Time Series Analysis Book Notes

Dustin Leatherman

February 11, 2020

Contents

1 Chapter 1 - Characteristics of Time Series . . . 2

1.1 Definitions . . . 2

1.2 Mean . . . 3

1.2.1 Population . . . 3

1.2.2 Sample . . . 3

1.3 Autocovariance . . . 4

1.3.1 Covariance of Linear Combos . . . 4

1.3.2 Moving Average . . . 4

1.3.3 Random Walk . . . 5

1.3.4 Cross-covariance . . . 5

1.4 Autocorrelation (ACF) . . . 5

1.4.1 Cross-correlation . . . 5

1.5 Stationary Time Series . . . 5

1.5.1 Strict Stationary . . . 5

1.5.2 Weakly Stationary . . . 6

1.5.3 Trend Stationarity . . . 6

1.5.4 Autocovariance Function Properties . . . 6

1.5.5 Joint Stationarity . . . 7

1.5.6 Linear Process . . . 7

1.5.7 Gaussian (Normal) Process . . . 7

1.6 Vector Time Series . . . 7

1.6.1 Mean . . . 7

1.6.2 Autocovariance Matrix . . . 8

1.7 Multidimensional Series . . . 8

1.7.1 Mean . . . 8

1.7.2 Autocovariance . . . 8

1.7.3 Autocorrelation . . . 8

1.7.4 Variogram . . . 8

2 Chapter 2 - Time Series Regression and Exploratory Data Analysis 9
2.1 Exploratory Data Analysis . . . 9
2.1.1 Trend Stationary Models . . . 9
2.1.2 Differencing . . . 10
2.1.3 Trig Identities to Discover a Signal in Noise . . . 11
2.1.4 Smoothing . . . 11

3 Chapter 3 - ARIMA Models 12
3.1 Autoregressive Moving Average (ARMA) Models . . . 12
3.1.1 Autoregressive Models . . . 12
3.1.2 Moving Average Models . . . 14
3.1.3 Autoregressive Moving Average (ARMA) Models . . . 15
3.2 Difference Equations . . . 17
3.2.1 Homogeneous Difference Equation (1) . . . 17
3.2.2 Homogeneous Difference Equation (2) . . . 17
3.3 Autocorrelation and Partial Autocorrelation . . . 17
3.4 Forecasting . . . 17
3.4.1 Best Linear Predictors (BLP) . . . 17
3.4.2 M-step-ahead Predictor . . . 18
3.4.3 Truncated Predictors . . . 18
3.4.4 Backcasting . . . 18
3.5 Estimation . . . 18
3.5.1 Method of Moments Estimators . . . 18
3.5.2 Newton-Raphson & Scoring Algorithms . . . 20

Time Series Analysis and Its Applications - 4th Edition

1 Chapter 1 - Characteristics of Time Series

1.1 Definitions

- **Filtered Series**: A linear combination of values in a time series.
- **Autoregression**: A time series where the current value x_t is dependent on a

function of previous values $x_{t-1}, x_{t-2}, \ldots$, etc. The order of Autoregression is dependent on the number of previous values.

- **Random Walk (with Drift)**: An AR(1) model with some constant δ called *drift*. When $\delta = 0$, this is called a Random Walk.

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

- **Signal-to-noise Ratio (SNR)**: $SNR = \frac{A}{\sigma_w}$
  - A: Amplitude of the Waveform
  - σ_w: Additive noise term
  - Note: A sinusoidal wave form can be written as $Acos(2\pi\omega t + \phi)$

- **Weak Stationarity**: A time series where the mean is constant. In this case, $h = |s - t|$ where h is

the separation between points x_s and x_t is important.

- **Note**: Many modeling practices attempt to reduce or transform a time series to white noise to then model it. This is known as *pre-whitening* and is typically done prior to performing Cross-Correlation Analysis (CCA).

1.2 Mean

1.2.1 Population

$\mu_{xt} = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx$

1. Moving Average $\mu_{vt} = E(v_t) = \frac{1}{3}[E(w_{t-1}) + E(w_t) + E(w_{t+1})] = 0$
2. Random Walk with Drift $\mu_{xt} = E(x_t) = \delta t + \sum_{j=1}^{t} E(w_j) = \delta t$

1.2.2 Sample

$$\begin{aligned}
\bar{x} =& \frac{1}{n}\sum_{t=1}^{n} x_t \\
var(\bar{x}) =& \frac{1}{n^2} cov(\sum_{t=1}^{n} x_t, \sum_{s=1}^{n} x_s) \\
=& \frac{1}{n}\sum_{h=-n}^{n}(1 - \frac{|h|}{n})\gamma_x(h)
\end{aligned} \tag{1}$$

1.3 Autocovariance

- the second moment product for all s and t.
- Measures linear dependence between two points on the same series observed at different times.

Population: $\gamma_x(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$
Sample: $\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})$ where $\hat{\gamma}(-h) = \hat{\gamma}(h) \forall h \in [0, n-1]$

- This estimator guarantees a non-negative result.

1.3.1 Covariance of Linear Combos

Let U and V be linear combinations with finite variance of the randome variables X_j and Y_k.

$$\begin{aligned} U =& \sum_{j=1}^{m} a_j X_j \\ V =& \sum_{k=1}^{r} b_k Y - k \end{aligned} \tag{2}$$

Then,

- $cov(U,V) = \sum_{j=1}^{m} \sum_{k=1}^{r} a_j b_k cov(X_j, Y_k)$
- $cov(U,U) = var(U)$

1.3.2 Moving Average

$\gamma_v(s,t) = cov(v_s, v_t) = cov(\frac{1}{3}(w_{s-1} + w_s + w_{s+1}), \frac{1}{3}(w_{t-1} + w_t + w_{t+1}))$

$$\gamma_v(s,t) = \begin{cases} \frac{3}{9}\sigma_w^2 & s = t \\ \frac{2}{9}\sigma_w^2 & |s-t| = 1 \\ \frac{1}{9}\sigma_w^2 & |s-t| = 2 \\ 0 & |s-t| > 2 \end{cases} \tag{3}$$

1.3.3 Random Walk

$\gamma_x(s,t) = cov(x_s, x_t) = cov(\sum_{j=1}^{s} w_j, \sum_{k=1}^{t} w_k) = min(s,t)\sigma_w^2$

- covariance of walk is dependent on time opposed to lag, unlike Linear combos and Moving Average.

1.3.4 Cross-covariance

Covariance between two time series x and y

Population: $\gamma_{xy}(s,t) = cov(x_s, y_t) = E[(x_s - \mu_{xs})(y_t - \mu_{yt})]$

Sample: $\hat{\gamma_{xy}}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y})$

1.4 Autocorrelation (ACF)

Measures the linear predictability of a time eseries at time t (x_t) using only the value x_s.

Population: $\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$

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

- For large sample sizes, the sample ACF is $\sim N(0, \frac{1}{n})$

1.4.1 Cross-correlation

Correlation between two different time series x and y

Population: $\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}}$

Sample: $\hat{\rho_{xy}}(h) = \frac{\hat{\gamma_{xy}}(h)}{\sqrt{\hat{\gamma_x}(0)\hat{\gamma_y}(0)}}$

- For large samples, $\hat{\rho_{xy}} \sim N(0, \frac{1}{n})$

1.5 Stationary Time Series

A measure of regularity over the course of a time series.

1.5.1 Strict Stationary

A time series for which the probabilistic behavior of every collection of values $(x_{t1}, x_{t2}, ..., x_{tk})$ is identical to that of the time shifted set $(x_{t1+h}, ..., x_{tk+h})$. i.e. $Pr(x_{t1} \leq c1, ..., x_{tk} \leq c_k) = Pr(x_{t1+h} \leq c1, ..., x_{tk+h} \leq c_k)$

Mean: $\mu_t = \mu_s$ for all s and t indicating that μ_t is *constant*.

Autocovariance: $\gamma(s,t) = \gamma(s+h, t+h)$

- The process depends only on time *difference* between s and t rather than the actual times.

This definition is too restrictive and unrealistic for most applcations.

1.5.2 Weakly Stationary

A time series for which

1. μ_t is constant and does not depend on time t
2. $\gamma(s,t)$ depends on s and t only through their difference $|s-t|$

If a time series is normal, then it implies it is strict stationary.

1. Autocorrelation Function (ACF) $\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}$
   - Moving Averages **are** Stationary
   - Random Walks are **not** Stationary since the mean depends on time

1.5.3 Trend Stationarity

When the Mean function is dependent on time but the Autocovariance function is not, the model can be considered as having a stationary behavior around a linear trend. a.k.a trend stationarity.

1.5.4 Autocovariance Function Properties

1. $\gamma(h)$ is non-negative definite meaning that that variance and linear combinations of such will never be negative.

   $0 \leq var(a_1 x_1 + ... + 1_n x_n) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_j a_k \gamma(j-k)$
2. $\gamma(h=0) = E[(x_t - \mu)^2]$ is the variance of the time series and thus Cauchy-Swarz inequality implies $|\gamma(h)| \leq \gamma(0)$
3. $\gamma(h) = \gamma(-h)$ for all h. i.e. symmetrical

1.5.5 Joint Stationarity

Both time series are stationary and the Cross-Covariance Function is a function only of lag h.

$\gamma_{xy}(h) = cov(x_{t+h}, y_t) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)]$

Cross-correlation Function (CCF) of a jointly stationary time series x_t and y_t is defined as $\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}$

Generally $cov(x_2, y_1) \neq cov(x_1, y_2)$ and $\rho_{xy}(h) \neq \rho_{xy}(-h)$; however, $\rho_{xy}(h) = \rho_{yx}(-h)$.

1.5.6 Linear Process

Linear combination of white noise variates w_t, given by $x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}$, $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$

1. Autocovariance for $h \geq 0$ $\gamma_x(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j$

   models that do not depend on the future are considered **causal**. In causal linear processes, $\psi_j = 0$ for $j < 0$

1.5.7 Gaussian (Normal) Process

A process is said to be Gaussian if the n-dimensional vectors $x = (x_{t1}, x_{t2}, ..., x_{tn})^T$ for every collection of distinct time points $t_1, t_2, ..., t_n$ and every positive integer n have a multivariate normal distribution.

- A Gaussian Process is Strictly Stationary. Gaussian Time series form the basis of modeling many time series.
- **Wold Decomposition**: A stationary non-deterministic time series is a causal linear process with $\Sigma\psi_j^2 < \infty$

1.6 Vector Time Series

$\underset{(p\times 1)}{x_t} = (x_{t1}, ..., x_{tp})^T$

1.6.1 Mean

1. Population $\vec{\mu} = E(x_t)$
2. Sample Vector $\bar{x} = n^{-1} \sum_{t=1}^{n} x_t$

1.6.2 Autocovariance Matrix

1. Population $\Gamma(h) = E[(x_{t+h} - \mu)(x_t - \mu)^T]$
   - $\Gamma(-h) = \Gamma^T(h)$ holds
2. Sample $\underset{(p \times p)}{\hat{\Gamma}(h)} = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})^T$
   - $\hat{\Gamma}(-h) = \hat{\Gamma}^T(h)$ holds

1.7 Multidimensional Series

In cases where a series is indexed by more than time alone, a *multidimensional process* can be used. For example, a coordinate may be defined as (s_1, s_2). Thus, $\underset{(r \times 1)}{s} = (s_1, ..., s_r)^T$ where s_i is the coordinate of the ith index.

1.7.1 Mean

- **Population**: $\mu = E(x_s)$
- **Sample**: $\bar{x} = (S_1 S_2 ... S_r)^{-1} \Sigma_{s1} \Sigma_{s2} ... \Sigma_{sr} x_{s1,s2,...,sr}$

1.7.2 Autocovariance

- **Population**: $\gamma(h) = E[(x_{s+h} - \mu)(s_x - \mu)]$ with multidimensional lag vector h, $h = (h_1, ..., h_r)^T$
- **Sample**: $\hat{\gamma}(h) = (S_1 S_2 ... S_r)^{-1} \Sigma_{s1} \Sigma_{s2} ... \Sigma_{sr} (x_{s+h} - \bar{x})(x_s - \bar{x})$

1.7.3 Autocorrelation

- **Sample**: $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$ with

$\hat{\gamma}$ defined above

1.7.4 Variogram

Sampling requirements for multidimensional processes are severe since there must be some uniformity across values. When observations are irregular in time space, modifications to the estimators must be made. One such modificaiton is the variogram.

$2V_x(h) = var(x_{s+h} - x_s)$

- **Sample Estimator**: $2\hat{V}_x(h) = \frac{1}{N(h)}\Sigma_s(s_{x+h} - x_s)^2$
  - $N(h)$: Number of points located within h

Issues

- negative estimators for the covariance function occur
- Indexing issues?

2 Chapter 2 - Time Series Regression and Exploratory Data Analysis

2.1 Exploratory Data Analysis

It is necessary for time series data to be stationary so lags are possible. It is tough to measure time series if the dependence structure is not regular. At bare minimum, the autocovariance and mean functions must be stationary for some period of time.

2.1.1 Trend Stationary Models

$x_t = \mu_t + y_t$

- x_t: Observations
- μ_t: Trend
- y_t: Stationary Process

Strong trends often obscure behavior of the stationary process so detrending is a good first step.

$$\begin{aligned} \hat{y_t} =& x_t - \hat{\mu_t} \\ =& x_t - (\beta_0 + \beta_1 t) \end{aligned} \tag{4}$$

Using $\hat{\mu_t} = \beta_0 + \beta_1 t$ detrends the data.

2.1.2 Differencing

$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_1 + y_t - y_{t-1}$

First Difference Notation: $\nabla x_t = x_t - x_{t-1}$

1. Backshift Used to specify a specific difference from a given point in a time series. When $k < 0$, it becomes a forward-shift operator.

   $B^k x_t = x_{t-k}$

   A given difference can be represented as: $\nabla^d x_t = (1 - B)^d x_t$

   (a) Example - Second Difference
   $\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2) x_t = x_t - 2x_{t-1} + x_{t-2}$

   (b) Example - Fractional Differencing
   $-0.5 < d < 0.5$
   $\nabla^{0.5} x_t = (1 - B)^{0.5} x_t$
   Typically used for environmental time series in hydrology.

2. Pros

   - No parameters estimated in differencing operation
   - Not viable when goal is to coerce data to stationarity

3. Cons

   - does not yield an estimate of the stationary process y_t
   - Detrending more viable if trend is fixed

4. Transformations Just as transformations can fix non-normality, so can they fix non-stationarity. The Box-Cox family transformations are useful.

$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda & \lambda \neq 0 \\ log X_t & \lambda = 0 \end{cases} \tag{5}$$

2.1.3 Trig Identities to Discover a Signal in Noise

$$\begin{aligned}
x_t =& Acos(2\pi\omega t + \phi) + w_t \\
acos(2\pi\omega t + \phi) =& \beta_1 cos(2\pi\omega t) + \beta_2 sin(2\pi\omega t) \\
\beta_1 =& acos(\phi) \\
\beta_2 =& -asin(\phi) \\
\omega = 1/50 & \\
x_t =& \beta_1 cos(2\pi t/50) + \beta_2 sin(2\pi t/50) + w_t
\end{aligned} \tag{6}$$

2.1.4 Smoothing

Let a Moving Average be defined as
$m_t = \sum_{j=-k}^{k} a_j x_{t-j}$
where
$a_j = a_{-j} \geq 0, \sum_{j=-k}^{k} a_j = 1$

1. Kernal smoothing

   Moving Average smoother that uses a weight function (kernel) to average observations.

   $$\begin{aligned}
   m_t =& \sum_{i=1}^{n} w_i(t) x_i \\
   w_i(t) =& K(\frac{t-i}{b}) / \sum_{j=1}^{n} K(\frac{t-j}{b})
   \end{aligned} \tag{7}$$

   where $K(.)$ is a kernel function.

   (a) Example - Original Kernel Function
   $K(z) = \frac{1}{\sqrt{(2\pi)}} exp(-z^2/2)$

2. Lowess

   KNN Regression followed by robust weighted regression to obtain smoothed values.

3. Splines

   Given the following:

$$\begin{aligned} x_t =& m_t + w_t \\ m_t =& \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 \\ t =& 1, ..., n \end{aligned} \tag{8}$$

Let t be divied into k intervals called *knots*. In each interval, fit a polynomial regression model. The most common is a *cubic spline* where the Order is 3 (as m_t is defined).

(a) Smoothing Spline
The following is a compromise between the model fit (smoothness) and the data (no smoothness).
$\sum_{t=1}^{n}[x_t - m_t]^2 + \lambda \int (m_t'')^2 dt$
$\lambda > 0$ controls the degree of smoothness.

3 Chapter 3 - ARIMA Models

3.1 Autoregressive Moving Average (ARMA) Models

3.1.1 Autoregressive Models

Autoregressive models are based on the idea that the current value of the series, x_t can be explained as a function of p past values, $x_{t-1}, x_{t-2}, ..., x_{t-p}$ where p determines the number of steps in the past needed to forecast the current value.

AR(P): Autoregressive model of the order P
$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t$
$\alpha = \mu(1 - \phi_1 - ... - \phi_p)$ where x_t is stationary, $w_t \sim wn(0, \sigma_w^2)$, and $\phi_1, \phi_2, ..., \phi_p$ are constants ($\phi_p \neq 0$).

This model can be expressed using the **Backshift** operator. This is known as the autoregressive operator.

$$\begin{aligned} \phi(B)x_t =& w_t \\ (1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)x_t =& w_t \end{aligned} \tag{9}$$

1. AR(1) An AR(1) model can be represented as a linear process given by $x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}$ assuming $|\phi| < 1$ and $var(x_t) < \infty$
   $E(x_t) = \sum_{j=0}^{\infty} \phi^j E(w_{t-j}) = 0$
   $\gamma(h) = cov(x_{t+h}, x_t) = \frac{\sigma_w^2 \phi^h}{1-\phi^2}$ for $h \geq 0$

2. Explosive AR Models When $|\phi| > 1$, the model is considered explosive because it grows without bound. A stationary model can be obtained by iterating k steps forward producing the model

   $x_t = -\sum_{j=1}^{\infty} \phi^{-j} w_{t+j}$

   This model is future dependent and thus useless. When a process does not depend on the future, its considered *causal.*

   However, explosive models have *causal* counterparts.

   Given

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

   - $|\phi| > 1$
   - $w_t \sim N(0, \sigma_w^2)$
   - $E(x_t) = 0$
   - $\gamma_x(h) = \frac{\sigma_w^2 \phi^{-2} \phi^{-h}}{1-\phi^{-2}}$

   The causal process is defined by

   $y_t = \phi^{-1} y_{t-1} + v_t$

   - $v_t \sim N(0, \sigma_w^2 \phi^{-2})$

   y_t is stochastically equal to x_t. i.e. all finite distributions of the processes are the same.

$$\begin{aligned} x_t =& 2x_{t-1} + w_t \\ \sigma_w^2 =& 1 \\ y_t =& \frac{1}{2} y_{t-1} + v_t \\ \sigma_v^2 =& \frac{1}{4} \end{aligned} \tag{10}$$

   For larger orders of AR models, it is more effective to match coefficients to find a stationary solution.

$$
\begin{aligned}
\phi(B)x_t =& w_t \\
\phi(B) =& 1 - \phi B \\
|\psi| <& 1 \\
x_t =& \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B) w_t \\
\psi(B) =& \sum_{j=0}^{\infty} \psi_j B^j \\
\psi_j =& \phi^j
\end{aligned} \tag{11}
$$

3.1.2 Moving Average Models

MA(q): $x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_q w_{t-q}$

- $w_t \sim wn(0, \sigma_w^2)$
- $[\theta_1, \theta_q], \theta_1 \neq 0$ are parameters

Operator Form

$$
\begin{aligned}
x_t =& \theta(B) w_t \\
\theta(B) =& 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q
\end{aligned} \tag{12}
$$

Unlike the *autoregressive process*, the moving average process is stationary for any values for parameters $[\theta_1, \theta_q]$

1. MA(1)

   Consider $x_t = w_t + \theta w_{t-1}$ with $E(x_t) = 0$

$$
\begin{aligned}
\gamma(h) =& \begin{cases} (1+\theta^2)\sigma_w^2 & h = 0 \\ \theta\sigma_w^2 & h = 1 \\ 0 & h > 1 \end{cases} \\
\rho(h) =& \begin{cases} \frac{\theta}{(1+\theta^2)} & h = 1 \\ 0 & h > 1 \end{cases}
\end{aligned} \tag{13}
$$

   (a) Properties

- $\forall\ \theta\ |\rho(1)| < 0.5$
- x_t and x_{t-1} correlated but not with \$x$_{\text{t - 2}}$, x$_{\text{t - 3}}$, …\$ This is unlike an AR(1) model where $cor(x_t, x_{t-k}) \neq 0$

2. Invertibility of MA Models A process with an infinite AR representation is called an *invertible* process.

   Discovering an Invertible Model

   (a) Reverse roles of x_t and w_t

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

   (b) Iterate Backwards k times to get the infinite AR representation of the model

   - $w_t = \sum_{j=0}^{\infty}(-\theta)^j x_{t-j}$ if $|\theta| < 1$

$$\begin{aligned} x_t =& \theta(B)w_t \\ \theta(B) =& 1 + \theta B \\ \pi(B)x_t =& w_t \text{ when } |\theta| < 1 \\ \pi(B) =& \theta^{-1}(B) \\ =& \sum_{j=0}^{\infty}(-\theta)^j B^j \end{aligned} \tag{14}$$

$$\begin{aligned} \theta(z) =& 1 + \theta z \text{ for } |z| \leq 1 \\ \pi(z) =& \theta^{-1}(z) \\ =& (1 + \theta z)^{-1} \\ =& \sum_{j=0}^{\infty}(-\theta)^j z^j \end{aligned} \tag{15}$$

3.1.3 Autoregressive Moving Average (ARMA) Models

A time series {\$x$_{\text{t}}$: \ t = 0, ± 1, ± 2, …\$} is ARMA(p, q) if it's *stationary* and

$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}$
$\phi_p \neq 0,\ \theta_q \neq 0,\ \sigma_w^2 > 0$
p: Autoregressive order q: Moving Average order

When $E(x_t) \neq 0$, $\alpha = \mu(1 - \phi_1 - ... - \phi_p)$
$x_t = \alpha + \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}$
Where $w_t \sim wn(0, \sigma_w^2)$
Operator Form: $\phi(B)x_t = \theta(B)w_t$

Be aware that this model can become complicated when multiplying both sides by another operator. This leads to *parameter redundancy* and can lead to (1) and over parameterized model and (2) incorrect inferences.

1. Properties

   - AR Polynomial: $\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p,\ \phi_p \neq 0$
   - MA Polynomial: $\theta(z) = 1 + \theta_1 z + ... + \theta_q z^q,\ \theta_q \neq 0$
     - z is a complex number
   - $\phi(z)$ and $\theta(z)$ have no common factors
   - ARMA(p, q) must be *causal.*

   (a) Causality An ARMA(p,q) model is said to be causal if

      - it can be written as a one sided linear process.
      - if and only if $\phi(z) \neq 0$ for $|z| < 1$

      $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B)w_t$

      - $\psi(B) = \sum j = 0^{\infty} \psi_j B^j$
      - $\sum_{j=0}^{\infty} |\psi_j| < \infty$
      - $\psi_0 = 1$

      The coefficients can be determined by solving
      $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)},\ |z| \leq 1$

   (b) Invertibility
      An ARMA(p, q) model is said to be invertible if the time series can be written as
      $\pi(B)x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = w_t$

      - $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$
      - $\sum_{j=0}^{\infty} |\pi_j| < \infty$
      - $\pi_0 = 1$

      Also if and only if $\theta(z) \neq 0$ for $|z| < 1$
      The coefficients can be determined by solving
      $\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)},\ |z| \leq 1$

3.2 Difference Equations

Let the sequence of numbers \$u$_0$, u$_1$,...\$ be defined assess \$u$_\mathrm{n}$ - α u$_\mathrm{n\ -\ 1}$ = 0, \ $\alpha \neq$ 0, \ n = 1,2,...\$

ARMA and ACFs **are** Difference Equations.

3.2.1 Homogeneous Difference Equation (1)

$u_n - \alpha u_{n-1} = 0$

Solution: $u_n = \alpha u_{n-1} = \alpha^n u_0$

Operator Notation: $(1 - \alpha B)u_n = 0$

Polynomial: $\alpha(z) = 1 - \alpha z$

3.2.2 Homogeneous Difference Equation (2)

\$u$_\mathrm{n}$ - α_1 u$_\mathrm{n\ -\ 1}$ - α_2 u$_\mathrm{n\ -\ 2}$ = 0, \ $\alpha_2 \neq$ 0, \ n = 2,3,...\$

Solution

- $u_n = c_1 z_1^{-n} + c_2 z_2^{-n}$ when $z_1 \neq z_2$
- $u_n = z_0^{-n}(c_1 + c_2 n)$

Polynomial: $\alpha(z) = 1 - \alpha_1 z - \alpha_2 z^2$

3.3 Autocorrelation and Partial Autocorrelation

Partial Autocorrelation is the marginal effect of a given lag on a comparison point. The linear effect of values between the two points are removed.

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

and

$\phi_{hh} = corr(x_{t+h} - \hat{x_{t+h}}, x_t - \hat{x_t}),\ h \geq 2$

3.4 Forecasting

Goal: Predict future values of a time series, \$x$_\mathrm{n\ +\ m}$, \ m = 1,2,...\$

x_3^2 is a one-step ahead of x_3 given $x_1,\ x_2$

3.4.1 Best Linear Predictors (BLP)

Linear predictors of the following form that minimize square prediction error.

$x_{n+m}^n = \alpha_0 + \sum_{k=1}^n \alpha_k x_k$ where $\alpha_0, \alpha_1, ..., \alpha_n$ are real numbers.

$E[(x_{n+m} - x_{n+m}^n)x_k] = 0,\ k = [0 : n],\ x_0 = 1$

The final form of the BLP is: $x_{n+m}^n = \mu + \sum_{k=1}^n \alpha_k(x_k - \mu)$

The equations within the aforementioned expectation are called *prediction equations*. They can be written in matrix notation: $\Gamma_n \phi_n = \gamma_n$.

- $\underset{(n\times n)}{\Gamma_n} = (\gamma(k-j))_{j,k=1}^n$
- $\phi_n = \Gamma_n^{-1}\gamma_n$

mean square one-step-ahead prediction error: $P_{n+1}^n = E(x_{n+1} - x_{n+1}^n)^2 = \gamma(0) - \gamma_n^T \Gamma_n^{-1} \gamma_n$

3.4.2 M-step-ahead Predictor

$x_{n+m}^n = \phi_{n1}^{(m)} x_n + \phi_{n2}^{(m)} x_{n-1} + ... + \phi_{nn}^{(m)} x_1$
Matrix Notation: $\Gamma_n \phi_n^{(m)} = \gamma_n^{(m)}$
$P_{n+m}^n = E(x_{n+m} - x_{n+m}^n)^2 = \gamma(0) - \gamma_n^{(m)T} \Gamma_n^{-1} \gamma_n^{(m)}$

3.4.3 Truncated Predictors

When dealing with an infinite time series, not all data for all points in time will be present so a *truncated predictor* would be used.

For ARMA(p, q) models, the truncated predictors for m = 1,2,... are $x_{n+m}^n = \phi_1 x_{n+m-1}^n + ... + \phi_p x_{n+m-p}^n + \theta_1 w_{n+m-1}^n + ... + \theta_q w_{n+m-q}^n$

3.4.4 Backcasting

Predicting x_{1-m} for \$m = 1,2,...\$ based on the data $[x_1, x_n]$
$x_{1-m}^n = \sum_{j=1}^n \alpha_j x_j$

3.5 Estimation

Goal is to estimate $[\phi_1, \phi_p]$ and θ_1, θ_q and σ_w^2 for an ARMA model.

3.5.1 Method of Moments Estimators

Equate sample moments to population moments and solve for parameters in terms of sample moments. Hit or miss in terms of estimators

Good for AR(p) models: $x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t$

1. Yule-Walker Estimators (AR(p) Models)

   Approx. Gaussian for large sample sizes for AR(p) models ($N(0, \sigma_w^2 \Gamma_p^{-1})$).

- Best for AR(p) models since they are linear models and Yule-Walker estimators are essentially least squares.

$$
\begin{aligned}
\Gamma_p \phi &= \gamma_p \\
\sigma_w^2 &= \gamma(0) - \phi^T \gamma_p \\
\underset{(p\times p)}{\Gamma_p} &= \gamma(k-j)_{j,k=1}^{p} \\
\underset{(p\times 1)}{\phi} &= (\phi_1, ..., \phi_p)^T \\
\underset{(p\times 1)}{\gamma_p} &= (\gamma(1), ..., \gamma(p))^T
\end{aligned}
\tag{16}
$$

Using *method of moments*, the following estimates are derived:

$$
\begin{aligned}
\hat{\phi} &= \widehat{\Gamma_p^{-1}} \hat{\gamma}_p \\
\widehat{\sigma_w^2} &= \widehat{\gamma(0)} - \hat{\gamma_p}^T \widehat{\Gamma_p^{-1}} \hat{\gamma_p}
\end{aligned}
\tag{17}
$$

Durbin-Levinson algo calculates $\hat{\phi}$ without inverting $\hat{\Gamma_p}$ by replacing $\gamma(h)$ with $\hat{\gamma}$. In doing so, $\underset{(h\times 1)}{\hat{\phi}} = (\widehat{\phi_{h1}}, ..., \widehat{\phi_{hh}})^T$ is calculated.

- $\widehat{\phi_{hh}}$ = sample PACF

(a) Large Sample Distribution of PACF For a causal AR(p) process (n → ∞) $\sqrt{n}\ \widehat{\phi_{hh}} \rightarrow N(0,1)$

2. MA(p) Let $$.

$$
\begin{aligned}
x_t &= 1_t + \theta w_{t-1},\ |\theta| < 1 \\
&= \sum_{j=1}^{\infty} (-\theta)^j x_{t-j} + w_t \\
\gamma(0) &= \sigma_w^2 (1 + \theta^2) \\
\gamma(1) &= \sigma_w^2 \theta \\
\widehat{\rho(1)} &= \frac{\widehat{\gamma(1)}}{\widehat{\gamma(0)}} = \frac{\hat{\theta}}{1 + \hat{\theta}^2}
\end{aligned}
\tag{18}
$$

Two solutions *exist* so the **invertible** one should be chosen

- If $|\rho\hat{(1)}| < 0.5$, solutions are real. else they dont exist

Invertible Estimate $\hat{\theta} = \frac{1-\sqrt{1-4\hat{\rho}(1)^2}}{2\hat{\rho}(1)}$

$\hat{\theta} \sim AN(\theta, \frac{1+\theta^2+4\theta^4+\theta^6+\theta^8}{n(1-\theta^2)^2})$

AN: Asymptotically Normal

(a) Maximum Likelihood and Least Square Estimation

$$\begin{aligned} S(\mu, \phi) =& (1-\phi^2)(x_1-\mu)^2 + \sum_{t=2}^{n}[(x_t-\mu)-\phi(x_{t-1}-\mu)]^2 \\ \hat{\sigma_w^2} =& n^{-1}s(\hat{\mu}, \hat{\phi}) \end{aligned} \tag{19}$$

- $\hat{\mu}$, $\hat{\phi}$ are MLEs of μ, ϕ respectively.
- $S(\phi, \phi)$: unconditional sum of squares
  - This can be minimized to find *unconditional least squares* estimators

Conditional MLE of $\hat{\sigma_w^2} = S_c(\hat{\mu}, \hat{\phi})/(n-1)$
Recall Standard Error of 1-step ahead forecast is $P_t^{t-1} = \gamma(0)\Pi_{j=1}^{t-1}(1-\phi_{jj}^2)$. For ARMA models, $\gamma(0) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2$
Thus, $P_t^{t-1} = \sigma_w^2([\sum_{j=0}^{\infty} \psi_j^2][\Pi_{j=1}^{t-1}(1-\phi_{jj}^2)]) = \sigma_w^2 r_t$

$$\begin{aligned} S(\beta) =& \sum_{t=1}^{n}[\frac{(x_t - x_t^{t-1}(\beta))^2}{r_t(\beta)}] \\ \hat{\sigma_w^2} =& n^{-1}S(\hat{\beta}) \end{aligned} \tag{20}$$

3.5.2 Newton-Raphson & Scoring Algorithms

AR(p) models are linear so traditional Regression works. MA(q) and ARMA(p, q) models are non-linear so numerical methods are required to calculate MLE for β.

Score Vector: $t^{(1)}(\beta) = (\frac{\partial l(\beta)}{\partial \beta_1}, ..., \frac{\partial l(\beta)}{\partial \beta_k})^T$

Hessian: $t^{(2)}(\beta) = (-\frac{\partial t^2(\beta)}{\partial \beta_i \partial \beta_j})_{i,j=1}^{k}$

When using the method of scoring, the *Hessian Matrix* is replaced by the *information matrix* $(E[t^{(2)}(\beta)])$

1. Gauss-Newton

   Conditional Sum of Square Error: $S_c(\beta) = \sum_{t=p+1}^{n} w_t^2(\beta)$

- Minimizing leads to *conditional leaste squares estimate*
- $q > 0$ is a nonlinear regression problem and requires numerical optimization
- Conditioning on a few samples with a large n has little influence on final parameter estimates

Unconditional Sum of Square Error

$$\begin{aligned} S(\beta) =& \sum_{t=-\infty}^{n} \hat{w_t}^2(\beta) \\ \hat{w_t}(\beta) =& E(w_t|x_1, ..., x_n) \end{aligned} \tag{21}$$

- t chosen using $t = -M + 1$ where M is large enough to guarantee $\sum_{t=-\infty}^{-M} \hat{w_t}^2(\beta) \approx 0$
- Numerical optimization needed even when $q = 0$

Gauss-Newton estimation is calculated by:

$\beta_{(j)} = \beta_{j-1} + \Delta(\beta_{j-1})$

- see p124 for more details.

Information Matrix: $\underset{(p+q)\times(p+q)}{\Gamma_{p,q}} = \begin{bmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta} \end{bmatrix}$

AR(p): $\Gamma_{\phi\phi}$ **MA(q)**: $\Gamma_{\theta\theta}$

2. Asymptotic Distribution Examples

   (a) AR(1) $\hat{\phi} \sim AN[\phi,\ n^{-1}(1-\phi^2)]$

   (b) MA(1) $\hat{\theta} = AN[\theta, n^{-1}(1-\theta^2)]$

   (c) ARMA(1,1) $\begin{bmatrix} \hat{\phi} \\ \hat{\theta} \end{bmatrix} \sim AN[\begin{bmatrix} \phi \\ \theta \end{bmatrix}, n^{-1} \begin{bmatrix} (1-\phi^2)^{-1} & (1+\phi\theta)^{-1} \\ sym & (1-\theta^2)^{-1} \end{bmatrix}^{-1}]$

   Fitting an AR(2) model when an AR(1) model is appropriate leads to *overfitting*. Variance is inflated. Though it can be useful to do this for diagnostic purposes.

   Note that MA(1) and AR(1) have similar variances. This is partially due to the fact that MA regressors are the differential processes $z_t(\theta)$ that have AR structure, and it is this structure that determines the asymptotic variance of the estimators.