}_{=0} \\
&= E\left[\sum_{j=1}^{h-1} \phi_j X_{n+h-j} | X_1, ..., X_n\right] + E\left[\sum_{j=h}^{p} \phi_j X_{n+h-j} | X_1, ..., X_n\right]
\end{aligned}$$

due to the uncorrelatedness of Z_{n+h} with respect to X_k. If $h = 1$, then the above equation becomes

$$\hat{X}_{n+1} = \sum_{j=1}^{p} \phi_j X_{n+1-j}$$

If $h = 2, 3, ..., p$ then remark that

$$\begin{aligned}
j < h &\implies n + h - j > n \\
j \geq h &\implies n + h - j \leq n
\end{aligned}$$

and so

$$\begin{aligned}
\hat{X}_{n+h} &= \sum_{j=h}^{p} \phi_j X_{n+h-j} + \sum_{j=1}^{h-1} \phi_j E\left(X_{n+h-j}|X_1, ..., X_n\right) \\
&= \sum_{j=1}^{h-1} \phi_j \hat{X}_{n+h-j} + \sum_{j=h}^{p} \phi_j X_{n+h-j}
\end{aligned}$$

If $h > p$, then $n + h - j > n$ and

$$\hat{X}_{n+h} = \sum_{j=1}^{p} \phi_j E\left(X_{n+h-j}|X_1, ..., X_n\right) = \sum_{j=1}^{p} \phi_j \hat{X}_{n+h-j}$$

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

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

In $AR(p)$, the $h-$step prediction is a linear combination of the previous steps. We either have the previous p steps in $X_1, ..., X_n$ so we substitute the values (like the $h = 1$ case), or we don't have all or some of them, in which case we recursively predict.

Given a dataset, ϕ_j can be estimated and $\hat{X}_{n+h}$ will be computed.

Example 10.1. 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 bee fitted. If the total sales of the last 3 years have been 9, 11 and 10 million dollars. Forecast this year's total sales (2013) as well as that of 2015.

We have

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

Now

$$\begin{aligned}
\hat{X}_{2013} &= X_{2012} - 0.21X_{2011} = 6.69 \\
\hat{X}_{2015} &= \hat{X}_{2014} - 0.21\hat{X}_{2013} = \hat{X}_{2014} - 0.21(6.69)
\end{aligned}$$

and since

$$\hat{X}_{2014} = \hat{X}_{2013} - 0.21\hat{X}_{2012} = 6.69 - 0.21 \times 9 = 4.8$$

then

$$\hat{X}_{2015} = 4.8 - 0.21(6.69) = 3.4$$

10.2 Forecasting MA(q)

MA processes are linear combinations of white noise. To do forecasting in $MA(q)$, we need to estimate $\theta_1, ..., \theta_q$ as well as "approximate" the innovations $Z_t, Z_{t+1}, ...$ First, consider the very simple case of $MA(1)$ where $X_t = Z_t + \theta Z_{t-1}, \{Z_t\} \sim WN(0, \sigma^2)$. We have

$$\begin{aligned} \hat{X}_{n+h} &= E\left[X_{n+h}|X_1, ..., X_n\right] \\ &= E\left[Z_{n+h}|X_1, ..., X_n\right] + \theta E\left[Z_{n+h-1}|X_1, ..., X_n\right] \end{aligned}$$

If $h = 1$, then the above equation is

$$\begin{aligned} \hat{X}_{n+1} &= \underbrace{E\left[Z_{n+1}|X_1, ..., X_n\right]}_{=0} + \theta E\left[Z_n|X_1, ..., X_n\right] \\ &= \theta E\left[Z_n|X_1, ..., X_n\right] \\ &= \theta Z_n \end{aligned}$$

and if $h > 1$ then the equation becomes

$$\hat{X}_{n+1} = E[Z_{n+h}] + \theta E\left[Z_{\underbrace{n+h-1}_{>n}}|X_1, ..., X_n\right] = 0$$

Now we need to plug in a value for Z_n. We "approximate" the $Z_i's$ by $U_i's$ as follows. Let $U_0 = 0$ and we estimate

$$\hat{Z}_t = U_t = X_t - \theta U_{t-1}, U_0 = 0$$

from the fact that $Z_t = X_t - \theta Z_{t-1}$. We can then get that

$$\begin{aligned} U_0 &= 0 \\ U_1 &= X_1 \\ U_2 &= X_2 - \theta X_1 \\ U_3 &= X_3 - \theta X_2 + \theta^2 X_1 \\ &\vdots \end{aligned}$$

Notice that as $i \to \infty$, U_i will need a convergence condition where $|\theta| < 1$ is sufficient. This was the invertibility condition for $MA(1)$. We see that the $U_i's$ are recursively calculable and for an invertible $MA(1)$ process, we have

$$\hat{X}_{n+h} = \begin{cases} \theta U_n & h = 1 \\ 0 & h > 1 \end{cases}, U_t = X_t - \theta U_{t-1}, U_0 = 0$$

Now consider an $MA(q)$ process $X_t = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}$. We have

$$\begin{aligned} \hat{X}_{n+h} &= E\left[X_{n+h}|X_1, ..., X_n\right] \\ &= E\left[Z_{n+h}|X_1, ..., X_n\right] + \theta_1 E\left[Z_{n+h-1}|X_1, ..., X_n\right] + ... + \theta_q E\left[Z_{n+h-q}|X_1, ..., X_n\right] \end{aligned}$$

If $h > q$ then the above equation's value is zero since we have $n + h - q > n$. If $0 < h \leq q$ then at least some of the terms in the above are non-zero. In particular,

$$\begin{aligned}\hat{X}_{n+h} &= \sum_{j=1}^{q} \theta_j E\left[Z_{n+h-1}|X_1, ..., X_n\right] \\ &= \sum_{j=h}^{q} \theta_j E\left[Z_{n+h-1}|X_1, ..., X_n\right]\end{aligned}$$

and for $j = h, h+1, ..., q$ we know $E[Z_{n+h-j}|X_1, ..., X_n] = Z_{n+h-j}$ and hence

$$\hat{X}_{n+h} = \sum_{j=h}^{q} \theta_j Z_{n+h-j}$$

Similar to $MA(1)$, we approximate $Z_i's$ by $U_i's$, provided the $MA(q)$ process is invertible. That is, $\theta(z) = 1+\theta_1 z+...+\theta_q z^q \neq 0$ for all $|z| \leq 1$. Therefore, assuming that

$$U_0 = U_{-1} = U_{-2} = ... = 0$$

then $U_t = X_t - \sum_{j=1}^{q} \theta_j U_{t-j}$ and

$$\begin{aligned}U_0 &= 0 \\ U_1 &= X_1 \\ U_2 &= X_2 - \theta_1 X_1 \\ U_3 &= X_3 - \theta_2 X_2 + \theta_2\theta_1 X_1 \\ &\vdots\end{aligned}$$

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 $U_0 = U_i = ... = 0$, $i < 0$ and $U_t = X_t - \sum_{j=1}^{q} \theta_j U_{t-j}$ for $t = 1, 2, 3, ...$

Example 10.2. Consider the $MA(1)$ process $X_t = Z_t + 0.5Z_{t-1}$ where $\{Z_n\} \sim WN(0, \sigma^2)$. If $X_1 = 0.3, X_2 = -0.1, X_3 = 0.1$, predict X_4, X_5. Notice that $\hat{X}_5 = \hat{X}_{3+2}$ which is a 2-step prediction based on the history $X_1 = X_2 = X_3$. Since this is an $MA(1)$ model, hence 1-correlated, $\hat{X}_5 = 0$. For X_4 we have

$$\hat{X}_4 = \sum_{j=1}^{1} = \theta_j U_{3+1-j} = \theta_1 U_3 = 0.5U_3$$

where

$$\begin{aligned}U_0 &= 0 \\ U_1 &= X_1 - 0.5U_0 = X_1 = 0.3 \\ U_2 &= X_2 - 0.5U_1 = -0.1 - (0.5)(0.3) = -0.25 \\ U_3 &= X_3 - 0.5U_2 = 0.1 - (0.5)(-0.25) = 0.225\end{aligned}$$

and hence $\hat{X}_4 = 0.5(0.225) = 0.1125$.

Example 10.3. Consider the $MA(1)$ process $X_t = Z_t + \theta Z_{t-1}$ with $\{Z_t\} \sim WN(0, \sigma^2)$ and $|\theta| < 1$. Show that the one-step predictor $\hat{X}_{n+1} = \theta U_n$ is equal to the predictor

$$\hat{\hat{X}}_{n+1} = -\sum_{j=1}^{n} (-\theta)^j X_{n-j+1}$$

This is by definition of U_n which we can write the closed form

$$U_n = X_n + \sum_{i=1}^{n-1}(-\theta)^i X_{n-i}, n \geq 2$$

and hence

$$\hat{X}_{n+1} = \theta U_n = \theta X_n - \sum_{i=1}^{n-1}(-\theta)^{i+1} X_{n-i} = -\sum_{i=0}^{n-1}(-\theta)^{i+1} X_{n-i} = -\sum_{j=1}^{n}(-\theta)^j X_{n-j+1} = \hat{\hat{X}}_{n+1}$$

Clearly for $n = 0, 1$ we have $\hat{X}_{n+1} = \hat{\hat{X}}_{n+1}$ as well. This shows that even in the MA process, the predictor may be written as a linear function of the "history".

10.3 Forecasting ARMA(p,q) Processes

For the causal and invertible $ARMA(p, q)$ process $\phi(B)X_t = \theta(B)Z_t$ where $\{Z_t\} \sim WN(0, \sigma^2)$, the predictors for $MA(q)$ and $AR(p)$ are "combined". We will not go into the theory of forecasting in ARMA processes and will use R for that matter. For example

$$\hat{X}_{n+1} = \sum_{j=1}^{p} \phi_j X_{n+1-j} + \sum_{i=1}^{q} \theta_i U_{n+1-i}$$

subject to some conditions.

Index

ARIMA process, 18
ARMA process, 13
autocorrelation function, 3
autocovariance function, 3
autoregressive process, 3, 9, 10

best linear predictor, 12
bias-variance tradeoff, 5
Box-Jenkins, 13

causal, 13, 14
classical decomposition, 2, 6
covariance function, 2

Difference Sign Test, 6
differencing, 7
double exponential smoothing, 9

even function, 3

first-order moving average, 3
forecasting, 21
future independent, 13

Gaussian time series, 10

innovation, 9
invertible, 14

level, 8
likelihood models, 21
linear process, 13

mean function, 2
moving average process, 9
MSE, 10

partial autocorrelation function, 16

q-correlated, 9
q-dependent, 9

Runs Test, 6

sample ACF, 4
sample autocorrelation function, 4
sample autocovariance function, 4
sample mean, 4
SARIMA process, 18
seasonal component, 2
Shapiro-Wilk Test, 6
simple exponential smoothing, 9
slope, 8
stochastic process, 1
strictly (strong) stationary, 2

time series model, 1
trend, 1

weak stationarity, 2
weakly stationary, 2

Yule-Walker equations, 20