STAT 5511
Time Series Analysis

Charles R. Doss

University of Minnesota

Fall, 2024

Introduction

- Welcome to STAT 5511, Time Series Analysis
- I'm Charles Doss.
- Course website is on Canvas, https://canvas.umn.edu/, wherein the syllabus, homework assignments, and *raw* grades (but not appropriately weighted course grades!) will be posted.
- Textbook is Shumway and Stoffer ("SS"), 4th edition (available online).
- Let's go over the Syllabus (find it on canvas).
- Class cancelled Nov 27 (Weds before NA Heritage / Thanksgiving holiday)

Structure of the course; Alternative time series texts

The course textbook is Shumway and Stoffer (SS).

- Structure: The class is to a very large degree based off of SS and structured to largely follow it through at least its first three chapters.
- Audience: No one course can appeal to everyone at once. Time series arise in an extremely large variety of disciplines; thus students who may be interested in this class may have a large variety of backgrounds. This course is broadly targeted to advanced STEM undergraduates or STEM master's students (already a broad group).

An alternative textbook: Cryer and Chan, *Time Series Analysis*

Table of Contents I

Table of Contents II

Table of Contents III

Table of Contents I

Table of Contents II

Table of Contents III

Note: The sections (e.g., "§1.1") in the table of contents refer to SS, 4th ed., and let you know roughly what in the textbook corresponds to the current lecture material.

Time Series
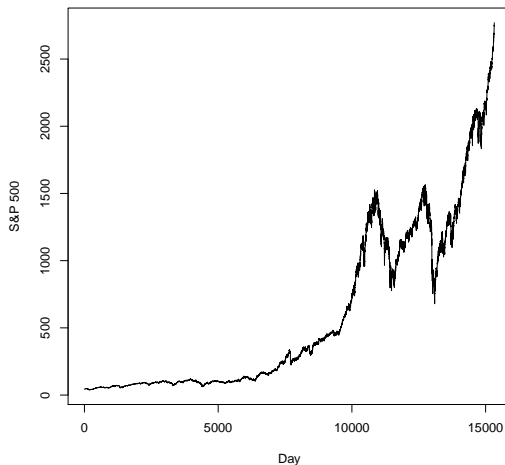
Examples

Time Series

Examples

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Time Series

Examples

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Time Series
Examples

S&P 500 differences

Day ('84–'91)

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.
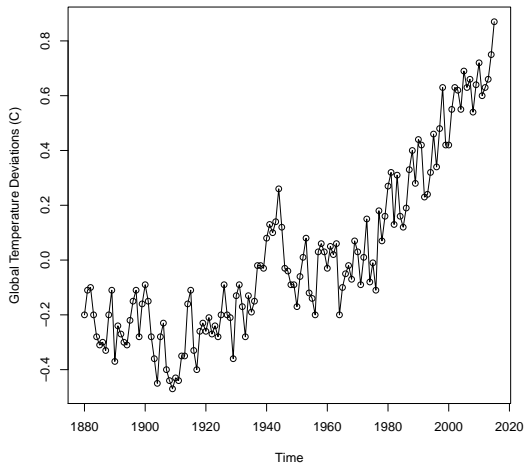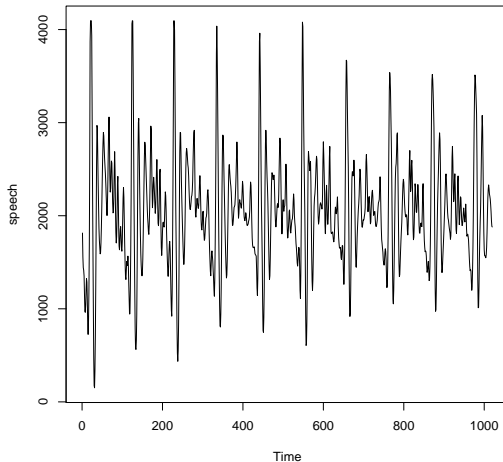
Time Series

Examples
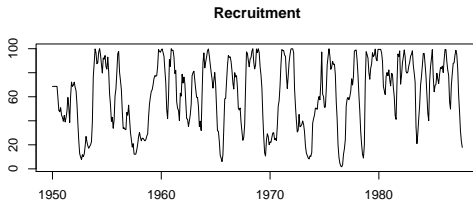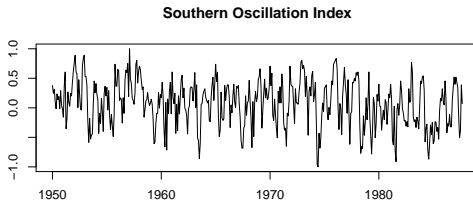
Time series characteristics

(The above data is (mostly) from the astsa package.)

- Not iid (independent and identically distributed).
- Dependence is related to time proximity.
- Require some *stationarity* (not yet defined rigorously!)

Our goals:

- Modeling (descriptive)
- Interpretation
- Estimation
- Forecasting/prediction
- Inference

What we do

Chapter 3: Time domain methods (similar to traditional regression). (AR, MA, AR(I)MA.)

Chapter 4: Spectral domain methods (periodicity).

Table of Contents I

Table of Contents II

Charles R. Doss   Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Table of Contents III

Complex numbers

There are dozens of textbooks one can study complex numbers or analysis from, at different difficulty levels. Here are two:

- "Introductory Complex Analysis", by Silverman
- "Complex Analysis" by Lang. (This one is closer to grad level.)

The facts we need about complex numbers/analysis are mostly only fairly elementary ones, mostly from the very beginning of most complex analysis texts.

Complex numbers

It turns out studying ARIMA models often involves understanding polynomials over the complex numbers, as we will see in not too long. So here we review some basic facts about complex numbers and complex analysis.

If $z \in \mathbb{C}$ that means $z = a + bi$, for some $a, b \in \mathbb{R}$, where $i = \sqrt{-1}$. So, if $z = a + bi$, then:

- Addition and multiplication work basically as you would expect: If $y = c + di$ then $z + y = (a + bi) + (c + di) = (a + c) + (b + d)i$. And $zy = (a + bi)(c + di) = (ac - bd) + (ad + bc)i$.
- Complex conjugate (definition): $\bar{z} = \overline{a + bi} = a - bi$.
- Modulus (definition): $|z|^2 = a^2 + b^2$. Thus, $z\bar{z} = |z|^2$.
- $(a + bi)^{-1}$ is itself a complex number. We can find a formula for it by multiplying / dividing by the complex conjugate.

For a real number r, recall the Geometric Series:
$1 + r + r^2 + \cdots + r^n + \cdots = \sum_{j=0}^{\infty} r^j$. Since for any $n \geq 1$, we have
$(1 + \cdots + r^n)(1 - r) = 1 - r^{n+1}$, we thus have

$$\sum_{j=0}^{\infty} r^j = \frac{1}{1-r} \quad \text{whenever} \quad |r| < 1.$$

The exact same algebra works for complex numbers. If $z = a + bi \in \mathbb{C}$ and
$|z| = \sqrt{a^2 + b^2} < 1$, then $\sum_{j=0}^{\infty} z^j = 1/(1-z)$.

Complex Analysis

Polynomials

The whole point of the number $i = \sqrt{-1}$, and of the complex numbers in general, is that i is a solution to the equation $x^2 = -1$; that is, i is a zero of the polynomial $x^2 + 1$. The Fundamental Theorem of Algebra says that in fact any polynomial $P(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$ always has n complex roots (some may be repeated; here $a_n \neq 0 + 0i$). Here the a_i are real or complex numbers.

This means we can factor the polynomial into its roots:
$P(z) = a_n(z - r_1)(z - r_2) \cdots (z - r_n)$. (Where the r_i are potentially complex numbers.)

No closed form solution for r_i in general if $n > 2$, but for a quadratic polynomial ($n = 2$; $P(z) = az^2 + bz + c$) the quadratic formula gives us the formula: $(-b \pm \sqrt{b^2 - 4ac})/2a$. If the term $b^2 - 4ac$ is negative then there is no real solution; but there are two imaginary ones.

Let $z \in \mathbb{C}$. We have been discussing so far the representation $z = a + bi$, which is the *Cartesian* representation.

There is another representation: the *polar coordinate* representation. We can write $z = re^{i\theta}$ where $r = |z| = \sqrt{a^2 + b^2}$ is the *radius* and θ is the *angle*. An important identity to remember here is Euler's identity:

$$e^{i\theta} = \cos(\theta) + i\sin(\theta).$$

Review

Mathematical statistics review: here are some textbook references that may be of help for mathematical statistics review. In increasing order of complexity:

1. Wackerly, Mendenhall, and Scheaffer, *Mathematical Statistics with Applications*. ($\approx$ STAT4101)
2. Berry and Lindgren, *Statistics Theory and Methods*. ($\approx$ STAT 5101).
3. DeGroot (or DeGroot and Schervish, 3rd edition and later), *Probability and Statistics*. ($\approx$ STAT5101)
4. Casella and Berger, *Statistical Inference*. ($\approx$ STAT8101)

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Math-stat review
Describing time series

A time series is a sequence of random variables, $\{X_t\}$, generally indexed by t; we write $\{x_t\}$ for a *realization* of the series.

Usually for us $t = 1, 2, 3, \ldots$ (meaning, discrete observation at uniform intervals), so the time series are $\{X_1, X_2, X_3, \ldots\}$ or $\{x_1, x_2, x_3, \ldots\}$.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

In general, any random variables (RVs), including time series variables, are described by their *joint cumulative distribution function* (CDF):

$$F(c_1, \ldots, c_T) = P(X_1 \le c_1, \ldots, X_T \le c_T).$$

We can describe individual (marginal) RVs X_t by their (marginal) CDFs,

$$F_t(c_t) = P(X_t \le c_t).$$

Assume X_t are all continuous. Then the joint density $f(c_1, \ldots, c_T)$ equals $\frac{\partial^T}{\partial c_1 \ldots \partial c_T} F(c_1, \ldots, c_T)$ and the marginal densities are $f_t(c_t) = \frac{\partial}{\partial c_t} F_t(c_t)$. Also $F_t(x) = \int_{-\infty}^{x} f_t(z) dz$. Similarly in the multivariate case. E.g., for the bivariate case, $F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(z_1, z_2) dz_2 dz_1$.

- Joint CDF/density: describes dependence behavior; but too complex
- Marginal CDF/density: simpler (e.g. can plot); but doesn't describe dependence.

CDF applies to any random variables; density only to continuous ones.

Definition

The mean of a continuous random variable or time series is

$$\mu_t \equiv \mu_{X_t} := E(X_t) = \int x f_t(x) dx.$$

Definition

The variance of a continuous random variable or time series is

$$\sigma_t^2 \equiv \sigma_{X_t}^2 := E((X_t - \mu_t)^2) = \int (x - \mu_t)^2 f_t(x) dx.$$

If X_t is discrete you sum the PMF instead of integrating in the last equalities.

The mean and variance serve as summaries of a (single, marginal) distribution.

Math-stat review
Covariance and correlation

Definition

The covariance between *two* random variables X and Y is

$$\mathrm{Cov}(X,Y) \equiv \sigma_{X,Y} := E((X - \mu_X)(Y - \mu_Y)).$$

When $\{X_t\}$ is a time series we define the *autocovariance function*

$$\gamma(t,s) \equiv \gamma_X(t,s) := \mathrm{Cov}(X_t, X_s).$$

The correlation between two random variables X and Y is
$\mathrm{Cor}(X,Y) \equiv \rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$. For a time series $\{X_t\}$ We define the
autocorrelation function (ACF),

$$\rho(t,s) \equiv \rho_X(t,s) = \mathrm{Cor}(X_t, X_s) = \frac{\gamma(t,s)}{\sqrt{\gamma(t,t)\gamma(s,s)}}.$$

Note $\gamma(t,s) = \gamma(s,t)$ and $\rho(t,s) = \rho(s,t)$.

The population correlation is normalized so that $-1 \le \rho \le 1$ (and similarly for the sample correlation, $-1 \le r \le 1$, to be defined later).

ρ measures *linear* dependence in the X and Y population.

Similarly, r measures the strength of of the linear relationship between the X's and Y's in the sample.

If $\rho = \pm 1$ then there is a perfect/exact linear relationship in the population; if $r = \pm 1$ then there is a perfect/exact linear relationship in the sample.

Scatterplots of samples of size $n = 100$ from populations with different correlations.

If $\rho = 0$ (equivalently $\text{Cov}(X, Y) = 0$) we say X and Y are uncorrelated. This means that there is no *linear* relationship between X and Y on average.

If X and Y are independent, this means there is *no relationship* between X and Y of any kind; X tells us nothing about Y and vice versa.

Recall: If X and Y are independent, then they are uncorrelated.

But the converse is not true! Variables can be uncorrelated but be dependent.

Here are two scatterplots of two different samples (size $n = 100$) from two different populations (call them "L" and "R").

Both populations have $\rho = 0$. But X and Y are independent in L but not in R. In R, there is a strong (quadratic) relationship between X and Y.
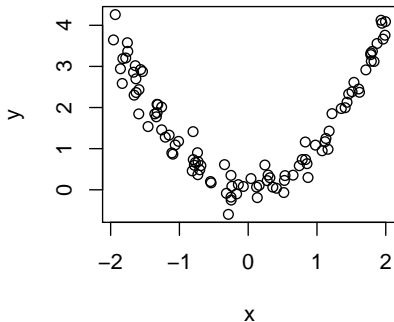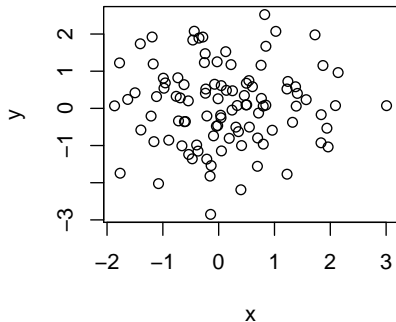
Math-stat review
Mean, variance, and covariance

Capital letters refer to (arbitrary) random variables, lower case to (nonrandom) constants.

Linearity of expectation: $E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i E(X_i)$.

Multilinearity of covariance:

$$\begin{aligned}
\mathrm{Cov}(W + X, Y + Z) = \mathrm{Cov}(W, Y) + \mathrm{Cov}(W, Z) \\
+ \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)
\end{aligned} \tag{1}$$

$$\mathrm{Cov}(aW, bY) = ab\,\mathrm{Cov}(W, Y). \tag{2}$$

Recall (directly from the definitions on the previous slide) that

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X).$$

Thus, by (1),

$$\mathrm{Var}(W + X) = \mathrm{Var}(W) + \mathrm{Var}(X) + 2\,\mathrm{Cov}(W, X) \tag{3}$$

and by (2),

$$\mathrm{Var}(aW) = a^2\,\mathrm{Var}(W).$$

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Capital letters refer to (arbitrary) random variables, lower case to (nonrandom) constants.

The general version of the (co)variance-of-sum formulas (1) and (3) from the previous slide but written out for more variables:

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{m} Y_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j), \qquad \text{so, in particular}$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

We also have the useful formulas for computation:

$$\mathrm{Var}(X) = E(X^2) - (EX)^2 \text{ and } \mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

Finally, if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$ (more on this later).

Math-stat review

Consider a random sample of size n of $(X_1, Y_1), \ldots, (X_n, Y_n)$ that is *independent and identically distributed (i.i.d.)* from some population distribution (e.g., (height, weight) of a person from the Twin Cities).

sample	population
$\overline{X}$	$\mu_x = EX$
$\overline{Y}$	$\mu_y = EY$
$s_x^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$	$\sigma_x^2 = E(X - \mu_x)^2$
$s_y^2 := \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$	$\sigma_y^2 = E(Y - \mu_y)^2$
$s_{x,y} := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$	$\sigma_{x,y} = E(X - \mu_x)(Y - \mu_y)$
$r = \dfrac{s_{x,y}}{s_x s_y}$	$\rho = \dfrac{\sigma_{x,y}}{\sigma_x \sigma_y}$

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) from any distribution with mean μ and variance σ^2. Let $\bar{X}_n := n^{-1} \sum_{i=1}^{n} X_i$.

Law of large numbers: $\bar{X}_n \to_p \mu$ as n gets large.

Central limit theorem: $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ approximately has a $N(0,1)$ distribution when n is large; mathematically, we write

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \to_d N(0,1) \quad \text{as} \quad n \to \infty.$$

Or, put another way,

$$\bar{X}_n \text{ is approximately distributed as } N(\mu, \sigma^2/n).$$

Math-stat review

If a random variable is normally distributed, $X \sim N(m, s^2)$, it has a "bell shaped" density[1].

If a random vector[2] $X = (X_1, \ldots, X_p)'$ is "multivariate normal" that means that every element of X is marginally normal, and it also specifies a density function for the joint distribution that has elliptical contours. The multivariate normal is parameterized by the mean vector $\mu = (E(X_1), \ldots, E(X_p))'$ and the covariance matrix,

$$\Sigma = \text{Cov}(X, X) = E((X - \mu)(X - \mu)').$$

We write $X \sim N_p(\mu, \Sigma)$ (sometimes with or without the p subscript).

---

[1] $f(x) = \exp(-(x - m)^2/(2s^2))/\sqrt{2\pi s^2}$.

[2] Standard convention: vectors are column vectors.

Theorem

If $X \sim N_p(\mu_x, \Sigma_x)$ and $Y = c + GX + VZ$ where $Z \sim N_q(0, Id)$ and is independent of X, and $G \in \mathbb{R}^{p \times q}$ is fixed as is $c \in \mathbb{R}^q$ and $V \in \mathbb{R}^{q \times q}$, then the joint distribution of (X, Y) is

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \right)$$

where

$$\mu_y = c + G\mu_x,$$

$\Sigma_{xx} = \Sigma_x$, $\Sigma_{xy} = \Sigma_x G' = \Sigma'_{yx}$, $\Sigma_{yy} = G\Sigma_x G' + VV'$.

These calculations use the identities $E(GX) = GE(X)$ and $\mathrm{Cov}(AX, BY) = A\,\mathrm{Cov}(X, Y)B'$.

Table of Contents I

Table of Contents II

Table of Contents III

Definition (Strong)

$\{X_t\}$ is *white noise* if (1) $E(X_t) = 0$ for all t and (2) X_t and X_s are independent, for all $t \neq s$.*

Definition (Weak)

$\{X_t\}$ is *white noise* if (1) $E(X_t) = 0$ for all t and (2) $\mathrm{Cor}(X_t, X_s) = 0$ for all $t \neq s$.*

*We also assume $\mathrm{Var}(X_t) = \sigma^2 < \infty$. (This just rules out very 'heavy tailed' noise.)

We write $X_t \sim \mathrm{WN}(0, \sigma^2)$ if $\mathrm{Var}(X_t) = \sigma^2$.

Ex: If $X_t \overset{\text{iid}}{\sim} (0, \sigma^2)$ (meaning, $E(X_t) = 0$, $\mathrm{Var}(X_t) = \sigma^2$) then X_t is white.

Ex: $X_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

But there are non-iid cases, which we will see also.
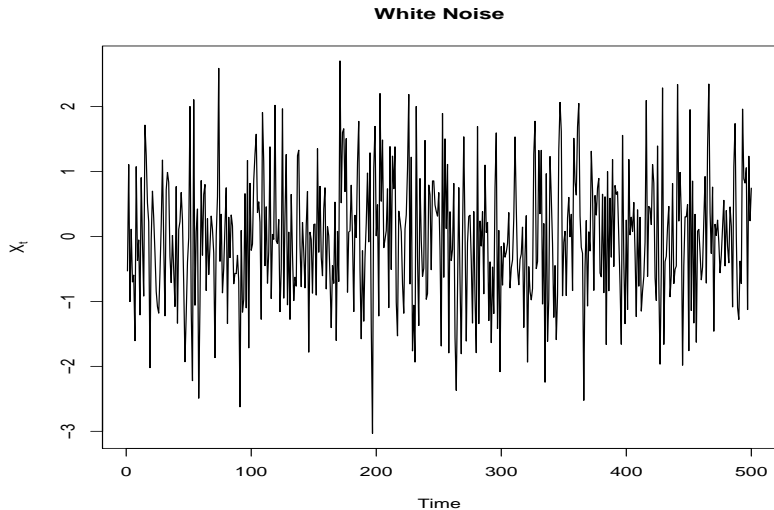
White Noise

The autocovariance function for white noise is

$$\gamma(s,t) = \begin{cases} \sigma^2 & s = t \\ 0 & s \neq t. \end{cases}$$

The ACF is

$$\rho(s,t) = \begin{cases} 1 & s = t \\ 0 & s \neq t. \end{cases}$$
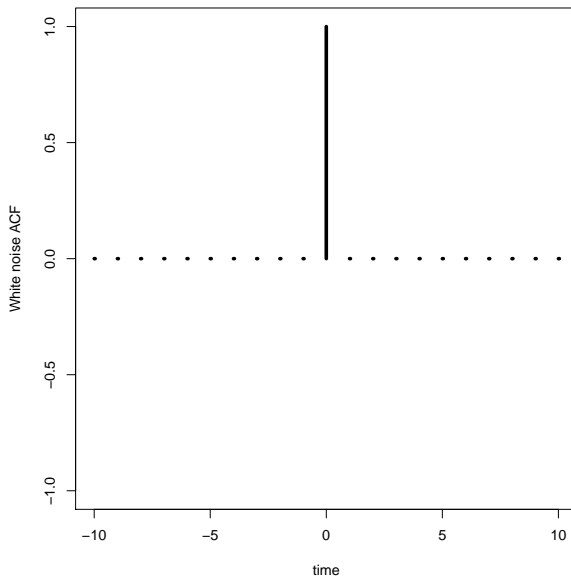
Basic time series models
Moving average

Let $\{W_t\} \sim \text{WN}(0, \sigma^2)$ be white noise. Let

$$V_t = \frac{1}{3}(W_{t-2} + W_{t-1} + W_t).$$

Then $\{V_t\}$ is an example of a *moving average* process.

Moving Average

Consider the moving average process $V_t = \frac{1}{3}(W_{t-2} + \cdots + W_t)$.

The mean function for V_t is $\mu_t = 0$.

The autocovariance function for this moving average process V_t is

$$\gamma(s, t) = \begin{cases} \frac{3}{9}\sigma^2 & s = t \\ \frac{2}{9}\sigma^2 & |s - t| = 1 \\ \frac{1}{9}\sigma^2 & |s - t| = 2 \\ 0 & |s - t| \geq 3. \end{cases}$$

We can then find the ACF.
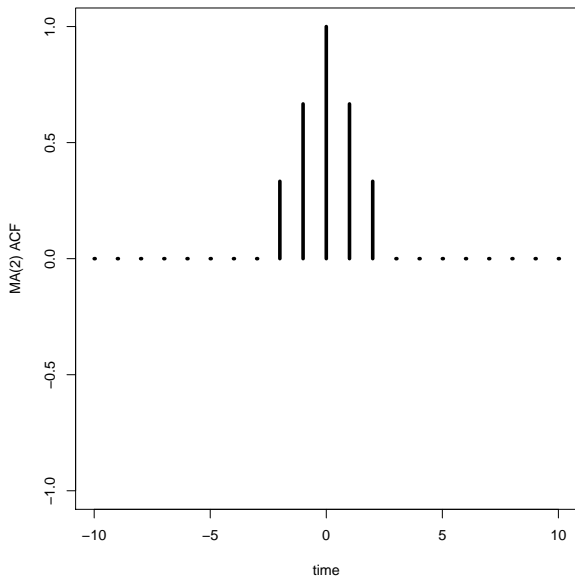
Let $W_t \sim \text{WN}(0, \sigma^2)$. Let

$$X_t = X_{t-1} - .9X_{t-2} + W_t$$

for $t = 0, 2, ...,$ (where we can let X_{-1} and X_{-2} be arbitrary random variables or constants). Then X_t is an example of an AR(2) process.

AR(2)
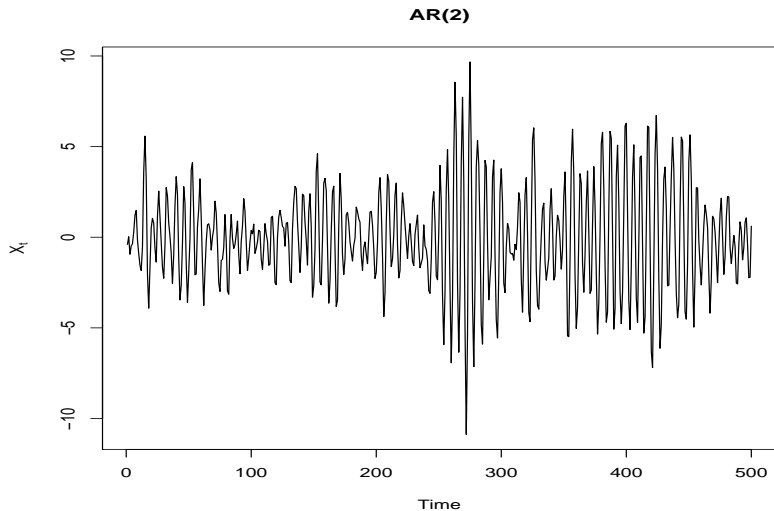
Let $W_t \sim \text{WN}(0, \sigma^2)$. Let

$$X_t = .4X_{t-1} + W_t.$$

Then X_t is an example of an AR(1) process.
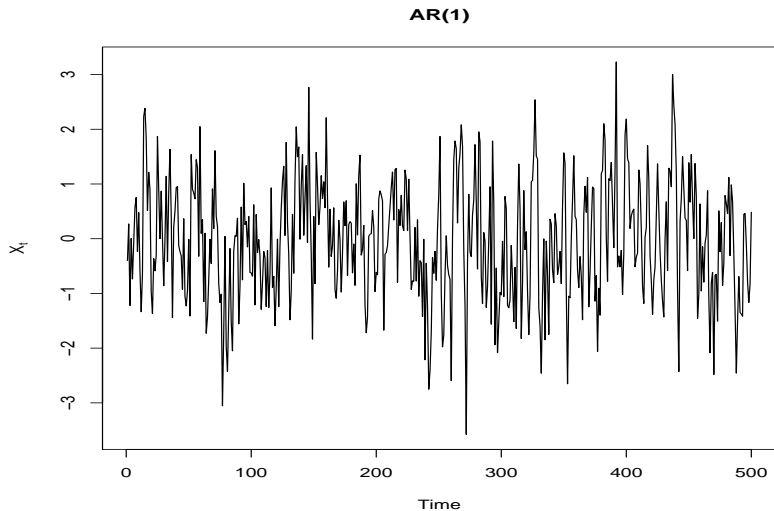
Let $W_t \sim \mathrm{WN}(0, \sigma^2)$. Let

$$X_t = \delta + X_{t-1} + W_t,$$

where $\delta \in \mathbb{R}$ is a constant scalar. (Assume $X_0 = 0$.) Then X_t is a random walk with drift if $\delta \neq 0$. If $\delta = 0$ it is a random walk (without drift).

Note by a recursive argument we can see that
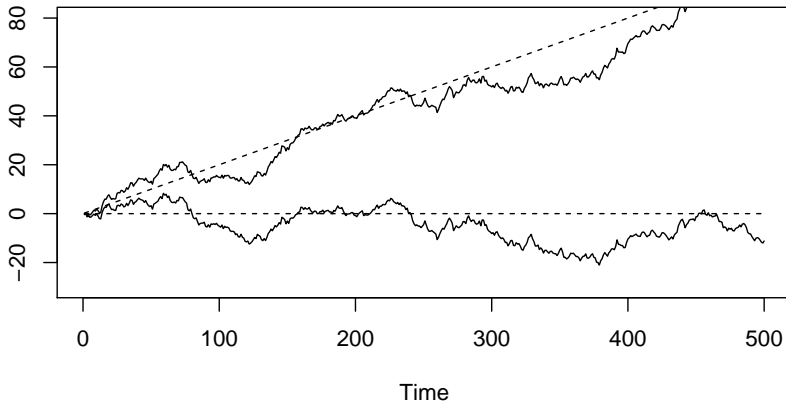
$$X_t = \delta t + \sum_{s=1}^{t} W_s.$$

Random walk with drift

Recall $X_t = \delta + X_{t-1} + W_t = \delta t + \sum_{s=1}^{t} W_s$.

The mean of X_t is: $\mu_t =$

The autocovariance of X_t is:

$$\gamma(t, s) =$$

Then the ACF of X_t is:

$$\rho(t, s) =$$
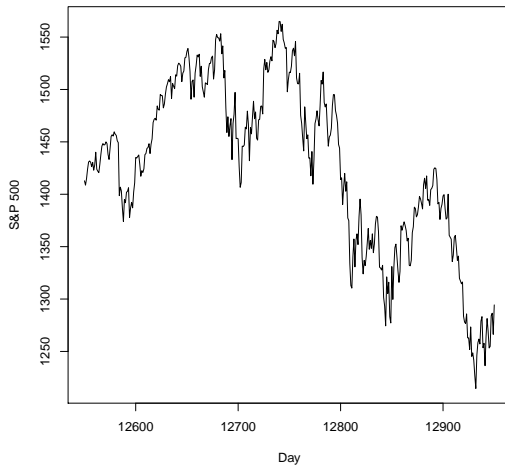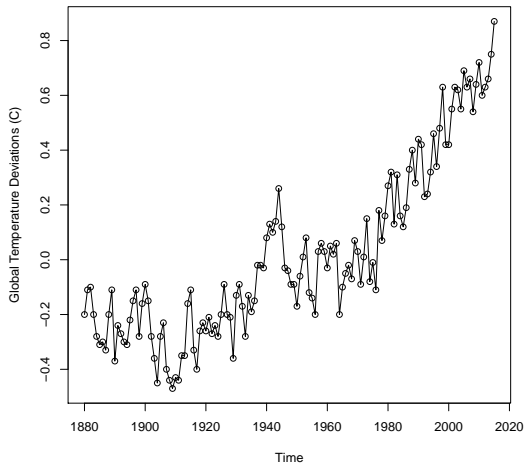
Basic time series models
Signal in noise

Let $W_t \sim$WN$(0, \sigma^2)$. Let

$$X_t = 2\cos(2\pi t/50 + .6\pi) + W_t.$$

Then X_t is an example of a "signal plus noise" (i.e., regression) series.

Consider $X_t = 2\cos(2\pi t/50 + .6\pi) + W_t$.

The mean function for X_t is: $\mu_t =$

The autocovariance function is:

$$\gamma(s,t) =$$

Table of Contents I

Table of Contents II

Table of Contents III

Stationarity

A *strictly (strongly) stationary process*[3] $\{X_t\}$ is one where the joint *distributions* of

$$X_{t_1}, \ldots, X_{t_k} \quad \text{and of} \quad X_{t_1+h}, \ldots, X_{t_k+h}$$

are identical for all time points $t_1, \ldots, t_k$ and any k.

Implications:

- $X_t =_d X_s$ for all s, t.
- $\mu_t = \mu_s = \mu$ is constant in t
- $\sigma_s^2 = \sigma_t^2 = \sigma^2$ is constant in t.
- $(X_1, X_2) =_d (X_7, X_8)$ (Joint distributions equal).
- $\gamma(t_1, t_1 + h) \equiv \text{Cov}(X_{t_1}, X_{t_1+h}) = \text{Cov}(X_{t_2}, X_{t_2+h}) \equiv \gamma(t_2, t_2 + h)$

---

[3]We use the words "process" and "series" interchangeably.

Stationarity

A *weakly stationary process* $\{X_t\}$ is a process where[4]

- $\mu_t = \mu$ is constant in t (does not depend on time)
- the autocovariance function $\gamma(s, t)$ (and ACF $\rho(s, t)$) depends on s and t only through $|s - t|$.

We default to weak stationarity, so in this class *"stationary" means weakly stationary*.

---

[4] $\{X_t\}$ must have a finite variance, $\mathrm{Var}(X_t) < \infty$.

Stationarity

Notational change: For a stationary processes, we write μ for μ_t. And we write

$$\gamma(h) = \gamma(0, h) = \gamma(t, t + h)$$
$$\rho(h) = \rho(0, h) = \rho(t, t + h).$$

Note $\gamma(h) = \gamma(-h)$ and $\rho(h) = \rho(-h)$. We have $\rho(h) = \gamma(h)/\gamma(0)$.

White Noise:

Moving Average:

Random walk with/without drift:

Signal in noise:

Linear processes

Definition

A *linear process* $\{X_t\}$ is given by a (potentially infinite) linear combination of white noise variables $W_t \sim \mathsf{WN}(0, \sigma^2)$:

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j} \quad \text{where} \quad \sum_{-\infty}^{\infty} |\psi_j| < \infty,$$

and $\mu, \psi_j \in \mathbb{R}$ are nonrandom constants.

Linear processes are automatically stationary (by their definitions).

We may see that for a linear process the stationary autocovariance function is

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

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Gaussian process

A process $\{X_t\}$ is a *Gaussian (Normal) series* if all the joint distributions of $X_{t_1}, \ldots, X_{t_k}$ (for any k) are jointly Normal.

For Gaussian processes, weak and strong stationarity coincide.

Table of Contents I

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Table of Contents II

Table of Contents III

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Estimation

If $\{X_t\}$ is assumed stationary, then μ, $\gamma(h)$, $\rho(h)$ don't depend on t. Then based on $X_1, \ldots, X_n$, we can estimate μ by $\bar{X}_n = \frac{1}{n} \sum_{t=1}^{n} X_i$. We have

- $E(\bar{X}_n) = \mu$,
- $\text{Var}(\bar{X}_n) = \frac{1}{n} \sum_{h=-n}^{n} (1 - \frac{|h|}{n}) \gamma(h)$.

What if in fact $\{X_t\} \overset{\text{iid}}{\sim} (\mu, \sigma^2)$?

To estimate $\gamma(h)$ we use

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

for $h = 0, \ldots, n-1$, and let $\hat{\gamma}(-h) = \hat{\gamma}(h)$.

We estimate $\rho(h) = \gamma(h)/\gamma(0)$ by

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

If $\{X_t\}$ is iid white noise then[5] $\widehat{\rho}(h) \overset{approx}{\sim} N(0, n^{-1})$ as n gets large for $h \neq 0$.

In fact, if $\{X_t\}$ is a stationary linear process, $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$ ($\sum |\psi_j| < \infty$) where W_t are iid white noise[6], then letting $\hat{\boldsymbol{\rho}}(H) := (\hat{\rho}(1), \ldots, \hat{\rho}(H))'$ and $\boldsymbol{\rho}(H) = (\rho(1), \ldots, \rho(H))$, we have that

$$\hat{\boldsymbol{\rho}}(H) \overset{approx}{\sim} N(\boldsymbol{\rho}(H), W/n)$$

where $W = (w_{ij})$ and $w_{ij} = \sum_{k=-\infty}^{\infty} (\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) - 2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i)$.

---

[5]Where X_t has finite fourth moment

[6]Where W_t has finite fourth moment

Estimation

Thus to test if there is correlation at, say, lag $h = 3$, i.e.

$$H_0 : \rho(3) = 0 \quad \text{vs} \quad H_1 : \rho(3) \neq 0,$$

we can use $\widehat{\rho}(3)$ as our test statistic and we reject the null if $|\widehat{\rho}(3)| \geq z_{\alpha/2}/\sqrt{n}$. (And similarly for any other $h \neq 0$.)

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.
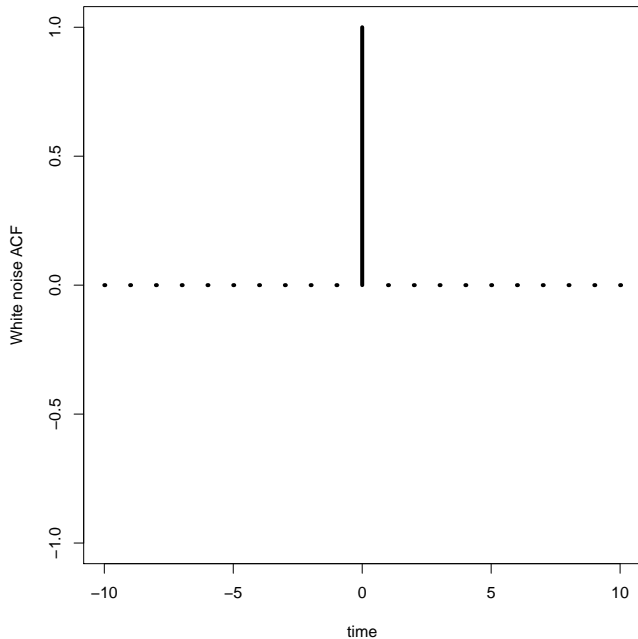
Vector-valued and multivariate series

Vector series are those indexed by t with multiple responses measured (at each t): $\boldsymbol{x}_t := (\boldsymbol{x}_{1,t}, \ldots, \boldsymbol{x}_{p,t})'$ (some p), $t \in \{1, 2 \ldots, \}$.[7]

Multivariate/multidimensional series are those indexed by multiple "time" indices (e.g., referring to time and space, or just to 2D or 3D space): $x_{\boldsymbol{s}} \in \mathbb{R}$ where $\boldsymbol{s} = (s_1, \ldots, s_r)$.

---

[7]We generally let vectors $\boldsymbol{x}$ be column vectors; we let $\boldsymbol{x}'$ denote the transpose of $\boldsymbol{x}$.

Multivariate series

Example: Soil temperature

Southern Oscillation Index

Recruitment
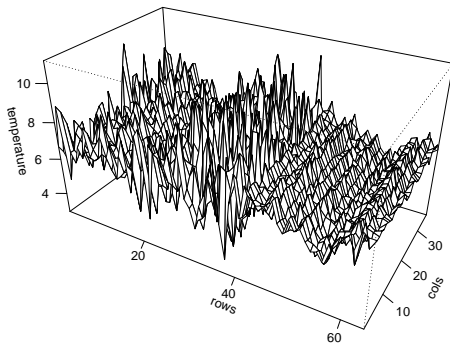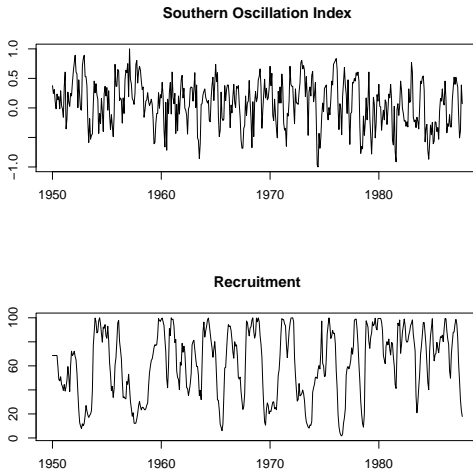
Vector-valued series

For a stationary p-dimensional vector-valued series, we let

$$E\boldsymbol{X}_t = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$$

and

$$\Gamma(h) = E\left((\boldsymbol{X}_{t+h} - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})'\right) \quad \text{for any } t,$$

be the (cross-) "autocovariance" matrix with (i,j)th entry given by

$$\gamma_{i,j}(h) = E((X_{t+h,i} - \mu_i)(X_{t,j} - \mu_j)), \text{ for } i,j = 1, \ldots, p.$$

- $\Gamma(-h) = \Gamma(h)'$ (because $\gamma_{i,j}(h) = \gamma_{j,i}(-h)$).
- Except for $\Gamma(0)$, $\Gamma(h)$ $(h \neq 0)$ is not generally a symmetric matrix.

The *cross-correlation* is

$$\rho_{i,j}(h) = \frac{\gamma_{i,j}(h)}{\sqrt{\gamma_{i,i}(0)\gamma_{j,j}(0)}}$$

Vector-valued series

We can estimate $\boldsymbol{\mu}$ by

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}_n = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{X}_t$$

and $\Gamma(h)$ by

$$\hat{\Gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (\boldsymbol{X}_{t+h} - \bar{\boldsymbol{X}}_n)(\boldsymbol{X}_t - \bar{\boldsymbol{X}}_n)' = \left(\widehat{\gamma}_{i,j}(h) \right)_{i,j=1}^{p}$$

for $0 \leq h < n$.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

The sample cross-correlation function is

$$\widehat{\rho}_{i,j}(h) := \frac{\widehat{\gamma}_{i,j}(h)}{\sqrt{\widehat{\gamma}_{i,i}(0)\widehat{\gamma}_{j,j}(0)}}.$$

Theorem

If one of $\{X_{i,t}\}$ or $\{X_{j,t}\}$ is iid white noise, and the other is a linear process based on white noise series that is iid and independent of the first then $\widehat{\rho}_{i,j}(h)$ is approximately $N(0, n^{-1})$ for large n.
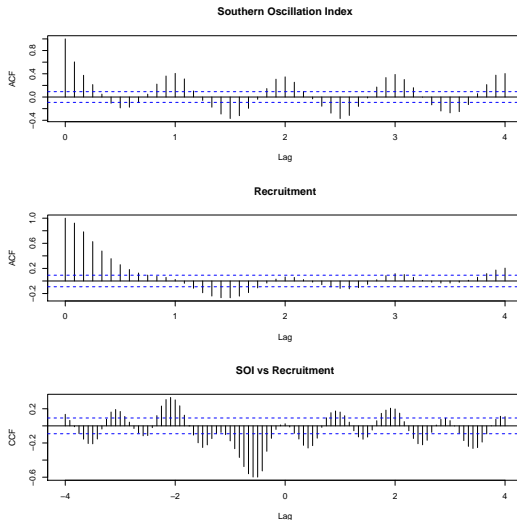
Vector-valued series

```
par(mfrow=c(3,1))
acf(soi, 48, main="Southern Oscillation Index")
acf(rec, 48, main="Recruitment")
ccf(soi, rec, 48, main= "SOI vs Recruitment", ylab="CCF")
```

Table of Contents I

Table of Contents II

Table of Contents III

Time series regression

The time series linear regression model:

$$X_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_q z_{tq} + W_t$$
$$= \beta' z_t + W_t$$

where $\{W_t\} \sim (0, \sigma^2)$ is

- iid, or
- white noise, or
- stationary.

We assume (for now) that $z_{t,i}$ are all fixed numbers.

We let $\beta' = (\beta_0, \beta_1, \ldots, \beta_q)$ and $z_t = (1, z_{t1}, \ldots, z_{tq})$. (And/so $z_{t0} = 1$.)

Time series regression

If W_t are iid we do estimation by ordinary least squares (OLS) which minimizes the sum of squares,

$$\widehat{\beta} := \operatorname*{argmin}_{\beta \in \mathbb{R}^{q+1}} \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \beta' z_i \right)^2.$$

Differentiating (w.r.t. β_i, $i = 0, \ldots, q$) and setting to 0 shows that the solution $\widehat{\beta}$ satisfies the normal equations

$$\sum_{i=1}^{n} z_i \left(X_i - z_i' \widehat{\beta} \right) = 0_{q+1}, \quad \text{so} \quad \left(\sum_{i=1}^{n} z_i z_i' \right) \widehat{\beta} = \sum_{i=1}^{n} X_i z_i,$$

or, assuming $\left(\sum_{i=1}^{n} z_i z_i' \right)$ (a $(q+1) \times (q+1)$ matrix) is invertible,

$$\widehat{\beta} = \left(\sum_{i=1}^{n} z_i z_i' \right)^{-1} \sum_{i=1}^{n} X_i z_i.$$

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Time series regression

Example: chicken price
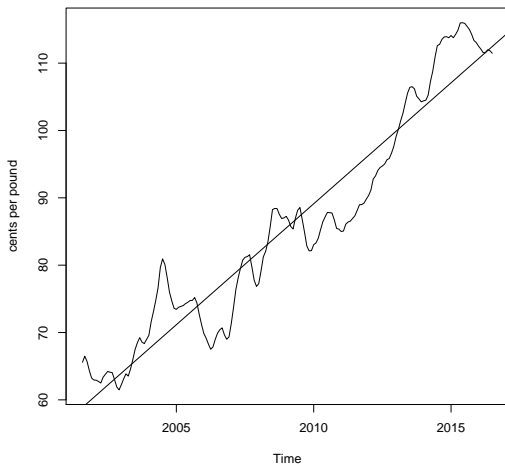
Regression

When $W_i \overset{\text{iid}}{\sim} (0, \sigma^2)$, the covariance matrix of $\widehat{\beta}$ is

$$\text{Cov}(\widehat{\beta}) = \sigma^2 (\sum_{i=1}^{n} z_i z_i')^{-1}.$$

(This formula would *not* generally hold if the W_i were a non-i.i.d. time series.) The diagonal entries of this matrix let us form CI's or do hypothesis tests for β_i (i.e., $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$).

Regression

$(W_i \overset{\text{iid}}{\sim} (0, \sigma^2)$ still.) We need to know σ^2.

Let $SSE = \sum_{i=1}^{n}(X_i - \widehat{\beta}'z_i)^2$ be the sum of squared (estimated) residuals. Here $X_i - \widehat{\beta}'z_i$ is an estimated residual, sometimes just referred to as "a residual."

We can estimate σ^2 using

$$MSE = \frac{SSE}{n - (q+1)} = \frac{\sum_{i=1}^{n}(X_i - \widehat{\beta}'z_i)^2}{n - (q+1)}$$

In a linear regression, we often want to test hypotheses about the coefficients, for instance to determine which predictors are relevant, i.e.

$$H_0 : \beta_{r+1} = \ldots = \beta_q = 0 \tag{4}$$

When the W_i's are i.i.d., we can test the null hypothesis (5) against the full model by using the traditional *F-statistic*. It is based on estimating under the null hypothesis model (where (5) holds) and under the full model, and comparing.

Recall $W_i \overset{\text{iid}}{\sim}$, and we want to test the null:

$$H_0 : \beta_{r+1} = \ldots = \beta_q = 0. \tag{5}$$

Let $\widehat{\beta}^{(r)}$ be the OLS estimates for $\beta_0, \ldots, \beta_r$ (leaving out $\beta_{r+1}, \ldots, \beta_q$).
Let $SSE_r := \sum_{i=1}^{n}(X_i - z_i'\widehat{\beta}^{(r)})^2$. Then the F-statistic for testing (5) is

$$F = \frac{(SSE_r - SSE)/(q-r)}{SSE/(n-(q+1))} =: \frac{MSR}{MSE}.$$

Under H_0, $F \sim F_{q-r, n-(q+1)}$. So if F is bigger than a $(1-\alpha)$ critical value of $F_{q-r, n-(q+1)}$ we reject H_0.

Let $\widehat{\sigma}_k^2 = SSE_k/n$.

AIC:

$$\log \widehat{\sigma}_k^2 + \frac{n + 2k}{n}.$$

BIC:

$$\log \widehat{\sigma}_k^2 + \frac{k \log n}{n}.$$

For a series $\{X_t\}$ we may often believe that

$$X_t = \mu_t + Y_t$$

where μ_t is a deterministic trend and Y_t is stationary. We want to remove the trend μ_t (detrend) to get to the stationary process Y_t.

For example we may model $\mu_t = \beta_0 + \beta_1 t$. We can regress X_t on time to get estimates of β_i and then let

$$\widehat{Y}_t = X_t - \widehat{\beta}_0 - \widehat{\beta}_1 t.$$

```
plot(resid(fit_chicken), ylab="cents per pound")
```
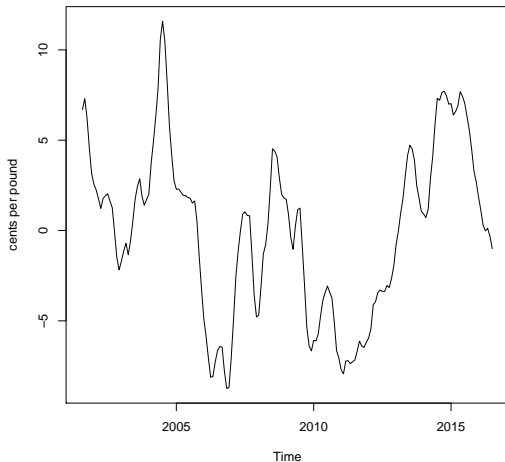
For a series $\{X_t\}$ we may often believe that

$$X_t = \mu_t + Y_t.$$

To detrend we can *difference*:

$$\nabla X_t = X_t - X_{t-1}.$$

This is a *nonparametric* approach to detrending.
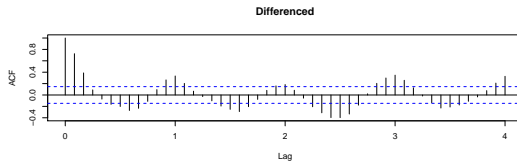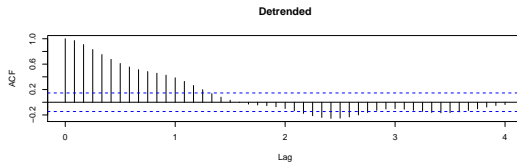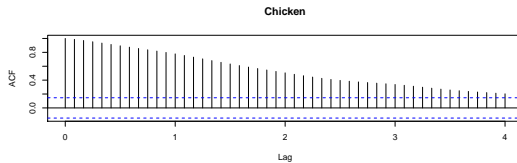
Time series regression
Scatterplot matrices (SOI auto-lags)

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.
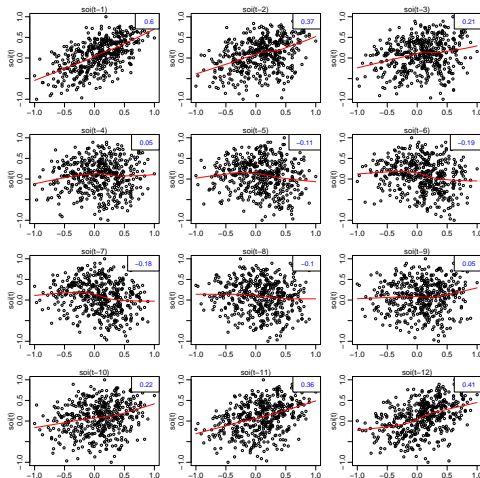
Time series regression
Scatterplot matrices (SOI & Recruitment cross)
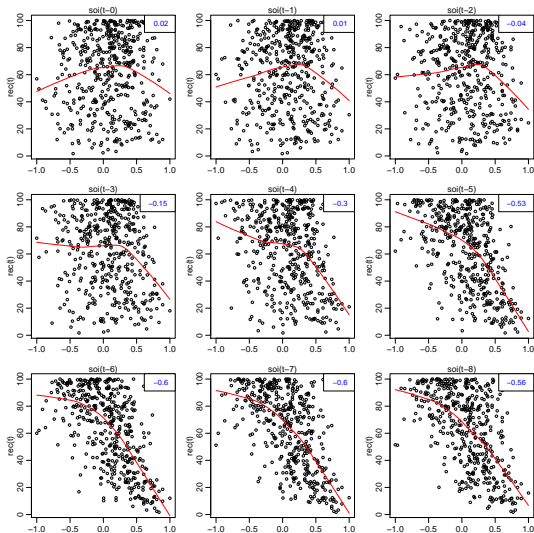
Time series regression

Smoothing

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Table of Contents I

Table of Contents II

Table of Contents III

Table of Contents I

Table of Contents II

Table of Contents III

ARIMA models

In chapter 3 of SS, we study so-called *ARIMA*(p, d, q) models.

1. the p is the autoregressive order (AR),
2. the q is the moving average order (MA),
3. the d is the "integrated" order (I).

Section 3.1 introduces AR, then MA, and then combines them to get ARMA models. (The "I" comes later.)

AR model

Definition

We say a mean 0 process $\{X_t\}$ is an *autoregressive process of order* p, or AR(p), if
$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + W_t$$
for some white noise process $\{W_t\}$. We assume $\phi_p \neq 0$, and for now center X_t so that $EX_t = 0$.

Definition

When talking about autoregressive processes it is useful to use the *backshift operator*,
$$BX_t := X_{t-1} \quad \text{and} \quad B^q X_t := X_{t-q}.$$

Note we can relate the difference operator, $\nabla X_t = X_t - X_{t-1}$ to the backshift operator: $\nabla = (1 - B)$. And $\nabla^k = (1 - B)^k$.

AR model

Definition

Given autoregressive coefficients $(\phi_1, \ldots, \phi_p)$, The *autoregressive operator* is

$$\phi(B) := 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

and we write

$$\phi(B) X_t = W_t.$$

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

The AR(1) model:

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

with $W_t \sim \mathrm{WN}(0, \sigma^2)$. Assume $|\phi| < 1$. (When $|\phi| \geq 1$, AR(1)'s have "explosive" behavior; we will return to this.)

We can check by recursive substitution that

$$X_t = \sum_{s=0}^{\infty} \phi^s W_{t-s},$$

so X_t is a linear process.[8]

We can use this representation for understanding properties of the AR(1) model.

---

[8]Here, t goes from $-\infty$ to ∞ in order for the process to be stationary.

With $|\phi| < 1$, we checked our AR(1) can be written as:

$$X_t = \sum_{s=0}^{\infty} \phi^s W_{t-s}.$$

This automatically shows that (when $|\phi| < 1$) the process is stationary.
We can also find the mean: $EX_t = 0$, and the ACF:

$$\gamma(h) = \phi^{|h|}\gamma(0),$$

and so

$$\rho(h) = \phi^{|h|}.$$

AR(1)

Consider an AR(1) model, $X_t = \phi_1 X_{t-1} + W_t$. Assume $|\phi_1| < 1$. Then we can use $\phi(B) = 1 - \phi_1 B$ to rederive the formula $X_t = \sum_{s=0}^{\infty} \phi_1^s W_{t-s}$.

Explosive AR(1) models and causality

Consider $X_t = X_{t-1} + W_t$. Is this process stationary?

Consider $X_t = \phi X_{t-1} + W_t$ with $|\phi| > 1$.

Causality

Definition

A linear process X_t is *causal* if it can be written as

$$X_t = \sum_{s=0}^{\infty} \psi_s W_{t-s}$$

for a white noise sequence W_t and numbers ψ_s satisfying $\sum_{s=0}^{\infty} |\psi_s| < \infty$.

Table of Contents I

Table of Contents II

Table of Contents III

Moving average (MA) model

Definition

We say a process $\{X_t\}$ is a mean 0 *moving average of order q* (MA(q)) if

$$X_t = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \cdots + \theta_q W_{t-q}$$

for some (generally unknown) parameter vector $\theta_1, \ldots, \theta_q$, a white noise process $W_t \sim \mathrm{WN}(0, \sigma^2)$, and where $\theta_q \neq 0$.

Note: by rescaling, we can change the variance of the white noise series $\{W_t\}$ to compensate for a coefficient in front of the individual term W_t.

We assume $\theta_q \neq 0$ so that the moving average is really of order q.

Some definitions change the signs of θ_i (replace θ_i by $-\theta_i$).

MA processes are linear, stationary, and causal for any parameter values, automatically.

MA operator

Definition

The *moving average operator* is

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q.$$

MA(1)

Consider $X_t = W_t + \theta W_{t-1}$, any $\theta \in \mathbb{R}$, where $\{W_t\} \sim \mathrm{WN}(0, \sigma^2)$.

It has $EX_t = 0$.

And

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

so

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

MA(1) sample paths
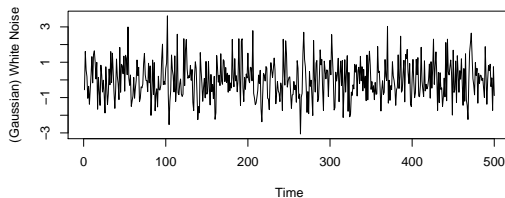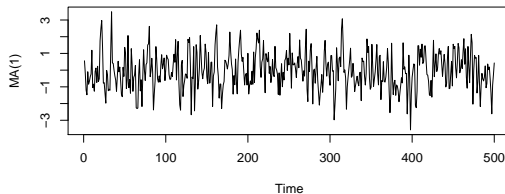
MA(1) sample paths
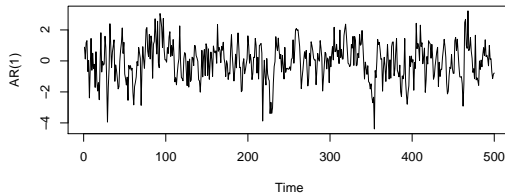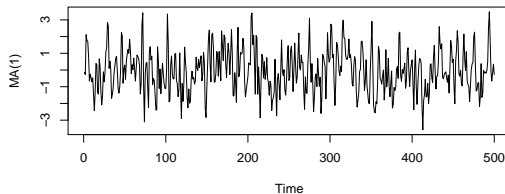
Non-uniqueness of MA models and invertibility

MA models as presently specified are not always identifiable (multiple parameter choices lead to the same probabilistic model).

For instance: in a MA(1) model, both $\theta = \theta_1$ and $\theta = 1/\theta_1$ have the same ACF (although $\gamma(h)$ is different).

One can then find a σ^2_{new} so that (θ_1, σ^2) and $(1/\theta_1, \sigma^2_{new})$ have the same autocovariance function (so same distributions if W_t is Gaussian).

Invertibility

Definition

An MA process $X_t = \theta(B)W_t$ is said to be invertible if there are constants π_j, where $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and we can write

$$\sum_{j=0}^{\infty} \pi_j X_{t-j} = W_t.$$

For an MA(1), if $|\theta| < 1$ then we can check by substitution or because $1/(1 + \theta B) = \sum_{i=0}^{\infty} (-\theta)^i B^i$ that

$$W_t = \sum_{i=0}^{\infty} (-\theta)^i X_{t-i}.$$

It turns out that requiring invertibility makes MA's unique. When we restrict to the model of invertible MA(q) time series, each parameter value is unique.

Charles R. Doss   Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Infinite stochastic sums
Mathematical details

Assume $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ for $\psi_j \in \mathbb{R}$. Let $\{X_t\}$ be *any* sequence of random variables such that $\sup_t E|X_t| < \infty$. Then the series

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

converges absolutely with probability 1^9. If $\sup_t E|X_t|^2 < \infty$ then the series converges in mean square to the same limit.

If $\{X_t\}$ is a stationary process with autocovariance function γ, then the limit process is stationary and has autocovariance function

$$\gamma_\psi(h) = \sum_{j,k=-\infty}^{\infty} \psi_j \psi_k \gamma(h - j + k).$$

---

[9]And, letting $Y_t = \psi(B)X_t$, we have $\sup_t E|Y_t| < \infty$.

Table of Contents I

Table of Contents II

Table of Contents III

ARMA(p,q) models

ARMA(p, q) models combine the AR(p) and MA(q) models. The general form is

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q},$$

where $\{W_t\} \sim \text{WN}(0, \sigma^2)$.

Using the AR and MA operators from before, we can write this as

$$\phi(B)X_t = \theta(B)W_t.$$

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Parameter redundancy

ARMA(p, q) setup: $\phi(B)X_t = \theta(B)W_t$.

As introduced so far, ARMA models have a problem with parameter redundancy or identifiability. (And it this is different than the issue we saw previously with MA models.) To see this we can start with $X_t = W_t$ where $W_t \sim \text{WN}(0, \sigma^2)$. So our observed data is white noise.

We can write $(1 - .5B)X_t = (1 - .5B)W_t$; or, in terms of a more standard model, we might write

$$X_t = .5X_{t-1} + W_t - .5W_{t-1}.$$

Is this ARMA$(1, 1)$ or ARMA$(0, 0)$? See SS, pg 84–85, to see how this concern would be problematic for estimation or forecasting: if we use a model that's too large we can get spurious results.

AR and MA polynomials

Definition

Given an ARMA(p, q) series with parameters $(\phi_1, \ldots, \phi_p)$ and $(\theta_1, \ldots, \theta_q)$, we define the *AR polynomial* and *MA polynomial* to be

$$\phi(z) := 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p$$
$$\theta(z) := 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q.$$

Here, $z \in \mathbb{C}$ is a complex number and $\phi(z)$ and $\theta(z)$ are complex polynomials.

Complex polynomials

A complex number $z \in \mathbb{C}$ can be written as $z = a + bi$ where $i = \sqrt{-1}$. The modulus of z is $|z| = \sqrt{a^2 + b^2}$.

Recall: the Fundamental Theorem of Algebra says that any polynomial $\eta(z) = a_0 + a_1 z + \cdots + a_d z^d$, where a_j are any complex coefficients (including of course real numbers), can be factored into

$$\eta(z) = c(z - z_1) \cdots (z - z_d)$$

where $c, z_1, \ldots, z_d$ are complex numbers.[10] Then $\eta(z_j) = 0$.

In our case, we will often have polynomials where a_0 is 1. In that case, we can write the factorization as

$$\eta(z) = (1 - y_1 z) \cdots (1 - y_d z)$$

for some $y_j \in \mathbb{C}$. In this form the roots are $1/y_j$.

---

[10] The z_j may not be unique, i.e. there may be repeated roots.

ARMA(p, q)

Definition

We say a time series $\{X_t\}$ is a mean-0 ARMA(p, q) series if

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q},$$

where $\phi(z)$ and $\theta(z)$ *have no common complex roots* (no common polynomial factors). A time series $\{Y_t\}$ is mean-μ ARMA(p, q) if

$$Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = \beta_0 + W_t + \theta_1 W_{t-1} + \cdots + \theta_q W_{t-q},$$

where $\beta_0 = (1 - \sum_{i=1}^{p} \phi_i)\mu$.

If $\phi(z)$ and $\theta(z)$ do have common polynomial factors, they should be factored out to yield simplified polynomials $\tilde{\phi}(z)$ and $\tilde{\theta}(z)$. Then the ARMA process can be represented in simplified form as
$\tilde{\phi}(B)X_t = \tilde{\theta}(B)W_t$.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

ARMA(p,q)

Example

Consider the model

$$X_t = .4X_{t-1} + .45X_{t-2} + W_t + W_{t-1} + .25W_{t-2}$$

with $W_t \sim \mathrm{WN}$. What is its ARMA order (p, q)?

Causality

Causality for an ARMA process X_t is just like for an AR process:

Definition

An ARMA process $\{X_t\}$ is causal if $X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}$ for some ψ_j's with $\sum |\psi_j| < \infty$.

Theorem

An ARMA(p, q) process $\phi(B)X_t = \theta(B)W_t$ is causal if and only if $\phi(z)$ has no zeros z with $|z| \le 1$ (i.e., has all zeros z with $|z| > 1$).
If the process is causal, then $X_t = \psi(B)W_t$ where
$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$ *for $|z| \le 1$.*

Causality

Example

Assume

$$X_t = .9X_{t-1} + W_t + .5W_{t-1}$$

Is $\{X_t\}$ causal?

Invertibility

Invertibility for an ARMA process is just like for an MA process:

Definition

An ARMA process $\{X_t\}$ is invertible if $W_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$ for some π_j's with $\sum_{j=0}^{\infty} |\pi_j| < \infty$.

Theorem

An ARMA(p,q) process $\phi(B)X_t = \theta(B)W_t$ is invertible if and only if $\theta(z)$ has no zeros in the unit circle (has all zeros strictly outside the unit circle). If the process is invertible then we can write $W_t = \pi(B)X_t$ where $\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \phi(z)/\theta(z)$ for $|z| \leq 1$.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Invertibility

Example

Assume

$$X_t = .9X_{t-1} + W_t + .5W_{t-1}$$

Is the process invertible?

AR(2)

The polynomial $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$ has roots

$$\frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}.$$

If these are larger in absolute value than 1 then the corresponding AR(2) process is causal.

This is equivalent to

$$\phi_1 + \phi_2 < 1, \quad \phi_2 - \phi_1 < 1 \quad \text{and} \quad |\phi_2| < 1.$$

Example:

$$X_t - \frac{9}{4}X_{t-1} - \frac{9}{4}X_{t-2} = W_t - 3W_{t-1} + \frac{1}{9}W_{t-2} - \frac{1}{3}W_{t-3}$$

where $W_t \overset{\text{iid}}{\sim} N(0,1)$.

- Find the roots of the AR and MA polynomials.
- Identify any parameter redundancy: find the values of p and q for which the model is ARMA(p, q) and write the model in its correct (non-redundant) form.
- Determine whether the model is causal, and
- determine whether it is invertible.

Theoretical analysis of a given ARMA series

Compute the first five coefficients $\psi_0, \ldots, \psi_4$ in the causal linear process representation $X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}$.

Solution:

$\psi_0 = 1$

$0 = \psi_1 + (3/4)\psi_0 = \psi_1 + 3/4 \quad \Rightarrow \psi_1 = -3/4$

$1/9 = \psi_2 + (3/4)\psi_1 = \psi_2 - (3/4)^2 \quad \Rightarrow \psi_2 = 1/9 + (3/4)^2 = 97/144$

$0 = \psi_3 + (3/4)\psi_2 = \psi_3 + (3/4)(97/144)...$

Then we can see that for any $j \geq 3$, $\psi_j = (-3/4)\psi_{j-1}$. Thus for $j \geq 3$, we have $\psi_j = (-3/4)^{j-2}(97/144)$.

In particular, for the first 5 ($j = 0$–4) coefficients, we have $\psi_0 = 1$, $\psi_1 = -\frac{3}{4}$, $\psi_2 = \frac{97}{144}$, $\psi_3 = -\frac{291}{576} = -\frac{97}{192}$ and $\psi_4 = \frac{873}{2304} = \frac{97}{256}$.

Now find the $\gamma(h)$ and $\rho(h)$ for $\{X_t\}$.

Recall the ACF of an AR(1) satisfies $\rho(h) = \phi\rho(h-1)$. If we didn't know the form of ρ, we could use this equation to solve for it: $\rho(0) = 1$ so $\rho(h) = 1 \cdot \phi^h$ for $h \geq 0$.

Now assume $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + W_t$ is AR(2) and assume the process is causal.

Table of Contents I

Table of Contents II

Table of Contents III

Homogeneous linear difference equations

Assume $\{\rho_h\}$ satisfies the difference equation:

$$\rho_h - \alpha_1 \rho_{h-1} - \alpha_2 \rho_{h-2} - \cdots - \alpha_p \rho_{h-p} = 0.$$

The solution depends on the *associated* (or *auxiliary*) *polynomial*: $\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p$. We also write $\alpha(B)\rho_h = 0$.

The solution depends on the roots of $\alpha(z)$. If z_j is a root of $\alpha(z)$ with multiplicity m_j, for $j = 1, \ldots, r$, then one can check algebraically that the solution has the form

$$\rho_h = z_1^{-h} P_1(h) + z_2^{-h} P_2(h) + \cdots + z_r^{-h} P_r(h),$$

where $P_j(h)$, $j = 1, \ldots, r$, is a polynomial in h of degree $m_j - 1$. The coefficients in P_j depend on the z_j as well as the p initial conditions $\rho_0, \ldots, \rho_{p-1}$.

AR(2)

For a causal AR(2) process X_t, recall we derived

$$\rho(h) = \phi_1 \rho(h-1) + \phi_2 \rho(h-2) \quad \text{for all } h \geq 1,$$

with initial condition $\rho(0) = 1$ and $\rho(-1) = \phi_1/(1-\phi_2)$.

Case: $z_1 = z_2$ **both real.** Then, the solution is $z_1^{-h}(c_1 + c_2 h)$ where $1 = c_1$ and $\phi_1/(1-\phi_2) = (c_1 + c_2)z_1^{-1}$. And So $c_2 = z_1\phi_1/(1-\phi_2) - 1$.

Second case: $z_1 \neq z_2$, **both real.** $\rho(h) = c_1 z_1^{-h} + c_2 z_2^{-h}$ where we can find c_i from ϕ_i, like above, using a system of linear equalities (plugging in the two give values of 1 and $\phi_1/(1-\phi_2)$ for u_0 and u_1). Again goes to 0 exponentially fast ($|z_i| > 1$).

AR(2)

For a causal AR(2) process X_t, recall we derived

$$\rho(h) = \phi_1 \rho(h-1) + \phi_2 \rho(h-2) \quad \text{for all } h \geq 1,$$

with initial condition $\rho(0) = 1$ and $\rho(-1) = \phi_1/(1-\phi_2)$.

Third case: $z_1 = \bar{z}_2$ **(and neither is real).** Then $\rho(h) = c_1 z_1^{-h} + c_2 z_2^{-h}$ just as in the previous case (but z_i are complex now). Can show that $c_2 = \bar{c}_1$ b/c $\rho(h)$ is real. Using the fact that $e^{-i\theta} + e^{i\theta} = 2\cos(\theta)$, we have

$$\rho(h) = |c_1||z_1|^{-h}(e^{i(\theta_c - h\theta_z)} + e^{-i(\theta_c - h\theta_z)}) = |c_1||z_1|^{-h}2\cos(\theta_c - h\theta_z),$$

where we denote $c_1 = |c_1|e^{i\theta_c}$, $z_1 = |z_1|e^{i\theta_z}$.

ACFs

Table of Contents I

Table of Contents II

Table of Contents III

Forecasting

Forecasting: We observe data, $X_1, \ldots, X_n$. Denote $\boldsymbol{X}_n := (X_n, \ldots, X_1)'$
Want to *forecast* or *predict* future unobserved values, $X_{n+1}, X_{n+2}, \ldots$.

We will begin with the so-called *Best Linear Predictor (Forecaster)*.
(Later, after PACF and estimation, we will discuss two other predictors,
the so-called *infinite past predictor* and the *truncated infinite past
predictor*, which reveal different intuition for ARMA prediction.)

Two cases:

1. model parameters are known,
2. model parameters are unknown.

We start with the simpler case 1. The *Best Linear Predictor (BLP)*, X_{n+m}^n, is the *linear predictor*

$$X_{n+m}^n = \alpha_0 + \sum_{k=1}^{n} \alpha_k X_k = \sum_{k=0}^{n} \alpha_k X_k,$$

(if we let $X_0 = 1$) that minimizes mean-squared error

$$E\left[(X_{n+m} - (X_{n+m}^n)^2\right] = E\left[(X_{n+m} - (\sum_{k=0}^{n} \alpha_k X_k)^2\right].$$

(Our goal now: find the values of $\alpha_0, \alpha_1, \ldots, \alpha_n \in \mathbb{R}$ for the BLP.)

Assume $\boldsymbol{X}_n = (X_n, \ldots, X_1)'$ come from a stationary process. For convenience let $X_0 := 1$. Then, for $m \geq 1$, the best linear predictor X_{n+m}^n of X_{n+m} is found by solving the $n + 1$ equations

$$E\left[(X_{n+m} - \sum_{j=0}^{n} \alpha_j X_j)X_k\right] = 0 \qquad k = 0, 1, \ldots, n$$

for the $n + 1$ values $\alpha_0, \alpha_1, \ldots, \alpha_n$.

The α_j's depend on m, the forecast horizon (i.e., $\alpha_j \equiv \alpha_j^{(m)}$).

We can/will often simplify to take $\mu = 0$.

Recall the prediction equations:

$$E\left[(X_{n+m} - \sum_{j=0}^{n} \alpha_j X_j)X_k\right] = 0 \qquad k = 0, 1, \ldots, n.$$

If $EX_j = \mu$ then, by the $k = 0$ equation (recall: $X_0 = 1$),
$\mu = \alpha_0 + \sum_{j=1}^{n} \alpha_j \mu$ so $\alpha_0 = \mu(1 - \sum_{j=1}^{n} \alpha_j)$.

So we can rewrite X_{n+m}^n as $X_{n+m}^n = \mu + \sum_{j=1}^{n} \alpha_j(X_j - \mu)$. Can also check that the other α_j's are unchanged, whether $\mu = 0$ or not. We sometimes will assume $\mu = 0$ and ignore the $k = 0$ equation.

Consider $m = 1$. Assume $EX_j = 0$ and $\{X_j\}$ is stationary. Define $\phi_{n,j}$ by $\phi_{n,j} := \alpha_{n+1-j}$, write $\phi'_n := (\phi_{n1}, \ldots, \phi_{nn})$, and so write

$$
\begin{aligned}
X^n_{n+1} &:= \phi_{n1}X_n + \phi_{n2}X_{n-1} + \cdots + \phi_{nn}X_1 \\
&= \phi'_n \boldsymbol{X}_n,
\end{aligned}
$$

where the ϕ_{nj} solve

$$
E\left[(X_{n+1} - \sum_{j=1}^{n} \phi_{nj}X_{n+1-j})X_{n+1-k}\right] = 0 \qquad k = 1, \ldots, n.
$$

The *mean-squared* 1-*step ahead prediction error* is

$$P_{n+1}^n := E(X_{n+1} - X_{n+1}^n)^2 = \gamma(0) - \boldsymbol{\gamma}_n' \Gamma_n^{-1} \gamma_n.$$

A general *prediction interval* (here m-step ahead, using P_{n+m}^n) is given by $X_{n+m}^n \pm c_{\alpha/2} \sqrt{P_{n+m}^n}$.

- Here $c_{\alpha/2}$ is chosen according to the distribution of $X_{n+m} - X_{n+m}^n$ (i.e., usually based on W_{n+m}), and could in general be complicated,

- But for instance if the W_{n+m} is Gaussian then $X_{n+m} - X_{n+m}^n$ is Gaussian and so the $c_{\alpha/2}$ is taken to be a normal critical value, e.g., for a 95% prediction interval we use

$$X_{n+m}^n \pm 1.96 \sqrt{P_{n+m}^n}.$$

Suppose $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t$ is a mean-0 causal series where $W_t \sim \mathsf{WN}(0, \sigma^2)$. We might guess that based on a sequence $X_1, \ldots, X_n$, $n > p$, that we would forecast by

$$X_{n+1}^n = \phi_1 X_n + \cdots + \phi_p X_{n+1-p}.$$

What is P_{n+1}^n?

We derived $\phi_n = \Gamma_n^{-1}\gamma_n$. To avoid matrix inversion, the *Durbin-Levinson Algorithm* can be used to compute the one-step coefficients $\phi_{n1}, \ldots, \phi_{nn}$ and P_{n+1}^n.

Forecasting
m steps ahead

Given a mean-0 (stationary, causal, invertible) series $\{X_t\}_{t=1}^n$, the m-step-ahead BLP is

$$X_{n+m}^n := \phi_{n1}^{(m)} X_n + \phi_{n2}^{(m)} X_{n-1} + \cdots + \phi_{nn}^{(m)} X_1$$

where for $k = 1, \ldots, n$,

$$\sum_{j=1}^n \phi_{nj}^{(m)} E(X_{n+1-j} X_{n+1-k}) = E(X_{n+m} X_{n+1-k})$$

or, for $k = 1, \ldots, n$,

$$\sum_{j=1}^n \phi_{nj}^{(m)} \gamma(k-j) = \gamma(m+k-1).$$

The m-step-ahead prediction error is ($\boldsymbol{\gamma}_n^{(m)} = (\gamma(m), \ldots, \gamma(m+n-1))$)

$$P_{n+m}^n = E(X_{n+m} - X_{n+m}^n)^2 = \gamma(0) - \boldsymbol{\gamma}_n^{(m)'} \Gamma_n^{-1} \boldsymbol{\gamma}_n^{(m)}.$$

In *backcasting* we want to predict the data series in the past, before our data begin (before we were able to start collecting data; e.g. in climate settings). So we have a mean-0 ARMA series $\{X_1, \ldots, X_n\}$ and we want to predict $X_0, X_{-1}, \ldots$; or in general, X_{1-m}. If the m-step *backcast* is $X_{1-m}^n = \sum_{j=1}^n \alpha_j X_j$, then the prediction equations are, for $k = 1, \ldots, n$,

$$\sum_{j=1}^n \alpha_j E(X_j X_k) = E(X_{1-m} X_k).$$

These are the same as for forward prediction,

$$\sum_{j=1}^n \alpha_j \gamma(k - j) = \gamma(m + k - 1),$$

for $k = 1, \ldots, n$! So $\alpha_j = \phi_{nj}^{(m)}$, $j = 1, \ldots, n$, and the backcast is

$$X_{1-m}^n = \phi_{n1}^{(m)} X_1 + \cdots + \phi_{nn}^{(m)} X_n.$$

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Partial ACF

The ACF is great for identifying the order q of an MA(q). It does not do well for identifying p in an AR(p).

The goal of the Partial ACF (PACF) is to consider dependence between X_{t+h} and X_t after adjusting for the terms $X_{t+h-1}, \ldots, X_{t+1}$.

The idea is to linearly "forecast" X_{t+h} based on $X_{t+h-1}, \ldots, X_{t+1}$. We denote this forecast as $\hat{X}_{t+h}$. Formally,

$$\hat{X}_{t+h} = \beta_1 X_{t+h-1} + \beta_2 X_{t+h-2} + \cdots + \beta_{h-1} X_{t+1}$$

where the β_j's minimize
$E(X_{t+h} - (\beta_1 X_{t+h-1} + \beta_2 X_{t+h-2} + \cdots + \beta_{h-1} X_{t+1}))^2$.

Similarly, let

$$\check{X}_t = \beta_1 X_{t+1} + \beta_2 X_{t+2} + \ldots + \beta_{h-1} X_{t+h-1}$$

be the "backcast" of X_t based on (the future variables) $X_{t+1}, \ldots, X_{t+h-1}$.

Partial ACF

Definition

The *partial autocorrelation function (PACF)* of a stationary process $\{X_t\}$ for $h = 1, 2, \ldots$, is

$$PACF(1) := \text{Cor}(X_{t+1}, X_t) = \rho(1)$$

$$PACF(h) := \text{Cor}(X_{t+h} - \hat{X}_{t+h}, X_t - \check{X}_t) \quad \text{for} \quad h \geq 2.$$

And $PACF(h) = PACF(-h)$.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

# PACF
AR(1)

Theorem

The partial autocorrelation function (PACF) *of a stationary process* $\{X_t\}$, *for* $h = 1, 2, \ldots,$ *satisfies*

$$PACF(h) = \phi_{hh},$$

where ϕ_{hh} *is the final coefficient for the one-step-ahead predictor based on* h *observations (and can be computed by the Durbin-Levinson algorithm).*

AR(p)

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Let $\hat{\phi}_{hh}$ be the last coordinate of $\hat{\phi} = \hat{\Gamma}_h^{-1}\hat{\gamma}_h$ where $\hat{\gamma}(i)$ is the sample autocovariance estimate.

Theorem

For a causal AR(p), $\hat{\phi}_{hh} \sim AN(0, n^{-1})$ for any $h > p$.

(This yields the dotted blue lines in the PACF plots from the acf() or acf2() functions in R.)

Consider an invertible MA(1) process.

Southern Oscillation Index

Recruitment
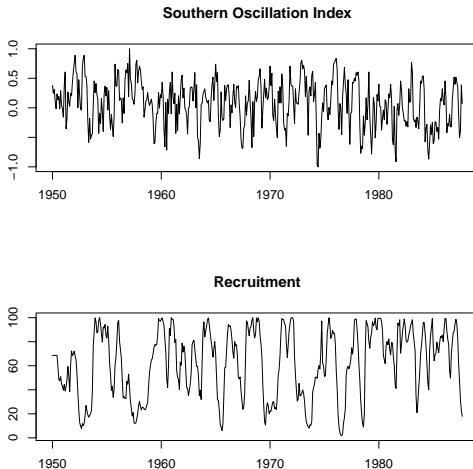
Let's return to forecasting. We developed the BLP for any ARMA, but focused in detail on how things work for the AR case. What about MA or ARMA models?

To answer this question, we develop a different approach, which turns out to be nearly identical at the end of the day. For theoretical purposes, we might consider pretending we had an infinite amount of data. (When we have a large sample size, this might be a good approximation.) For a causal, invertible $\{X_t\}$, the *infinite past predictor* of X_{n+m}, denoted by $\tilde{X}_{n+m}$, is

$$\tilde{X}_{n+m} := E\left[X_{n+m} | X_n, X_{n-1}, \ldots, X_1, X_0, X_{-1}, \ldots\right].$$

Its mean-squared error is denoted

$$\tilde{P}_{n+m} := E\left(X_{n+m} - \tilde{X}_{n+m}\right)^2.$$

Recall $\tilde{X}_{n+m} := E\left[X_{n+m}|X_n, X_{n-1}, \ldots, X_1, X_0, X_{-1}, \ldots\right]$ and $\tilde{P}_{n+m}^n := E\left(X_{n+m} - \tilde{X}_{n+m}\right)^2$.

We can use the causal and invertible representations of $\{X_t\}$ to derive general formulas for $\tilde{X}_{n+m}$ and $\tilde{P}_{n+m}$. Assume X_t is mean 0, and let

$$\tilde{W}_t := E(W_t|X_n, \ldots, X_1, X_0, \ldots) = E(W_t|W_n, \ldots, W_1, W_0, \ldots).$$

Then for $m \geq 1$ we have

$$\tilde{X}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{X}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j X_{n+m-j}, \quad \text{and}$$

$$\tilde{P}_{n+m} = \sigma^2 \sum_{j=0}^{m-1} \psi_j^2.$$

The infinite past predictor $\tilde{X}_{n+m}$ is not realistic of course. So, the *truncated (m-step-ahead) predictor* (based on n observations), $\tilde{X}_{n+m}^n$ is

$$\tilde{X}_{n+m}^n := -\sum_{j=1}^{m-1} \pi_j \tilde{X}_{n+m-j}^n - \sum_{j=m}^{n+m-1} \pi_j X_{n+m-j}$$

This is similar to the infinite past predictor. It also turns out to be often very similar or, in the case or AR models, identical to the BLP X_{n+m}^n.

For AR(p) models (with $n > p$): $\tilde{X}_{n+m} = \tilde{X}_{n+m}^n = X_{n+m}^n$.

Truncated predictor for ARMA: we do not need to explicitly find the π_j's (the invertible representation). Rather, it turns out that the truncated predictor for ARMA models can be computed using the ARMA specification directly.

For (stationary, causal, invertible) ARMA(p, q) series, the truncated predictors for $m = 1, 2, \ldots$, can be found by s

$$\widetilde{X}_{n+m}^n = \phi_1 \widetilde{X}_{n+m-1}^n + \cdots + \phi_p \widetilde{X}_{n+m-p}^n + \theta_1 \widetilde{W}_{n+m-1}^n + \cdots + \theta_q \widetilde{W}_{n+m-q}^n,$$

where $\widetilde{X}_t^n = X_t$ for $1 \leq t \leq n$ and $\widetilde{X}_t^n = 0$ for $t \leq 0$, and where $\widetilde{W}_t^n = 0$ for $t \leq 0$ or $t > n$, and for $1 \leq t \leq n$,

$$\widetilde{W}_t^n := \phi(B)\widetilde{X}_t^n - \theta_1 \widetilde{W}_{t-1}^n - \cdots - \theta_q \widetilde{W}_{t-q}^n.$$

(Note that $\widetilde{W}_t^n$ is not the same as $\widetilde{W}_t$.)

SS, Example 3.24, Forecasting an ARMA(1,1):

Forecasting
Prediction using data

Predicting for ARMA(p, q) based on data (and not knowing the real model): estimate the $p + q$ parameters and plug-in.

In R can use the predict() function which implements the best linear predictor based on MLE's.

Table of Contents I

Charles R. Doss   Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Table of Contents II

Table of Contents III

Estimation

There are many factors to consider in estimation: efficiency (effectiveness), computation, complication/simplicity. We will discuss

- maximum likelihood estimators (MLEs),
- unconditional least-squares estimators (ULSEs),
- and conditional least-squares estimators (CLSEs).

Finally, we will also discuss

- method of moments estimators, known as Yule-Walker (YW) estimators.

The "likelihood" in "maximum likelihood" is based on W_t being i.i.d. Normal/Gaussian. What this means is that if the white noise W_t is far from i.i.d. or has very heavy tails, then these estimators will not be effective.

To begin with, we discuss MLE's, ULSE's, and CLSE's, which are all closely related, in the AR(1) case.

We write

$$f(x_1, ..., x_n) = f(x_1)f(x_2|x_1)f(x_3|x_2, x_1) \cdots f(x_n|x_{n-1}, ...x_1)$$
$$= f(x_1)f(x_2|x_1)f(x_3|x_2) \cdots f(x_n|x_{n-1})$$

because $\{X_t|(X_{t-1}, ..., X_1)\} \sim N(\mu + \phi(X_{t-1} - \mu), \sigma^2)$ ("Markov property") ie $X_t|(X_{t-1}, \ldots, X_1)$ is the same as $X_t|X_{t-1}$ for an AR(1). Thus

$$f(x_t|x_{t-1}) = f_W((x_t - \mu) - \phi(x_{t-1} - \mu))$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}((x_t - \mu) - \phi(x_{t-1} - \mu))^2\right\}.$$

(where f_W is the $N(0, \sigma^2)$ density of W_t).

The unknown parameters for a stationary AR(1) (with arbitrary mean and $W_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$) are (ϕ, μ, σ^2). Then it turns out, we can write the likelihood as

$$
\begin{aligned}
L(\mu, \phi, \sigma^2) &= (2\pi\sigma^2)^{-n/2}(1-\phi^2)^{1/2} \exp\left[-\frac{S(\mu, \phi)}{2\sigma^2}\right] \\
&= (2\pi\sigma^2)^{-n/2}(1-\phi^2)^{1/2} \exp\left[-(1-\phi^2)(x_1-\mu)^2/2\sigma^2\right] \times \\
&\qquad \exp\left[-S_c(\mu, \phi)/2\sigma^2\right]
\end{aligned}
$$

where

$$
S(\mu, \phi) = (1-\phi^2)(x_1-\mu)^2 + \sum_{t=2}^{n}((x_t-\mu) - \phi(x_{t-1}-\mu))^2, \text{ and}
$$

$$
S_c(\mu, \phi) = \sum_{t=2}^{n}((x_t-\mu) - \phi(x_{t-1}-\mu))^2.
$$

To estimate σ^2 we can use the MLE, $\widehat{\sigma}^2 = n^{-1}S(\widehat{\mu}, \widehat{\phi})$. (Other variants are possible, but this suffices.) For μ and ϕ,

- the MLE's are the maximizers of $L(\mu, \phi, \sigma^2)$,
- the ULSE's are the minimizers of $S(\mu, \phi)$, and
- the CLSE's are the minimizers of $S_c(\mu, \phi)$.

Let $\widehat{\phi}$ and $\widehat{\theta}$ be either the MLE's, ULSE's, or CLSE's.

Theorem (Asymptotic normality of MLE's, ULSE's, CLSE's)

For causal, invertible, mean 0 ARMA process $\phi(B)X_t = \theta(B)W_t$, under appropriate conditions on $\{W_t\}$,

$$\begin{pmatrix} \widehat{\phi} \\ \widehat{\theta} \end{pmatrix} \sim AN\left(\begin{pmatrix} \phi \\ \theta \end{pmatrix}, \frac{\sigma^2}{n}\Gamma^{-1} \right), \text{ where}$$

$$\Gamma := \mathrm{Cov}(X, X),$$

where X is a $(p+q)$-length vector we will specify shortly.

- (Var-Cov matrices): For any random vector $X = (X_1, \ldots, X_m)$,

$$\mathrm{Cov}(X, X) = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_m) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_m) \\ \vdots & & & \\ \mathrm{Cov}(X_m, X_1) & \mathrm{Cov}(X_m, X_2) & \cdots & \mathrm{Cov}(X_m, X_m) \end{pmatrix}$$
$$= (\mathrm{Cov}(X_i, X_j))_{i,j=1}^m$$

- (Confidence intervals): The asymptotics let us form confidence intervals (CI's) or hypothesis tests for the unknown parameters. Let $\Sigma = \Gamma^{-1}\sigma^2/n$. From the asymptotics on previous slide, a $(1 - \alpha) \times 100\%$ two-sided CI for, say, ϕ_1 is given by: $\widehat{\phi}_1 \pm z_{\alpha/2}\sqrt{\Sigma_{1,1}}$. (Similarly for any of the other ϕ_i or θ_i parameters, using the appropriate diagonal entry of Σ.)

Let $\widehat{\phi}$ and $\widehat{\theta}$ be either the MLE's, ULSE's, or CLSE's from n data points.

Theorem (Asymptotic normality of MLE's, ULSE's, CLSE's)

For causal, invertible, mean 0 ARMA process $\phi(B)X_t = \theta(B)W_t$, where W_t is i.i.d. WN$(0, \sigma^2)$,

$$\begin{pmatrix} \widehat{\phi} \\ \widehat{\theta} \end{pmatrix} \sim approx. N\left(\begin{pmatrix} \phi \\ \theta \end{pmatrix}, \frac{\sigma^2}{n} \Gamma^{-1} \right), \text{ where}$$

$$\Gamma := \begin{pmatrix} \Gamma_{\phi,\phi} & \Gamma_{\phi,\theta} \\ \Gamma'_{\phi,\theta} & \Gamma_{\theta,\theta} \end{pmatrix} = \text{Cov}((Z,Y),(Z,Y)),$$

$$Z = (Z_1, \ldots, Z_p); \quad \phi(B)Z_t = W_t$$
$$Y = (Y_1, \ldots, Y_q); \quad \theta(B)Y_t = W_t.$$

Note, another way to describe Γ is to say that the i,j entry of the $p \times p$ matrix $\Gamma_{\phi,\phi}$ is $\gamma_z(i - j)$ (autocovariance of $\{Z_t\}$), the, i,j entry of the $q \times q$ matrix $\Gamma_{\theta\theta}$ is $\gamma_y(i - j)$ (autocovariance of $\{Y_t\}$), and the i,j entry of the $p \times q$ matrix $\Gamma_{\phi,\theta}$ is $\gamma_{zy}(i - j)$ (for $i = 1, \ldots, p$, $j = 1, \ldots, q$).

Examples (see page 126, 4ed, SS):

Same (optimal) limit for all three estimators, meaning that if n is large they are all effectively the same (so, if n large, we can use CLSE's if we want).

Method of moments estimators are based on equating the population moments, in a given model, with sample moments. In ARMA settings, they are efficient (meaning, have low/optimal variance) for AR estimation but not for MA estimation. We consider just the AR case.

The Yule-Walker estimators of $\phi_1, \ldots, \phi_p, \sigma^2$ (based on asuming Y_t is a causal AR(p) process) are

$$\hat{\phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p \quad \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\gamma}_p' \hat{\Gamma}_p^{-1} \hat{\gamma}_p. \tag{6}$$

If $\{Y_t\}$ is not mean 0 then we estimate μ by $\hat{\mu} = \bar{Y}$ (and the intercept by $\hat{\delta} = \hat{\mu}(1 - \hat{\phi}_1 - \cdots - \hat{\phi}_p)$).

Note: by the definition of $\widehat{\phi}$ in (6), we see the definitions of $\widehat{\phi}_h$ and $\widehat{\phi}_{hh}$ (final 1-step-ahead coefficient, based on $n = h$ data points) are identical. And thus the sample PACF is defined to be $\widehat{\text{PACF}}(h) = \widehat{\phi}_h$.

Yule-Walker (method of moments)

Asymptotics

Theorem

If $\{X_t\}$ is a causal AR(p) process then

$$\hat{\phi} \sim AN\left(\phi, n^{-1}\sigma^2\Gamma_p^{-1}\right)$$

and $\hat{\sigma}^2 \rightarrow_p \sigma^2$ as $n \rightarrow \infty$.

Furthermore this applies when the fitted order m is greater than the true order p. In particular, since the YW AR(h) estimate of ϕ_h is $\hat{\phi}_{hh}$, we have the following.

Theorem

For a causal AR(p), $\hat{\phi}_{hh} \sim AN(0, n^{-1})$ for any $h > p$.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Table of Contents I

Table of Contents II

Table of Contents III

State space models: the dynamic linear model

Let

$$X_t = \Phi X_{t-1} + W_t \tag{7}$$
$$Y_t = A X_t + V_t \tag{8}$$

where W_t is $p \times 1$ iid $N_p(0, Q)$, so also $X_t \in \mathbb{R}^p$, $\Phi \in \mathbb{R}^{p \times p}$, $Y_t, V_t \in \mathbb{R}^q$, with V_t iid $N_q(0, R)$, and A is a $q \times p$ measurement (observation) matrix.

In this "dynamic linear model" or "state-space model", we assume we observe Y_t but not the "state" variable X_t (nor the noise variables, W_t, V_t).

Some assumption is needed about X_0, e.g., $X_0 \sim N_p(\mu_0, \Sigma_0)$. The equation (8) is the "measurement equation" and (7) is the "state equation". Assume that X_0, W_t, V_t are all independent.

State space models: the dynamic linear model

We may extend the dynamic linear model to include possibly a vector of observed "exogenous" (covariate) variables $u_t \in \mathbb{R}^r$ that affect the observations but are not affected by the observations:

$$X_t = \Phi X_{t-1} + \Upsilon u_t + W_t$$
$$Y_t = A X_t + \Gamma u_t + V_t,$$

where Υ is $p \times r$ and Γ is $q \times r$. These exogenous covariates don't change much about the computation of the Kalman filters or the likelihood, and in practice are usually present. For basic derivations we may leave them off though.

Filtering, prediction, smoothing

Usually when a model is formulated in a state-space fashion, estimating/predicting the underlying state (random) variables is an important goal. This is also necessary for computing the likelihood (and doing estimation). We let $X_t^s := E(X_t|Y_{1:s})$ (where $Y_{1:s} := \{Y_1, \ldots, Y_s\}$).

Three different settings:

- X_t^s with $s < t$: *prediction*,
- X_t^s with $s = t$: *filtering*,
- X_t^s with $s > t$: *smoothing*.

And

$$P_t^s := E[(X_t - X_t^s)(X_t - X_t^s)'] \equiv E[(X_t - X_t^s)(X_t - X_t^s)'|Y_{1:s}].$$

Kalman filter and prediction

Leave off the external covariates u_t. Fix initial condition $X_0^0 = x_0^0$ and $P_0^0 = \Sigma_0$. Assume you have computed X_{t-1}^{t-1} and P_{t-1}^{t-1}. For $t = 1, \ldots, n$,

$$X_t^{t-1} = \Phi X_{t-1}^{t-1} \tag{9}$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi' + Q, \tag{10}$$

$$X_t^t = X_t^{t-1} + K_t(Y_t - A X_t^{t-1}) \tag{11}$$

$$P_t^t = [I - K_t A] P_t^{t-1} \tag{12}$$

$$K_t = P_t^{t-1} A' [A P_t^{t-1} A' + R]^{-1}. \tag{13}$$

Here X_t^t is the Kalman *filter*. Prediction, i.e. X_t^n with $t > n$, is via iterating the state update equations (9) and (10) (and noting *for $t > n$* that $X_t^t = X_t^{t-1}$, similarly for P). Important benefit: simple to update, i.e. to go from X_{t-1}^{t-1} to X_t^t. For intermediate calculations: let $\epsilon_t := Y_t - E(Y_t | Y_{1:(t-1)})$. Then $\epsilon_t = Y_t - A X_t^{t-1}$. And then $\Sigma_t := \text{Cov}(Y_t | Y_{1:t-1}) = \text{Cov}(\epsilon_t) = (R + A P_t^{t-1} A')$.

Kalman filter and prediction

The Kalman filter derivation is based on full knowledge of model parameters (no estimation has been done).

Multivariate Gaussian fact review

Conditional normal distributions have a specific form.

Theorem

Let $\epsilon = (\epsilon_1, \ldots, \epsilon_m)'$, $X = (X_1, \ldots, X_n)'$ and suppose

$$\begin{pmatrix} X \\ \epsilon \end{pmatrix} \sim N_{n+m} \left(\begin{pmatrix} \mu_x \\ \mu_\epsilon \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{x\epsilon} \\ \Sigma_{\epsilon x} & \Sigma_{\epsilon\epsilon} \end{pmatrix} \right)$$

(where $\Sigma_{\epsilon x} = \Sigma'_{x\epsilon}$). Then[a] $X | \epsilon$ is normal, with mean $\mu_{X|\epsilon} = \mu_x + \Sigma_{x\epsilon} \Sigma_{\epsilon\epsilon}^{-1}(\epsilon - \mu_\epsilon)$ and covariance $\Sigma_{X|\epsilon} = \Sigma_{xx} - \Sigma_{x\epsilon} \Sigma_{\epsilon\epsilon}^{-1} \Sigma_{\epsilon x}$.

---
[a]Assume $\Sigma_{\epsilon\epsilon}^{-1}$ exists.

Multivariate fact review

These calculations use the identities $E(AX) = AE(X)$ and $\text{Cov}(AX, BY) = A\,\text{Cov}(X, Y)B'$ (where A, B, are fixed matrices of appropriate dimensions and X, Y are random vectors).

DLM likelihood

We can use the (Gaussian) likelihood for estimation in any DLM. Assume the noise variables are Gaussian. Then the $\{X_t\}$ and $\{Y_t\}$ variables are together/jointly Gaussian. It is a fact then that $\{Y_t\}$ are normal (marginally). We can write the likelihood as

$$f(y_1, ..., y_n) = f(y_1)f(y_2|y_1)f(y_3|y_2, y_1) \cdots f(y_n|y_{n-1}, ...y_1).$$

Each (conditional) density is normal. All we need are the conditional means and covariances (of Y_t (not X_t) given $Y_{1:t-1}$). We can also get these from Kalman filtering.

DLM likelihood

Let $\epsilon_t := Y_t - E(Y_t|Y_{1:(t-1)})$. Then $\epsilon_t = Y_t - AX_t^{t-1}$. And then $\Sigma_t := \mathrm{Cov}(Y_t|Y_{1:t-1}) = \mathrm{Cov}(\epsilon_t) = (R + AP_t^{t-1}A')$.

We can write the likelihood as

$$f(y_1, ..., y_n) = f(y_1)f(y_2|y_1)f(y_3|y_2, y_1) \cdots f(y_n|y_{n-1}, ...y_1).$$
$$= f(y_1)f(\epsilon_2|y_1)f(\epsilon_3|y_2, y_1) \cdots f(\epsilon_n|y_{n-1}, ...y_1).$$

The log likelihood, $\log L(\theta)$ in the end, is

$$-\sum_{t=1}^{n} \log \det \Sigma_t(\theta) - \sum_{t=1}^{n} \epsilon_t(\theta)'\Sigma_t(\theta)^{-1}\epsilon_t(\theta).$$

(where factor of $1/2$ and additive constants are removed).

DLM and ARMA models

ARMA models can be written in state-space form. There are many ways this can be done. Here is one.

Assume we have (one-dimensional) zero-mean ARMA(p, q) process Y_t with $\phi(B)Y_t = \theta(B)W_t$ and $W_t \sim N(0, \sigma_a^2)$. Let $m = \max(p, q+1)$. Rewrite the model as

$$Y_t = \sum_{i=1}^{m} \phi_i Y_{t-i} + W_t + \sum_{j=1}^{m-1} \theta_j W t - j$$

with $\phi_i = 0$ for $i > p$ and $\theta_j = 0$ for $j > q$. (Always $\theta_m = 0$, note.)

Then let X_t be an m dimensional state vector with

$$X_t = TX_{t-1} + R\eta_{t-1} \tag{14}$$
$$Y_t = ZX_t \tag{15}$$

with

$$T = \begin{pmatrix} \phi_1 & 1 & 0\ldots & 0 \\ \phi_2 & 0 & 1\ldots & 0 \\ \vdots & & & \vdots \\ \phi_{m-1} & 0 & 0\ldots & 1 \\ \phi_m & 0 & 0 & 0 \end{pmatrix} \tag{16}$$

and $R = (1, \theta_1, \ldots, \theta_{m-1})'$, with $Z = (1, 0, \ldots, 0)_{1 \times m}$ and $\eta_t \sim N(0, \sigma_a^2)$.

In practice, many implementations of ARMA techniques (e.g., MLE computation, best linear prediction) are done by putting the ARMA model in a state-space form and then using the state-space results. Kalman prediction is best linear prediction for instance.

Table of Contents I

Table of Contents II

Table of Contents III

Nonstationary (integrated) models

Not all series are stationary.
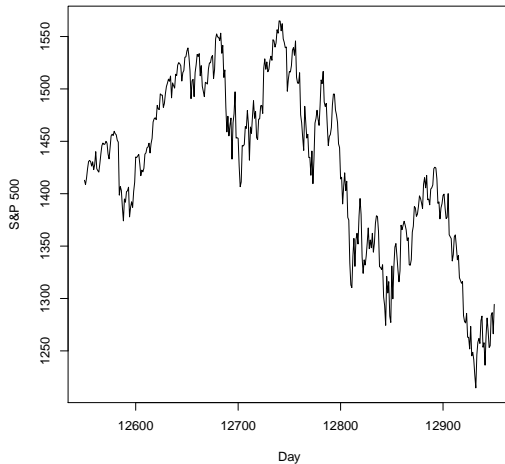
ARIMA models

Recall $\nabla = 1 - B$, the differencing operator:

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1}.$$

If $Y_t = \nabla X_t$, then $\nabla Y_t = \nabla(\nabla X_t) = \nabla^2 X_t$.

Definition

For $p, d, q \geq 0$, we say that $\{X_t\}$ is an ARIMA(p, d, q) process if $Y_t = \nabla^d X_t$ is an ARMA(p, q) process. Thus,

$$\phi(B)(1 - B)^d X_t = \theta(B)W_t.$$

If $d > 0$ then X_t is nonstationary.

Example: Random walk. If $\{X_t\}$ is a random walk, $X_t = X_{t-1} + W_t$, then $Y_t = \nabla X_t = X_t - X_{t-1} = W_t$ is white noise. Thus a random walk is ARIMA(0,1,0).

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Identifying d

Usually d is either 0 (stationary ARMA) or 1, very rarely 2. We identify if $d > 0$ by examining the ACF: very slow decay indicates nonstationarity ($d > 0$).

Table of Contents I

Charles R. Doss   Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Table of Contents II

Table of Contents III

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Fractionally integrated time series (ARFIMA)

Taking $d = 1$ (using first differences) on a series can be too severe. A different way to model long memory processes is to do *fractional differencing*: take $|d| < .5$, formally speaking. That is

$$(1 - B)^d X_t = W_t.$$

The binomial expansion for $(1 - z)^{-d}$, $d < 1$, is $(1 - z)^{-d} = \sum_{j=0}^{\infty} \psi_j z^j$ where $\psi_j = \Gamma(j + d)/\Gamma(j + 1)\Gamma(d)$, and Γ is the Gamma function (the continuous analog of the factorial, i.e. $\Gamma(n) = (n - 1)\Gamma(n - 1)$)[11].

Thus

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

with $\psi_j = \Gamma(j + d)/\Gamma(j + 1)\Gamma(d)$.

---

[11]The actual definition of Γ is $\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx$

Fractionally integrated time series (ARFIMA)

The definitions of ψ_j (on the previous slide) ends up implying for $0 < d < .5$ that $\sum_{h=0}^{\infty} |\rho(h)| = \infty$ (where ρ is the ACF). For this reason these are known as *long memory processes*.

Note: with $-.5 < d < .5$, ARFIMA processes are stationary. If $0 < d < .5$ they are long memory. If $-.5 < d < 0$ they are anti-persistent (mean reverting) but not long memory. With $|d| \geq .5$, they are non-stationary.

Fractionally integrated time series (ARFIMA)

Note that when $d = 1$, we have

$$X_t = \sum_{j=0}^{t-1} \psi_j W_{t-j}$$

with $\psi_j = 1$ for all j.

When $|d| < .5$, the same representation holds but the ψ_j are not all 1 (rather, from the earlier slide, $\psi_j = \Gamma(j + d)/\Gamma(j + 1)\Gamma(d) \leq 1$).

Analyzing ARIMA series

Once we've identified d ($d \in \{0, 1, 2, \ldots\}$), we have $Y_t = \nabla^d X_t$ is ARMA(p,q). Thus we analyze the d-difference'd process Y_t using our tools for ARMA processes.

When estimating/analyzing a series in R, the `astsa::sarima()` or `arima()` functions do estimation and (with argument for d passed in) can automate the difference taking. Sometimes it may be better/simpler to just difference (see the `diff()` function) yourself though (and pass in the differenced series to `arima`/`sarima`, with d set to 0).

Model Selection

We've focused so far on assuming we know the values of p and q ahead of time. In reality, discovering the values of p and q (i.e., *model selection*) is usually the hardest part.

So far, we have discussed model selection using the ACF and PACF. But these are mostly useful if the true model is either pure AR or pure MA, but not if it is ARMA(p, q) with both p and q nonzero.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

For causal and invertible processes there are information criteria (IC) to use for model selection. If a mean/intercept is included, then let $\beta = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ be the vector of parameters. Let $L(\widehat{\beta}, \sigma^2(\widehat{\beta})) \equiv L(p, q)$. Then

$$AIC = -2 \log L(p, q) + 2(p + q + 1),$$

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

$$BIC = -2 \log L(p, q) + (p + q + 1) \log n$$

where n is the sample size. Then the (p, q) with smallest IC value is best (according to that IC).

One very basic rule of thumb[12] is that models within 2 IC units are basically indistinguishable. Note: `sarima()` outputs the above IC's divided by n (so "2" IC units becomes "$2/n$" IC units).

IC's are very useful in model selection, but their output is not the final word; it is just one piece of the puzzle. Furthermore, after doing model selection with IC's, we need still to do diagnostic checking (to be discussed next). An important caveat: these IC's are all based on having i.i.d. Gaussian white noise!

IC's *cannot* be used to compare across different values of d, or across different transformations of the data (which we will discuss shortly). (E.g., you cannot choose $d = 0$ or $d = 1$ based on comparing IC's.)

---

[12]Which must be taken with a grain of salt.

(Standardized) Residual diagnostics

After fitting a model we need to assess the fit which we do via the residuals and potentially via looking at the estimates' p-values. In time series models, (standardized) residuals are from the one-step-ahead predictions:

$$\widehat{W}_t := X_t - X_t^{t-1} \quad \text{or} \quad \widehat{e}_t := \widehat{W}_t / \sqrt{P_t^{t-1}}.$$

Think of them as you would think of (standardized) residuals in linear regression.

In R, the `sarima()` function provides some automated diagnostics as output, which we focus on. If the automated formatting is not sufficient you can use the output from the `sarima()` function to create modified plots of the individual diagnostics on your own.

Most of the diagnostics we have talked about or are familiar from linear regression. One is new though: the *Ljung-Box-Pierce (LBP) Q-statistic*. The LBP statistic is like an aggregated look at the ACF from lags $h = 1$ up to lag H, where H is a tuning parameter you have to choose (maybe $H = 10, 20, 30$ depending on how large n is). Usually we look at a variety of H values. The actual formula is

$$Q = n(n + 2) \sum_{h=1}^{H} \frac{\rho^2(h)}{n - h}.$$

This statistic can be compared to a χ^2_{H-p-q} for $H \geq p + q + 1$ (and n much larger than H). It is testing whether $H_0 : \rho(h) = 0$ for all $h = 1, \ldots, H$ holds.
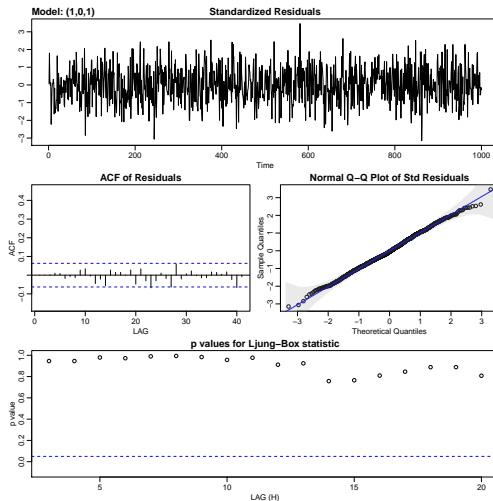
(Standardized) Residual diagnostics

Figure: `sarima()` diagnostics of model residuals

(Standardized) Residual diagnostics

- Examine standardized residuals for any obvious departures from being i.i.d. $(0, 1)$; departures can mean model too small, can affect quality of estimates, and especially validity of p-values, IC's, and of LBP test.
- Examine ACF for remaining correlation structure (suggesting larger model needed); lagged scatterplots also useful for verifying i.i.d. structure
- Examine QQ plot for departure from normality; non-normal residuals can signify
  - missing terms in fit,
  - means the i.i.d. assumption harder to check
  - and if residuals are heavy tailed, means estimates, p-values, IC's, and LBP diagnostic are unreliable
- Check LBP test(s)

ARIMA data analysis recipe

- plot the data
- (potentially) transform the data [skip this step for now]
- identify d, and take appropriate differences if necessary
- identify plausible choices (maybe multiple!) for p, q using AICc, BIC, and ACF/PACF
- estimate parameters
- examine diagnostics (and potentially repeat the above steps until picking one/more best models)
    - plot residuals
    - ACF of residuals
    - Normal QQ-plot of residuals
    - Ljung-Box-Pierce (LBP) Q-statistics for a range of reasonable H
    - Examine significance of fitted coefficients

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

ARIMA data analysis

See R code.

In cases where, even for the best fitting model, the standardized residuals do not appear to be $N(0,1)$, or other diagnostics fail, transformations may be of help. The main reason for doing a transformations is that the variance is proportional to the mean (just like "variance-stabilizing transformations" in linear regression). In some cases, domain-specific information may suggest a transformation as well. Transformations can on occasion be helpful with asymmetry.

Model fitting and diagnostics
Transformations of the data

Transformations are usually implemented before the process of identifying d and potentially differencing. The most common transformations (on a variable X) are the power or Box-Cox transformations:

$$(X^\lambda - 1)/\lambda \qquad\qquad \text{for } \lambda \neq 0$$
$$\log X \qquad\qquad (\text{if “}\lambda = 0\text{”})$$

Usually we would consider λ in a range of -2 to 2 or so.

The same issues arise as in linear regression:

- the data must be positive,
- transformations behave differently depending on the centering of the data,
- and IC's cannot be used to compare models across different transformations.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Model fitting and diagnostics
Transformations of the data

See R code (varve example and simulations).

Model fitting
Further "tips and tricks"

The connection between the PACF and the Yule-Walker estimates can be useful; Yule-Walker fits can sometimes be more informative versions of the PACF.

One can glean extra information from ACF and PACF of the residuals to be used in model fitting.

ARIMA data analysis recipe

- plot the data
- potentially do variance-stabilizing transform on the data
- identify d, and take appropriate differences if necessary
- identify plausible choices (maybe multiple!) for p, q using AICc, BIC, ACF/PACF, other tips&tricks
- estimate parameters
- examine diagnostics (and potentially repeat the above steps until picking a best fit (or multiple fits if the data cannot distinguish)
    - plot residuals
    - ACF of residuals
    - Normal QQ-plot of residuals
    - Ljung-Box-Pierce (LBP) Q-statistics for a range of reasonable H
    - Examine significance of fitted coefficients[13]

---

[13]But be careful.

See R code.

Table of Contents I

Table of Contents II

Table of Contents III

Seasonal ARMA models

Sometimes data have seasonal behavior (weekly, monthly, quarterly, etc.)
– we may know a priori the season or may observe it in the data

Seasonal ARMA models

If $\{X_t\}$ is a firm's monthly sales number, the previous few months may not capture the dynamics of the series correctly: e.g., sales in Nov and Dec may not predict sales in Jan very well, if Jan sales are systematically different than Dec and Nov sales (e.g., because of holiday seasons). We may want to look back at the most recent Jan sales data.

E.g., $X_t - \phi X_{t-12} = W_t$. Or $X_t - \phi_1 X_{t-1} - \phi_{1,s}(X_{t-12} - \phi_1 X_{t-13}) = W_t$.

Note these are ARMA still; e.g., the first of the above models is ARMA$(12, 0)$.

Here, seasonality is captured by variance (time series dynamics) rather than a seasonal time mean/trend.

Pure seasonal ARMA models

A seasonal model is ARMA. However, we do not want to use 12 parameters if prior knowledge suggests 1 will do, and seasonality is often interpretable. Thus we model seasonality explicitly. If $X_t - \Phi_1 X_{t-12} - \Phi_2 X_{t-24} = W_t$ we can write $(1 - \Phi_1 B^{12} - \Phi_2 B^{24})X_t = W_t$.

Definition (Pure seasonal ARIMA model)

Given P, D, Q and $s \geq 1$, a time series $\{X_t\}$ is an ARIMA$(P, D, Q)_s$ process if $\Phi(B^s)\nabla_s^D X_t = \Theta(B^s)W_t$. Here $\nabla_s^D = (1 - B^s)^D$ is the seasonal difference, and Φ is an order P polynomial and Θ is an order Q polynomial and Φ and Θ have no common roots.

- The process is causal if and only if the roots of $\Phi(z)$ are outside the unit circle.
- The process is invertible if and only if the roots of $\Theta(z)$ are outside the unit circle.

The ACF and PACF for an ARMA$(P, Q)_s$ are zero for $h \neq ks$ ($k \in \mathbb{N}$). For $h = ks$, the ACF or PACF is identical to what it would be at lag k in the corresponding (non-seasonal) ARMA(P, Q).

That is: the ACF of $\Phi(B^s)\nabla_s^D X_t = \Theta(B^s)W_t$ can be found from the ACF of $\Phi(B)\nabla^D X_t = \Theta(B)W_t$. (Similarly for the PACFs.)

(Examples on the next slide.)

Seasonal examples

Example (AR$(1)_{12}$ model)

$$X_t - \Phi X_{t-12} = W_t.$$

If $|\Phi| < 1$, then $\gamma(0) = \sigma^2/(1 - \Phi^2)$, $\gamma(12k) = \sigma^2\Phi^k/(1 - \Phi^2)$, and $\gamma(h) = 0$ otherwise.
If $\Phi = 1$, then $\nabla_{12}X_t = W_t$ and so X_t is seasonally non-stationary: $\gamma(12k)$ will decay very slowly.

Example (MA$(1)_4$ model)

$$X_t = W_t + \theta W_{t-4}.$$

Then $\gamma(0) = \sigma^2(1 + \Phi^2)$, $\gamma(\pm 4) = \sigma^2\theta$, $\gamma(h) = 0$ otherwise.

Multiplicative seasonal ARIMA models

Definition

Let $p, d, q, P, D, Q \geq 0$, $s \geq 1$. A multiplicative seasonal ARMA model, denoted ARIMA$(p, d, q) \times (P, D, Q)_s$, is of the form

$$\Phi(B^s)\phi(B)\nabla_s^D \nabla^d X_t = \delta + \Theta(B^s)\theta(B)W_t,$$

where $\{W_t\} \sim \mathrm{WN}(0, \sigma^2)$. Here $\nabla_s^D = (1 - B^s)^D$ and $\Phi(z) = 1 - \Phi_1 z - ... - \Phi_P z^P$, $\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Example:

SARIMA data analysis recipe

- plot the data
- (potentially) do variance-stabilizing transform on the data
- (potentially) identify s
- identify D and/or d (and take appropriate differences)
- identify plausible choices (maybe multiple!) for p, q, P, Q using AICc, BIC, ACF/PACF, any other tips/tricks
- estimate parameters
- examine diagnostics (and potentially repeat the above steps until picking a best fit (or multiple fits if the data cannot distinguish)
    - plot residuals
    - ACF of residuals
    - Normal QQ-plot of residuals
    - Ljung-Box-Pierce (LBP) Q-statistics for a range of reasonable H
    - Examine significance of fitted coefficients

SARIMA
R examples

Table of Contents I

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Table of Contents II

Table of Contents III

Regression with correlated errors

Consider the regression model with r covariates,

$$Y_t = \sum_{j=1}^{r} \beta_j z_{t,j} + X_t = z_t' \beta + X_t \qquad t = 1, \ldots, n;$$

or, in matrix notation,

$$Y = Z\beta + X,$$

where $Y = (Y_1, \ldots, Y_n)'$, $X = (X_1, \ldots, X_n)'$, $\beta = (\beta_1, \ldots, \beta_r)'$, and $Z = [z_1|z_2|\cdots|z_n]'$.

Here, the covariates $z_{t,j}$ are all *fixed numbers*. (We know exactly what covariate values we will use in our model.)

In traditional regression, we assume the X_t's are i.i.d.; now, we can use time series modelling to allow them to have dependence in the form of an ARMA model.

Regression with correlated errors

The `stats::arima()` or `astsa::sarima()` functions have an argument `xreg` that takes covariates and automates fitting, using (Gaussian) maximum likelihood. You also have to input the ARMA model for the error series $\{X_t\}$.

How do we decide on the ARMA model for $\{X_t\}$? Methodology:

- regress Y on z's as if errors are uncorrelated (using OLS). Retain residuals $\hat{X}_t = Y_t - \sum_{j=1}^r \hat{\beta}_j z_{t,j}$.
- Identify the ARMA model based on the residuals.
- Then fit the MLE based on the given ARMA model (discarding the earlier fit you did).

This is repeated using different ARMA fits to generate the residuals, as necessary; IC's can/should also be compared across different fits.

Transfer modeling

In the previous "regression with correlated errors" slides, we regressed Y_t on some covariates that were prespecified. But what if we think Y_t is related to a time series Z_t but we cannot (a priori) specify which lags of Z_t are important? We might start by writing

$$Y_t = \alpha(B)Z_t + X_t$$

where Z_t, and X_t are both time series potentially having ARMA structure. And $\alpha(B) = \alpha_0 + \alpha_1 B + \alpha_2 B^2 + \dots$. We assume that the series $\{Z_t\}$ and the series of errors $\{X_t\}$ are independent of each other.

A common assumption is that $\alpha(B) = \delta(B)B^d/\omega(B)$ (where usually δ and ω are of low order, and d is just extra notation for a time shift). Then we can rewrite the equation as

$$\omega(B)Y_t = \delta(B)B^d Z_t + \omega(B)X_t.$$

So this allows Y_t to depend on itself in AR fashion as well as on Z_t. And $\omega(B)X_t$ is still ARMA, as before.

If we knew the order or form for $\alpha(B)$ we could just use our methodology for "regression with correlated errors" (MLE on ARMA parameters of Z and on coefficients in α (and σ^2) simultaneously).

How do we find a model for $\alpha(B)$?

Transfer modeling

A methodology for analysis (assuming $\alpha(B) = \delta(B)B^d/\omega(B)$):

1. First find the order of Z_t, by our usual methods (ACF/PACF, etc.), and estimate the ARMA process on Z_t, yielding $\hat{\pi}(B)$;

2. Then "multiply through" by $\hat{\pi}(B)$, yielding $\tilde{Y}_t$ and "prewhitened" Z_t's (aka W_t's; can also be computed using one step ahead residuals);

3. Use CCF of $\tilde{Y}_t$ and the prewhitened Z_t's to try to understand α (and "d"). Thus we settle on the *order* of $\omega(B)$, $\delta(B)$, and on d.

4. Then use d and the order of $\omega(B)$ and $\delta(B)$ to run regression with correlated errors (of the untransformed Y_t on the untransformed Z_t), as we learned previously.

Table of Contents I

Table of Contents II

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Table of Contents III

Spectral analysis

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Spectral analysis

Southern Oscillation Index

Recruitment

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Spectral analysis
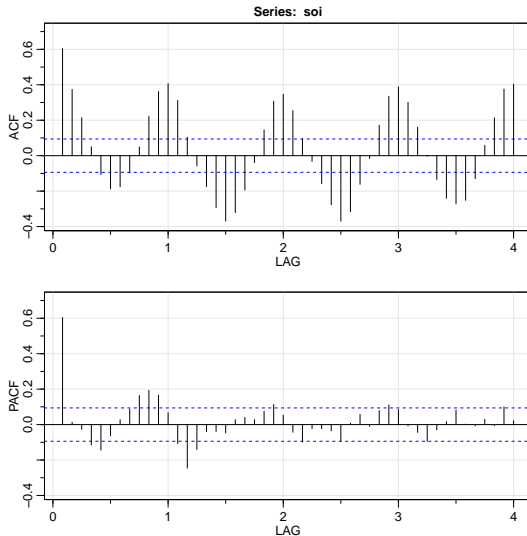
Spectral analysis

The SARIMA models are "time-domain" models. These essentially consider regression of X_t on past (earlier-in-time) values of the series.

Spectral analysis is about regressing X_t on sinusoids, which capture *periodic* or *frequency-domain* behavior of the series.

- Period: γ has period n if $\gamma(x) = \gamma(x + n)$; measured in units of time.
- Frequency: frequency is $1/$ period, measured in units of $\frac{1}{\text{time}}$.

$$\tan \theta = \frac{\sin \theta}{\cos \theta}$$
$$\sin^2 \theta + \cos^2 \theta = 1$$
$$\sin(a + b) = \sin a \cos b + \cos a \sin b$$
$$\cos(a + b) = \cos a \cos b - \sin a \sin b$$
$$e^{i\theta} = \cos \theta + i \sin \theta$$

Note: $z \mapsto e^{2\pi i \omega z}$ has frequency ω and period $1/\omega$.

Spectral analysis

Definition

If a time series $\{X_t\}$ has autocovariance γ with $\sum_{h=0}^{\infty} |\gamma(h)| < \infty$ then we define its *spectral density* to be

$$f(\nu) := \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \nu h}$$

for $-\infty < \nu < \infty$.

(This is the Fourier transform (of a discrete sequence).)

The spectral density is based on the *population* (not on a data sample).
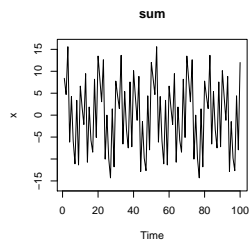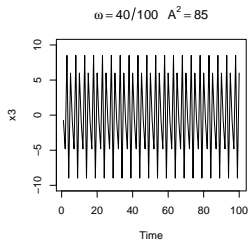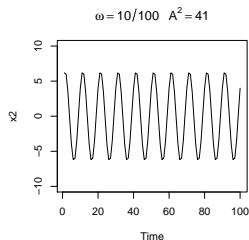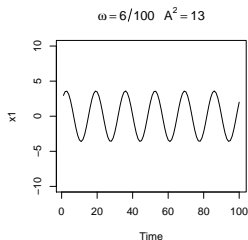
The spectral density $f(\nu)$ and the autocovariance function $\gamma(h)$ contain the same information in different formats.

Spectral analysis

Facts about the spectral density:

1. $f(\nu) < \infty$
2. $f(\nu) \geq 0$.
3. f is periodic, with period 1 (because this is true for $e^{-2\pi i \nu h}$).
   Thus, we restrict the domain of f to $[-1/2, 1/2]$.
4. f is even, i.e. $f(\nu) = f(-\nu)$.
   Thus, we often restrict the domain of f to $[0, 1/2]$.
5. $\gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i \nu h} f(\nu) d\nu$.

White noise:

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h)e^{-2\pi i \omega h} = \gamma(0) = \sigma^2.$$

Thus the spectral density is constant, equal weight to all frequencies. The name "white noise" comes from the analogy with white light, which is also a uniform mixture of all frequencies across the visible spectrum (think of a prism breaking white light into a rainbow of colors, where each color in the rainbow actually corresponds to an individual wave frequency).

Spectral analysis

Theorem

If $\{X_t\}$ is ARMA(p, q), $\phi(B)X_t = \theta(B)W_t$ then

$$f(\omega) = \sigma^2 \frac{|\theta(e^{-2\pi i \omega})|^2}{|\phi(e^{-2\pi i \omega})|^2}.$$

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Spectral analysis

AR(1): If $X_t = \phi_1 X_{t-1} + W_t$, $W_t \sim \mathrm{WN}(0, \sigma^2)$, then one can check that

$$f(\omega) = \frac{\sigma^2}{1 - 2\phi_1 \cos(2\pi\omega) + \phi_1^2}.$$

This shows that if $\phi_1 > 0$ (positive autocorrelation) the spectrum is dominated by low frequency components. This is smooth in the time domain.

If $\phi_1 < 0$ (negative autocorrelation) the spectrum is dominated by high frequency components. Rough in the time domain.
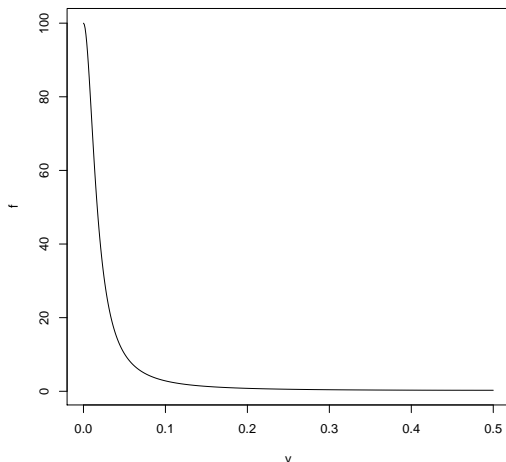
See plots in R.

Figure: AR(1) spectral density, $\phi = .9$

Spectral analysis

Figure: AR(1) spectral density, $\phi = -.9$

     Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.
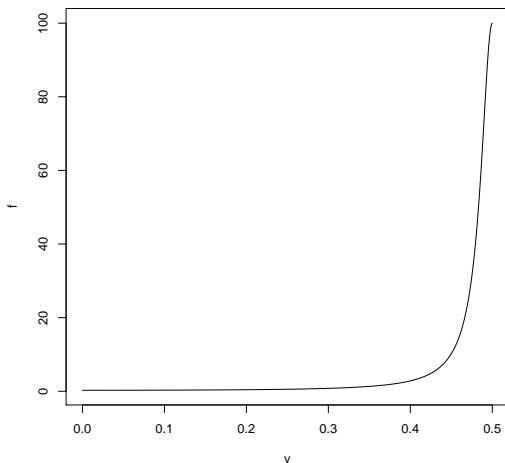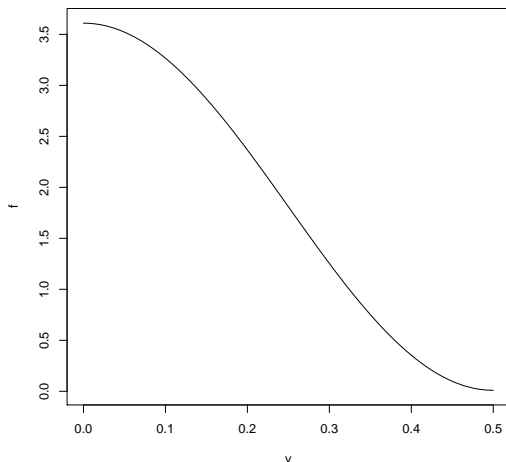
Spectral analysis

Figure: MA(1) spectral density, $\theta = .9$

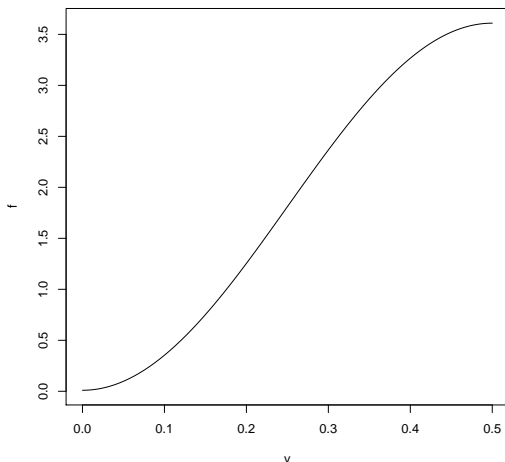Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Figure: MA(1) spectral density, $\theta = -.9$

Spectral analysis

Suppose we have a *data series* of length n, $\{X_t\}_{t=1}^n$. For simplicity, assume n is odd. We can consider a regression with X_t as the response, and n predictor variables:

$$\cos\left(\frac{1 \cdot 2\pi t}{n}\right), \sin\left(\frac{1 \cdot 2\pi t}{n}\right);$$
$$\cos\left(\frac{2 \cdot 2\pi t}{n}\right), \sin\left(\frac{2 \cdot 2\pi t}{n}\right);$$
$$\cos\left(\frac{3 \cdot 2\pi t}{n}\right), \sin\left(\frac{3 \cdot 2\pi t}{n}\right); \qquad ...,$$
$$\cos\left(\frac{[(n-1)/2]\,2\pi t}{n}\right), \sin\left(\frac{[(n-1)/2]\,2\pi t}{n}\right),$$

with t going from 1 to n, and a constant term.

(If n is even then also use a term $\cos\left(2\pi t/2\right) = (-1)^t$, but for simplicity of formulas let's assume n is odd.)

Spectral analysis

Then we have the same number of predictors as data points, and we can write *exactly* that

$$X_t = a_0 + \sum_{j=1}^{(n-1)/2} [a_j \cos(2\pi tj/n) + b_j \sin(2\pi tj/n)]$$

for some coefficients a_j, and b_j (these are the "regression coefficients").

Using standard results from simple linear regression (Ch 2 SS) $a_j = \sum_{t=1}^n X_t z_{t,j} / \sum_{t=1}^n z_{t,j}^2$ where $z_{t,j} = \cos(2\pi tj/n)$ and the same formula holds for b_j but with $z_{t,j} = \sin(2\pi tj/n)$. It can be checked that $\sum_{t=1}^n z_{t,j}^2 = n/2$ in either case if $j \neq 0, 1/2$, so that

$$a_j = \frac{2}{n} \sum_{t=1}^n X_t \cos(2\pi tj/n) \quad \text{and} \quad b_j = \frac{2}{n} \sum_{t=1}^n X_t \sin(2\pi tj/n)$$

Spectral analysis

Definition

The *scaled periodogram* is then $P(j/n) = a_j^2 + b_j^2$.

- It indicates which frequencies are large in magnitude.
- It is the sample variance at the given frequency component.
- The (scaled) periodogram is related to the *discrete Fourier transform* which can be computed quickly by the *fast Fourier transform* (algorithm).

Spectral analysis

Definition

For a *data sample*, the *discrete Fourier transform (DFT)* is $d(\cdot)$ where

$$d(j/n) = n^{-1/2} \sum_{t=1}^{n} X_t e^{-2\pi i t j/n}$$

$$= n^{-1/2} \sum_{t=1}^{n} X_t \left(\cos(2\pi t j/n) - i \sin(2\pi t j/n) \right)$$

The scaled periodogram satisfies $P(j/n) = (4/n)|d(j/n)|^2$.

Definition

The *periodogram* is $I(j/n) := |d(j/n)|^2 = P(j/n)n/4$.

The periodogram is basically the "sample spectral density".

Definition

Given data points $X_1, \ldots, X_n$, recall the DFT is

$$d(j/n) \equiv d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} X_t \cos(2\pi\omega_j t) - iX_t \sin(2\pi\omega_j t)$$

where $\omega_j := j/n, \ j = 0, 1, \ldots, n-1$ are the *Fourier frequencies* or the *fundamental frequencies.* Also define the cosine and sine transform,

$$d_c(\omega_j) := n^{-1/2} \sum_{t=1}^{n} X_t \cos(2\pi\omega_j t) \text{ and } d_s(\omega_j) := n^{-1/2} \sum_{t=1}^{n} X_t \sin(2\pi\omega_j t).$$

Thus $d(\omega_j) = d_c(\omega_j) - id_s(\omega_j)$ and the periodogram is
$I(\omega_j) = d_c^2(\omega_j) + d_s^2(\omega_j)$.

Spectral analysis

As mentioned before, $d(\omega_j)$ can be computed rapidly by the *fast Fourier transform (FFT)*; this is especially true if n is a *highly composite integer* (has many factors).

Can/should always check output of functions (e.g., `spec.pgram()`, `fft()`) to see which n was actually used. The function `nextn()` tells, for a given sample size n, the value n' that will actually be used. (That is $n' = $ `nextn(n)`.)

One always has to be careful of the defined scaling of a given FFT implementation and of $d(j/n)$!

Best way to be sure you understand the definitions is to play with them in R, which we do now. We can use the `fft()` or `spec.pgram()` functions (or, `spectrum()` which just calls `spec.pgram()`) to compute the periodogram.

See R code.

Spectral analysis

Periodogram

> ## Theorem
>
> *For any fixed distinct m frequencies $\omega_k, k = 1, \ldots, m$, and any fundamental frequencies $\omega_{k:n} \to \omega_k$ as $n \to \infty$, then under appropriate assumptions,*
>
> $$\frac{2I(\omega_{k:n})}{f(\omega_k)} \to^d \chi_2^2.$$
>
> *The joint limits of $I(\omega_{k:n})$ and $I(\omega_{j:n})$, $j \neq k$, are independent.*

The assumptions are: we can write X_t as

$$X_t := \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}, \qquad \text{where } \sum_{j=-\infty}^{\infty} |\psi_j| < \infty,$$

where $W_t \overset{\text{iid}}{\sim} (0, \sigma^2)$. Assume $\gamma(h)$ satisfies $\sum_{h=-\infty}^{\infty} |h||\gamma(h)| < \infty$. We assume $f(\omega_k) > 0$, for $k = 1, \ldots, m$.

Spectral analysis
Smoothing the periodogram

A key idea: the periodogram asymptotic result said that different periodogram values were approximately independent. Therefore we can average them together, like a sample mean, to reduce variance.

Definition (Averaged/smoothed periodogram)

Fix m. Let $L = 2m + 1$. Let

$$\bar{f}(\omega_j) := \frac{1}{L} \sum_{k=-m}^{m} I(\omega_{j+k}).$$

be the averaged (or smoothed) periodogram of a stationary linear time series.

Idea: need $L \to \infty$ as $n \to \infty$, but $L/n \to 0$, in such a way that $f(\omega_{j+k}) \approx f(\omega_j)$.

Theorem (Brockwell Davis; pp 351)

Let X_t be a stationary linear process based on iid white noise with finite fourth moment. If $n \to \infty$, $L \to \infty$, $L/n \to 0$, then[a] $\bar{f}(\omega) \to_p f(\omega)$.

[a]under a condition on the linearity coefficients ψ_j: $\sum |\psi_j||j|^{1/2} < \infty$.

We let $B = L/n$ be the "bandwidth". Actually, `spec.pgram()` returns as bandwidth the value $B\sqrt{12} = \sqrt{12}L/n$.[14]

---

[14]See footnote 9 page 190, SS, 4ed

We then also have that

$$2L\bar{f}(\omega)/f(\omega) \text{ is approximately } \sim \chi^2_{2L}.$$

We can then form confidence intervals. For notational simplicity assume ω is a fundamental frequency. A CI for $f(\omega)$ is

$$\left[\frac{2L\bar{f}(\omega)}{\chi^2_{2L,1-\alpha/2}}, \frac{2L\bar{f}(\omega)}{\chi^2_{2L,\alpha/2}} \right]$$

or taking logs, a CI for $\log f(\omega)$ is

$$\left[\log \bar{f}(\omega) - a_L, \log \bar{f}(\omega) + b_L \right]$$

where $a_L = -\log 2L + \log \chi^2_{2L,1-\alpha/2}$, $b_L = \log 2L - \log \chi^2_{2L,\alpha/2}$.

If we use the FFT with a length n' padded sequence of X, then we have to replace the 'degrees of freedom' parameter $2L$ with $d = 2Ln/n'$. So, e.g., for the CI for $\log f(\omega)$, we have

$$a_L = -\log d + \log \chi^2_{d,1-\alpha/2} \qquad b_L = \log d - \log \chi^2_{d,\alpha/2}.$$

Spectral analysis

See R code.

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

What if we want to form confidence intervals for two parameters, $f(\omega_1)$ and $f(\omega_2)$ simultaneously, but we want to say that the intervals *both* cover the truth with say 95% confidence.

We have seen before the concept of *leakage* related to having fundamental frequencies not match the frequencies present in our data.

Tapering counters this effect by replacing our data X_t by $Y_t := h_t X_t$ for some taper sequence h_t (fixed, known numbers that we decide on).

See R.

Imagine there is an (unobserved) continuous signal C_t, for all $t \geq 0$. We observe it at n discrete time points $\Delta, 2\Delta, ..., n\Delta$, for some sampling interval $\Delta > 0$.

The *Nyquist frequency* is

$$Q := \frac{1}{2\Delta}$$

where Δ is the *sampling interval (length)* and $1/\Delta$ is the *sampling frequency*. Thus 2Δ corresponds to a cycle every two data points. We need Q *larger than any frequency components in the continuous signal* C_t.

One important use for spectra is for regression of one series on another. We start with the idea of *coherency* which is like correlation (and correlation is related to the regression coefficient in simple linear regression).

We assume we have two stationary time series, $\{X_t\}$, $\{Y_t\}$. Recall the cross-covariance function, $\gamma_{xy}(h) := E(X_{t+h} - \mu_X)(Y_t - \mu_Y)$.

Definition

Define[a] the *cross-spectrum* $f_{xy}(\omega)$ by

$$f_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) e^{-2\pi i \omega h}, \quad -1/2 \leq \omega \leq 1/2.$$

[a]Assuming the cross-covariance function is absolutely summable.

Unlike the spectrum $f_x(\omega)$ or $f_y(\omega)$, the cross spectrum is complex-valued not real-valued, and we sometimes write $f_{xy}(\omega) =: c_{xy}(\omega) - iq_{xy}(\omega)$ where c_{xy} is the cospectrum and q_{xy} is the quadspectrum.

Note $f_{xy}(\omega) = \overline{f_{xy}(\omega)}$.

Define

$$\rho^2_{xy}(\omega) := \frac{|f_{xy}(\omega)|^2}{f_x(\omega)f_y(\omega)}.$$

Then $0 \le \rho^2_{xy}(\omega) \le 1$. Also, $\rho^2_{xy}(\omega) = 1$ for all ω if and only if one series is a linear combination ("filter") of the other, with no added noise.

We can estimate coherence by

$$\hat{\rho}_{yx}^2(\omega) = \frac{|\hat{\hat{f}}_{yx}(\omega)|^2}{\hat{f}_x(\omega)\hat{f}_y(\omega)}$$

(Here the estimators may all be averaged or kernel-smoothed. For simplicity, we use the same L and same kernel weights for the three estimators.

One can test if $\rho_{yx}^2(\omega) = 0$ or not using:

$$F = \frac{\hat{\rho}_{yx}^2(\omega)}{1 - \hat{\rho}_{yx}^2(\omega)}(L-1)$$

which is approximately $F_{2,df-2}$ where $df = 2Ln/n'$ (where n' is the padded sample size the fast Fourier transform uses).

Consider two series, stationary, and mean 0 for simplicity, Y_t, X_t. Consider

$$Y_t = \sum_{r=-\infty}^{\infty} \beta_r X_{t-r} + \nu_t$$

where ν_t is a stationary noise process. We want to estimate the β_r's.

$\hat{B}(\omega_k) = \frac{\hat{f}_{yx}(\omega_k)}{\hat{f}_{xx}(\omega_k)}$, $\omega_k = k/M$, with M an even number $M \ll n$.

Then $\hat{\beta}_t = M^{-1} \sum_{k=0}^{M-1} \hat{B}(\omega_k) e^{2\pi i \omega_k t}$, $t = 0, \pm 1, \pm 2, \ldots, \pm(M/2 - 1)$. We take $\hat{\beta}_t = 0$ for $|t| \geq M/2$.

We may additionally threshold to 0 (set to 0) coefficient values that are small.

Table of Contents I

Table of Contents II

Companion slides for STAT 5511 lectures. Not a substitute for attending lecture.

Table of Contents III

The ARMA(p, q) model is $\phi(B)X_t = \theta(B)W_t$ where $\{W_t\} \sim \mathrm{WN}(0, \sigma^2)$.

Table of Contents I

Table of Contents II

Table of Contents III