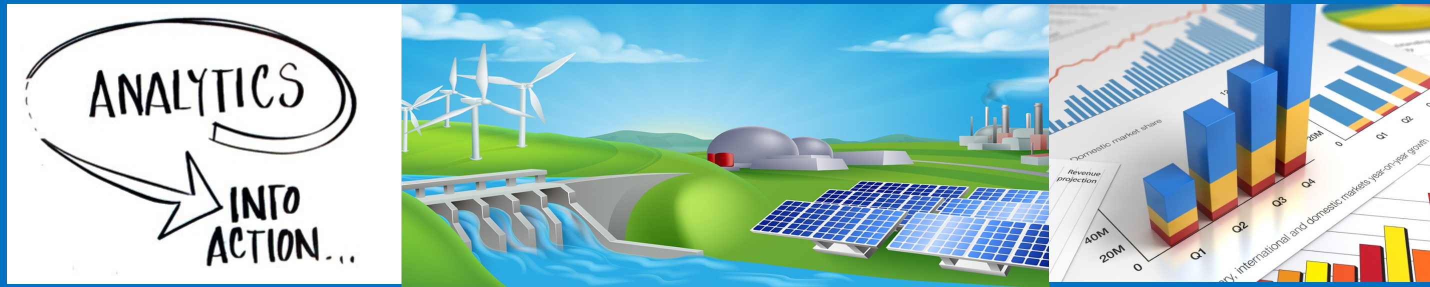


ENV 797 - TIME SERIES ANALYSIS FOR ENERGY AND ENVIRONEMNT APPLICATIONS

M5 – ARIMA Models

Prof. Luana Medeiros Marangon Lima, Ph.D.

Nicholas School of the Environment - Duke University

Learning Goals

- Discuss Models for Stationary Time Series
  - Autoregressive Model (AR)
  - Moving Average Model (MA)
  - ARMA Model
  - ARIMA Model
- Learn how to implement those models in R

What do we know so far?

Filtering and Preparation

- Initial Plots
- Trend
- Seasonality
- Stationarity Test
- Outliers
- Missing Data

Estimation

- Auto Correlation Function
- **Partial Autocorrelation Function**
- **Model Parameter estimation**

Forecasting

- Forecast
- Model accuracy
- Model Selection

Introduction

- Basic concepts of parametric time series models the ARMA or ARIMA models
  - AR stands for Auto Regressive; and
  - MA stands for Moving Average
  - And the I stands for Integrated (more on that later)
- Traditional Box-Jenkins models
- To model a time series with the Box-Jenkins approach, the series has to be stationary
- Recall: series is stationary if tends to wonder more or less uniformly about some fixed level

Review: Achieving Stationarity

- Is the trend stochastic or deterministic?
  - Run the tests
  - If stochastic: use differencing
  - If determinist: use regression
- Check if variance changes with time
  - If yes: make it constant with log transformation

AR models

Auto Regressive Models

- The simplest family of these models are the autoregressive (AR)
- They generalize the idea of regression to represent the linear dependence between a dependent variable y_t and an explanatory variable y_{t-1}, such that:

$$y_t = \underset{\beta_0}{c} + \underset{\beta_1}{\phi} y_{t-1} + a_t$$

where c and ϕ are constants to be determined and a_t are i.i.d. $N(0, \sigma^2)$

First order autoregressive process

Auto Regressive Models

- From the unit root test, the condition $-1 < \phi < 1$ is necessary for the process to be stationary, **but why?**
- Suppose $y_o = h$ where h is constant

$$y_1 = c + \phi h + a_1$$

$$y_2 = c + \phi y_1 + a_2 = c + \phi(c + \phi h + a_1) + a_2 = c(1+\phi) + \phi^2 h + \phi a_1 + a_2$$

$$y_3 = c(1 + \phi + \phi^2) + \phi^3 h + \phi^2\, a_1 + \phi a_2 + a_3$$

General Form

$$\boxed{y_t = c\sum_{i=0}^{t-1}\phi^i + \phi^t h + \sum_{i=0}^{t-1}\phi^i\, a_{t-i}}$$

$$E[a_t] = 0 \Rightarrow E[y_t] = c\sum_{i=0}^{t-1}\phi^i + \phi^t h$$

Auto Regressive Models

- Hence the process is stationary if this function does not depend on t

$$E[y_t] = c\sum_{i=0}^{t-1}\phi^i + \phi^t h$$

The first term is a geometric progression with ratio ϕ, thus

$$\sum_{i=0}^{t-1}\phi^i \approx \frac{1-\phi^{t-1}}{1-\phi} \approx \frac{1}{1-\phi}\ if\ |\phi|<1$$

Second term needs to converge to zero, this is only true if

$$|\phi|<1$$

Review: Geometric Progression

- Sequence of numbers where each term is found by multiplying the previous one by a fixed ratio

  Ex.: $a, ar, ar^2, ar^3, ar^4, ar^5$,… where $r \neq 0$

- The sum of the first n element of a geometric progression is given by

$$\sum_{k=1}^{n} ar^{k-1} = a\sum_{k=1}^{n} r^{k-1} = a\frac{(1-r^n)}{1-r}$$

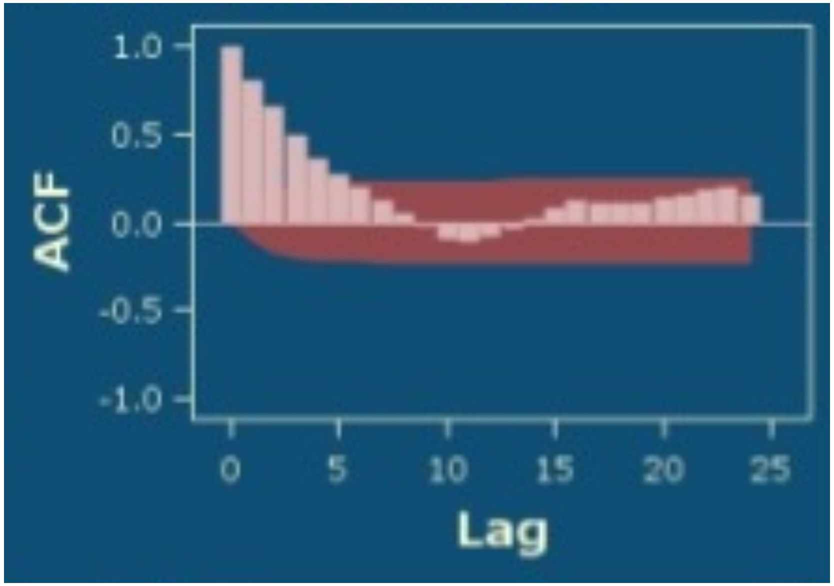
Auto Regressive Models (cont'd)

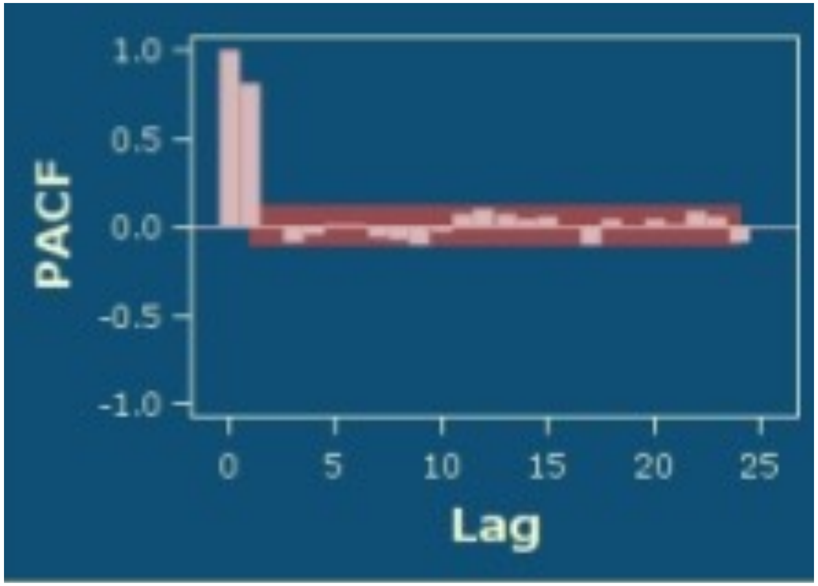
- This linear dependence can be generalized so that the present value of the series, y_t, depends not only on y_{t-1}, but also on the previous p lags, $y_{t-2} \dots, y_{t-p}$
- Thus, AR process of order p is obtained

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + a_t$$

ACF and PACF for AR Process

- For AR models ACF will decay exponentially with time
- The PACF will identify the order of the AR model





p = 1

MA models

Moving Average Models

- The AR process have infinite non-zero autocorrelation coefficients that decay with the lag
- Therefore, we say AR processes have a relatively **"long memory"**
- There is another family of model, that have a **"short memory"**, the moving average or MA process
- The MA processes are a function of a finite and generally small number of its past residuals

Moving Average Models

- A first order moving average process MA(1), is defined by

$$y_t = \mu + a_t - \theta a_{t-1}$$

where μ is the process mean and a_t are i.i.d. $N(0, \sigma^2)$

- Or

$$\tilde{y}_t = a_t - \theta a_{t-1} \quad \text{where} \quad \tilde{y}_t = y_t - \mu$$

- Note: This process will always be stationary for any value of θ

MA(q) Process Basic Concepts

- A q-order moving average process, denoted MA(q) takes the form

$$y_t = \mu + a_t - \theta_1 a_{t-1} - \cdots + \theta_q a_{t-q}$$

- Assume that error terms are i.i.d (independent and identically distributed)

$$a_i \sim N(0, \sigma)$$

mean μ

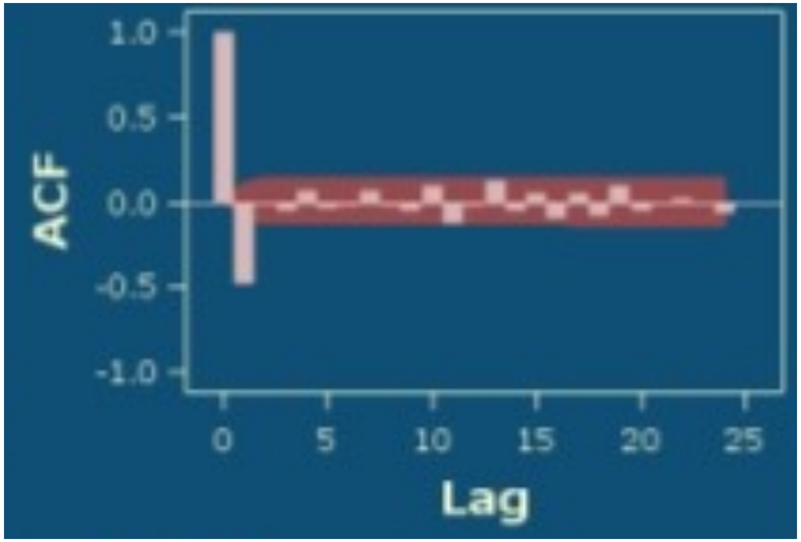
standard deviation

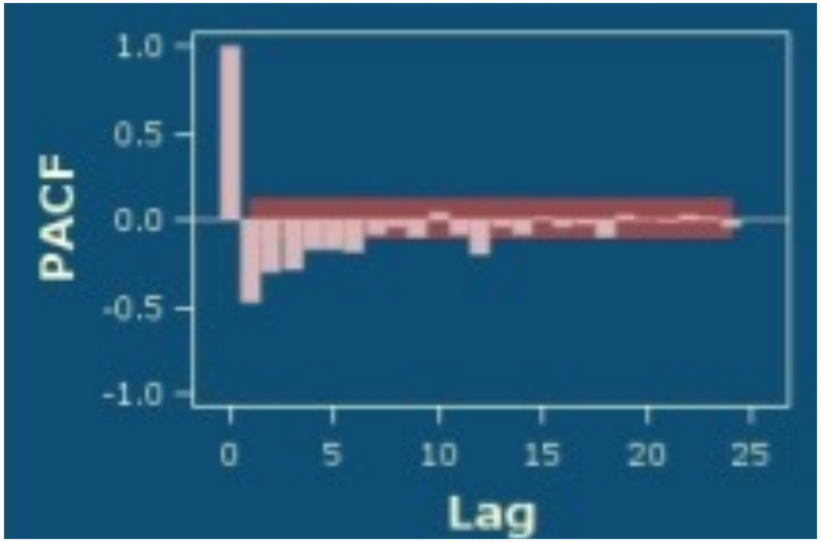
$$cov\left(a_i, a_j\right) = 0 \quad if\ i \neq j$$

$$cov\left(a_i, a_j\right) = \sigma^2 \quad if\ i = j$$

ACF and PACF for MA Process

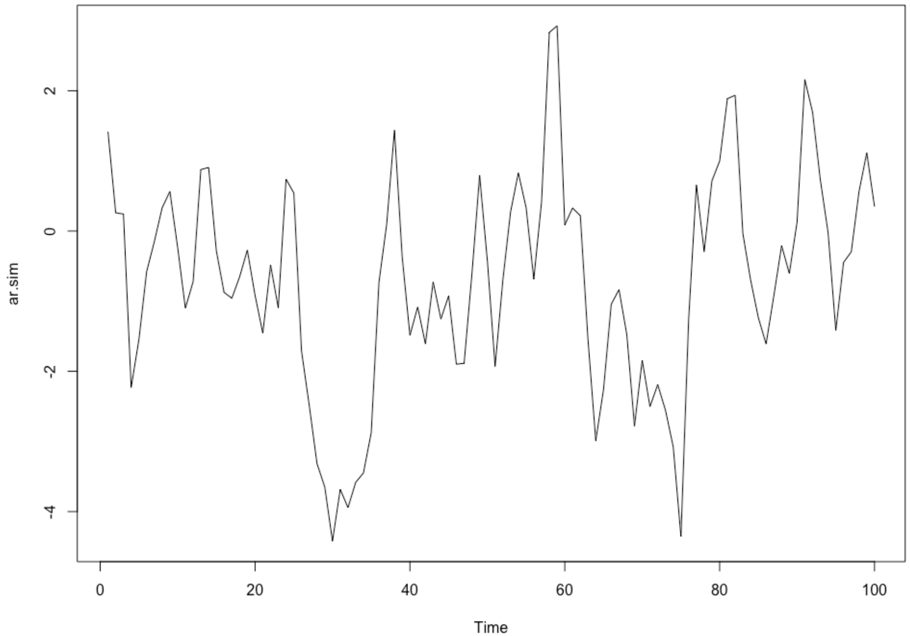
- For MA models ACF will identify the order of the MA model
- The PACF will decay exponentially

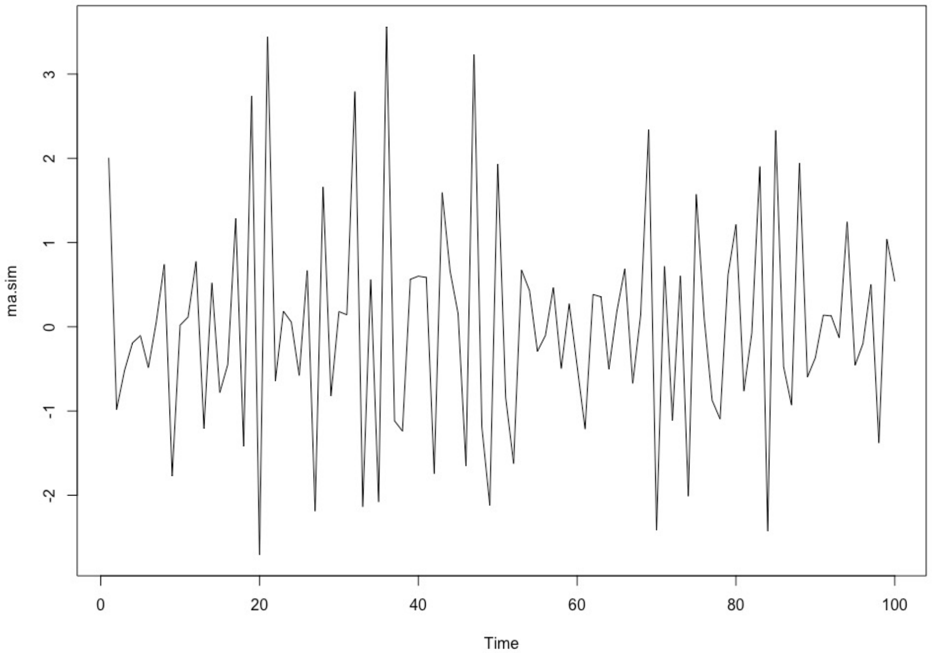




q = 1

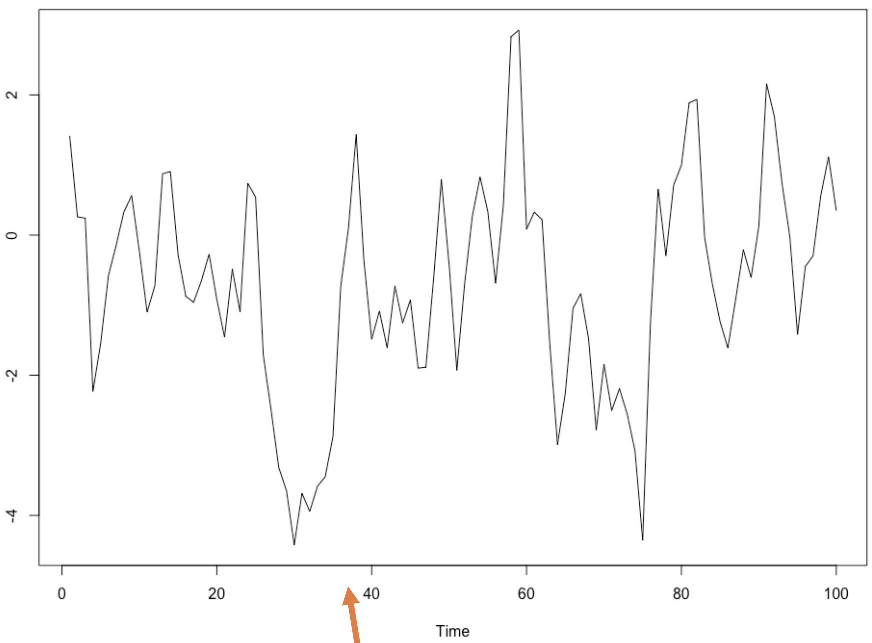
AR vs MA - Comparing Series Plots





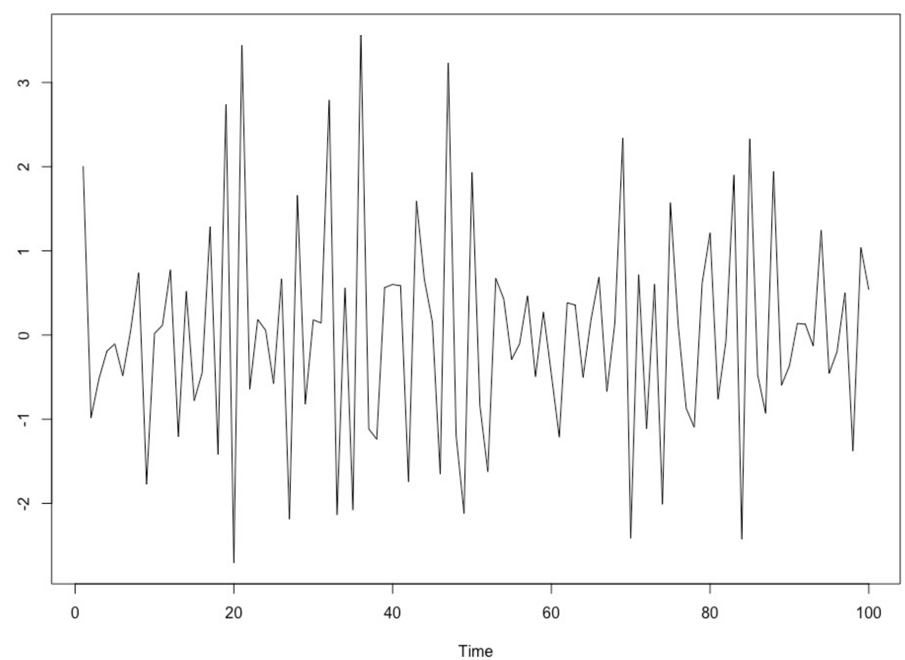
AR vs MA - Comparing Series Plots

AR

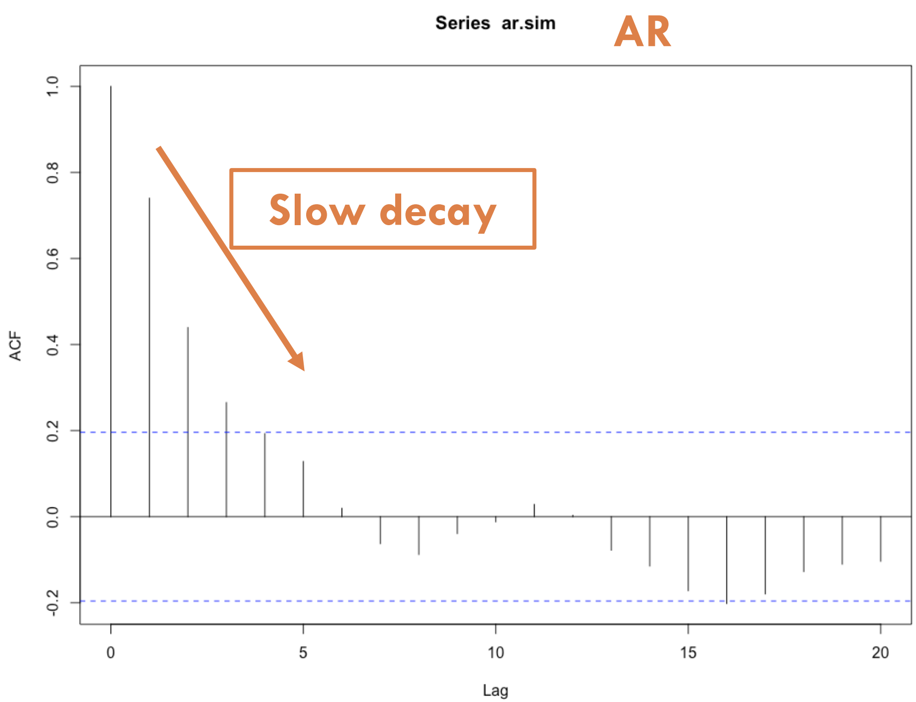


Dependency on previous observations

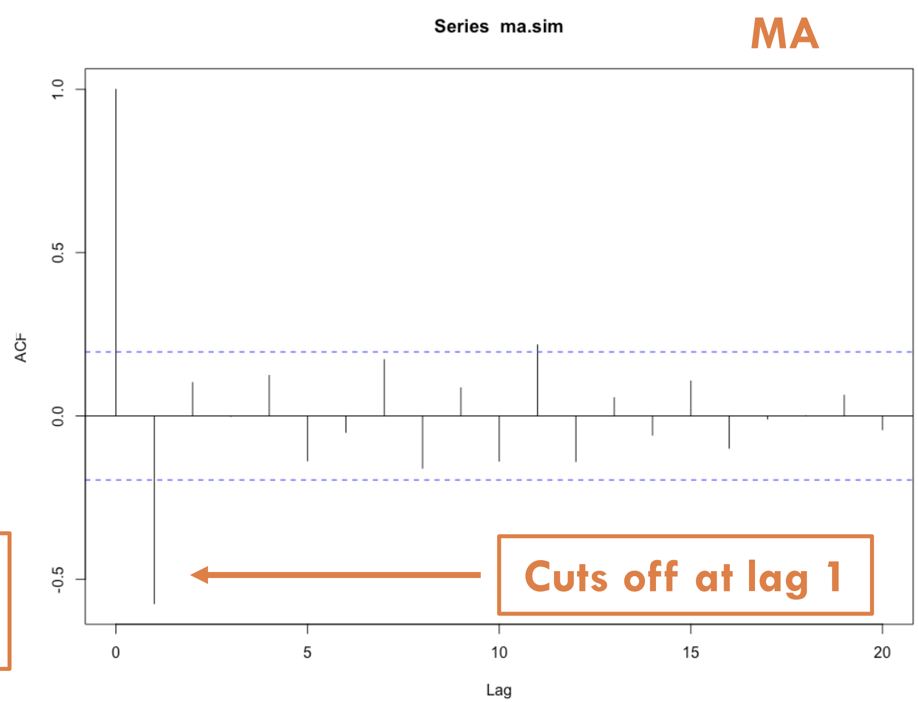
MA



AR vs MA - Comparing ACF Plots

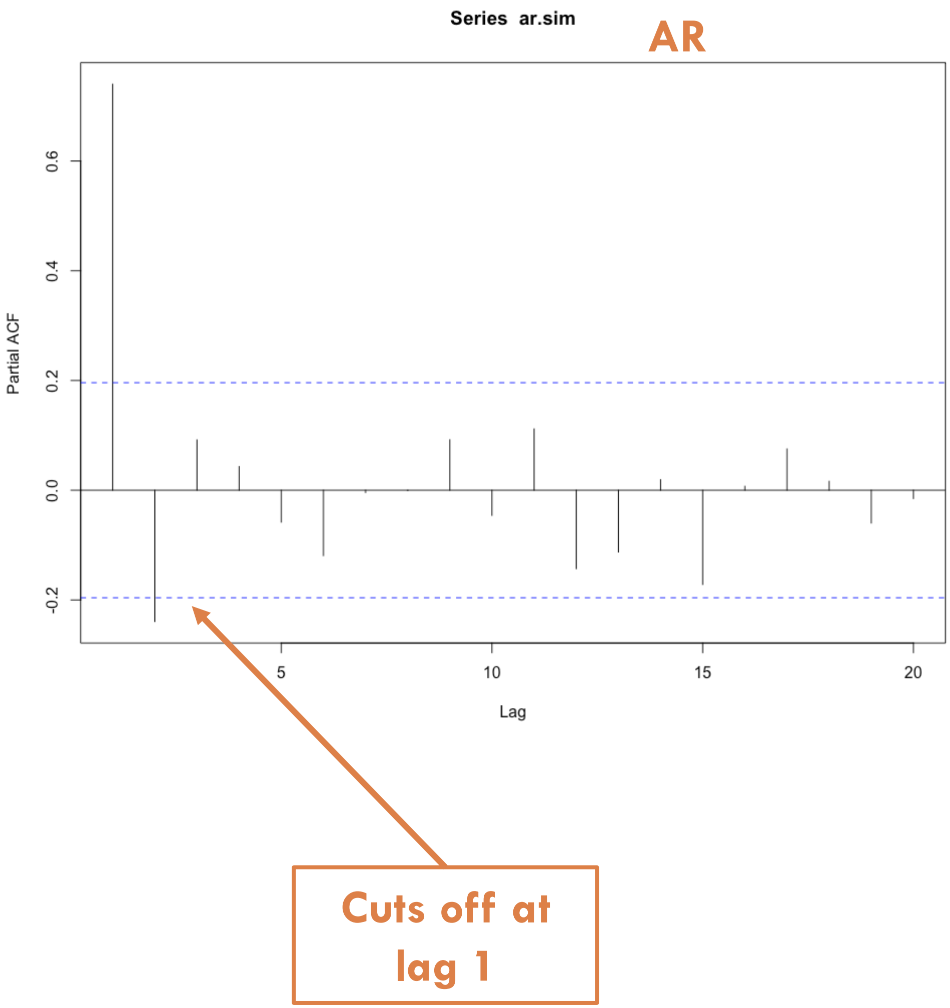


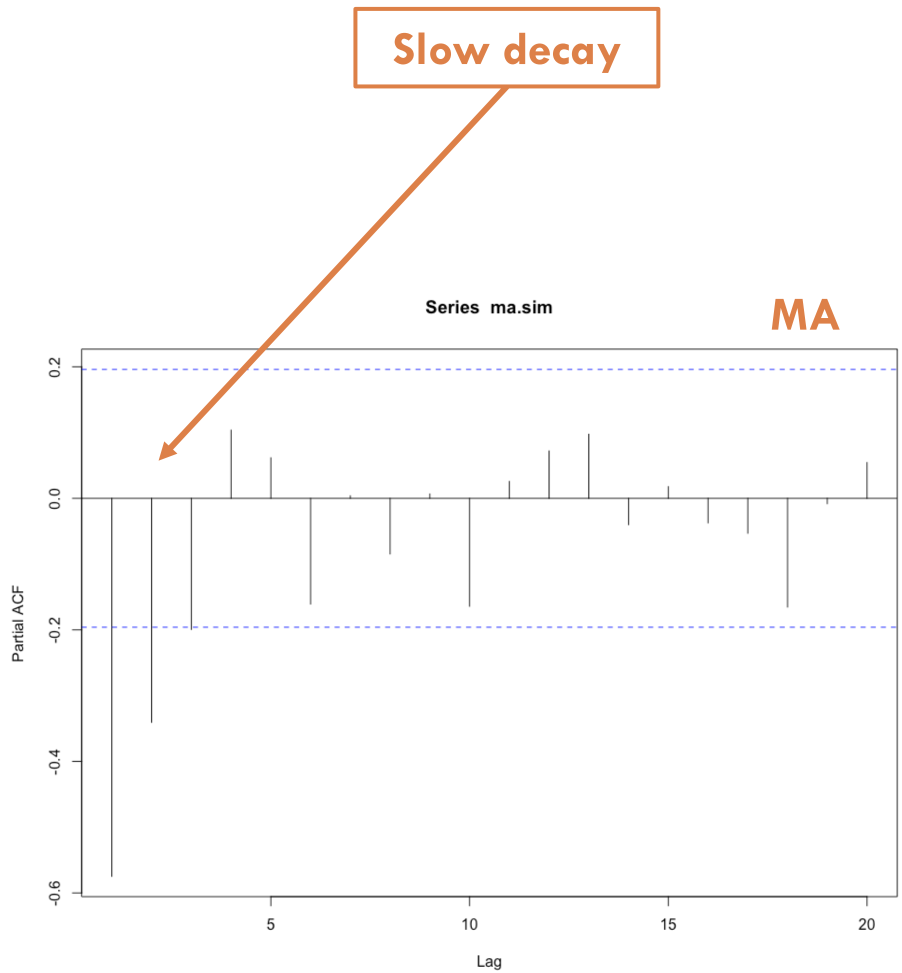
Often if the stationary series has positive autocorrelation at lag 1 AR terms work best



Often if it has negative autocorrelation at lag 1, MA terms work best

AR vs MA - Comparing PACF Plots





In summary…

- **AR Process**
  - Series current values depend on its own previous values
  - AR(p) – current value depend on its own p-previous values
  - p is order of the AR process
- **MA Process**
  - The current deviation from mean depends on previous deviations
  - MA(q) – current deviation depends on q-previous deviations
  - q is the order of the MA process
- But we can also have **ARMA Process**
  - Takes into account both of the above factors when making predictions

ARMA models

ARMA Process

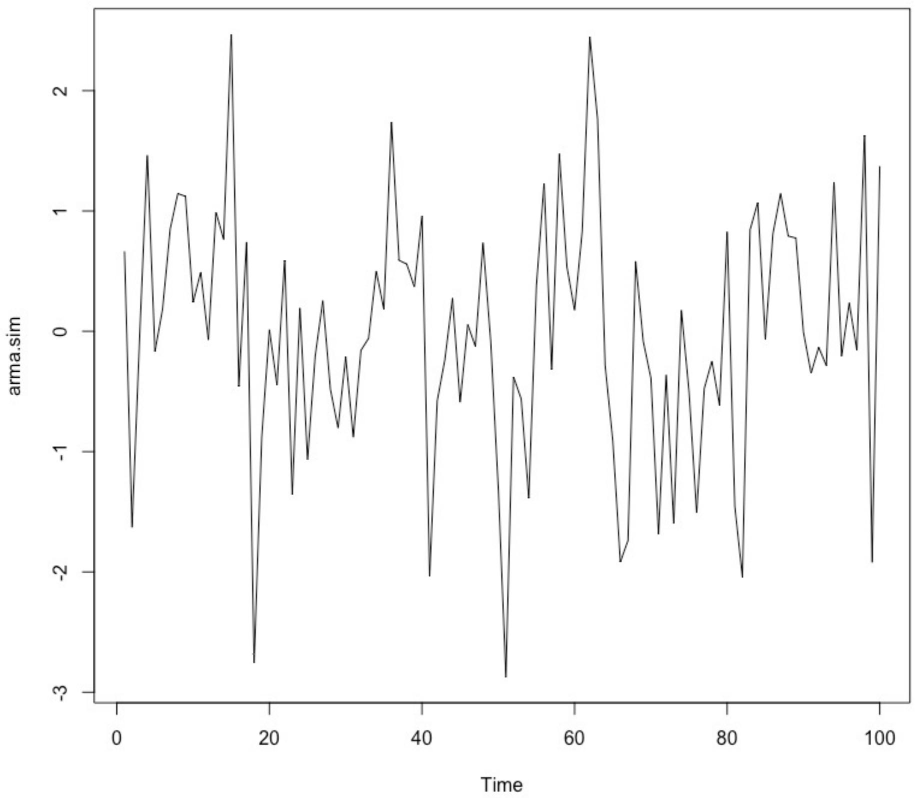
- The simplest process, the ARMA(1,1) is written as

$$\tilde{y}_t = \phi_1 \tilde{y}_{t-1} + a_t - \theta_1 a_{t-1}$$

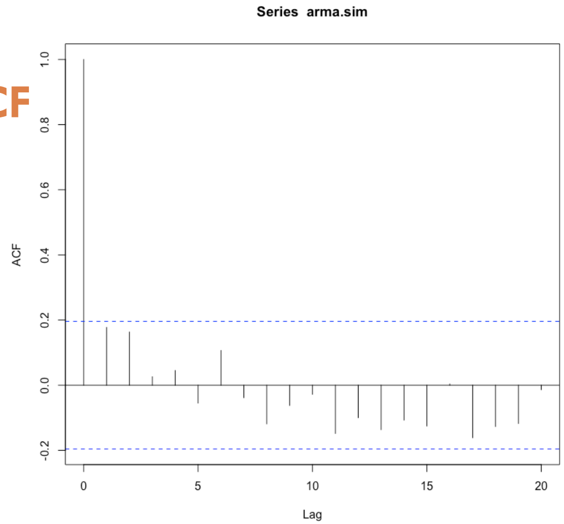
where $|\phi_1| < 1$ for the process to be stationary

- The ACF and PACF of the ARMA processes are the result of superimposing the AR and MA properties
  - In the ACF initial coefficients depend on the MA order and later a decay dictated by the AR part
  - In the PACF initial values dependent on the AR followed by the decay due to the MA part

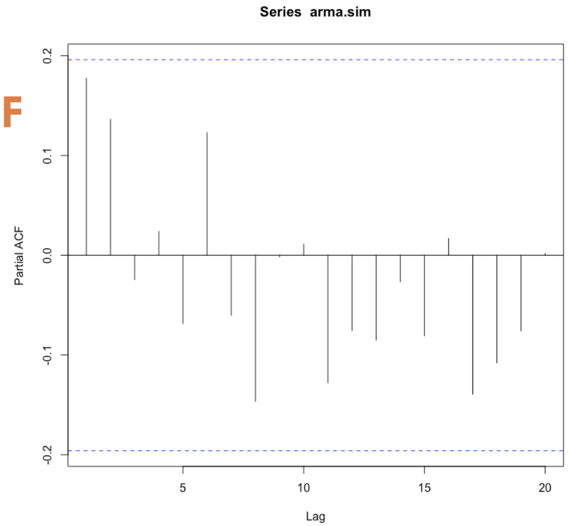
ARMA Model Plots



ACF



PACF



ARIMA models

ARIMA Models

- **A**uto-**R**egressive **I**ntegrated **M**oving **A**verage
- We know the AR and MA part already
- The Integrated part refers to a series that needs to be differenced to achieve stationarity
- The non-seasonal ARIMA model is described by three numbers

ARIMA$(\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{q})$

p: $number\ of\ autoregressive\ terms$
d: $number\ of\ differences$ (non-seasonal)
q: $number\ of\ moving\ average\ terms$

ARIMA Models

- Equation

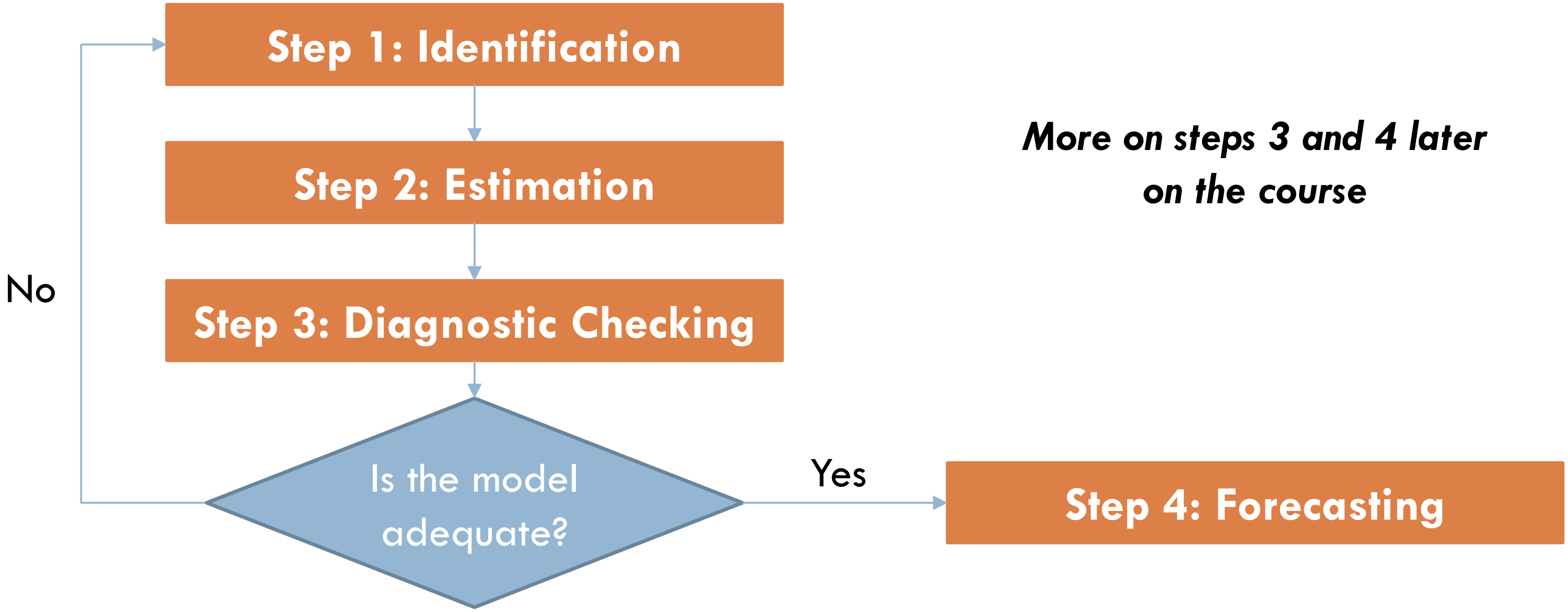
$$\hat{y}_t = \underset{\text{constant}}{\mu} + \underbrace{\phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}}_{\text{AR terms convention (+)}} + \underset{\text{Error term}}{a_t} \underbrace{- \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}}_{\text{MA terms convention (-)}}$$

- $\hat{y}_t$ is an estimate for the differenced version of the series therefore

$$If\ d = 0: \quad \hat{Y}_t = \hat{y}_t$$
$$If\ d = 1: \quad \hat{Y}_t = \hat{y}_t + Y_{t-1}$$
$$\vdots$$

Drawbacks

- There is no systematic approach for identification and selection
- The identification is mainly trial-and-error



More on steps 3 and 4 later on the course

ARIMA class models in R

Fit ARIMA Models in R

- *arima()* from package "stats"

```
arima(x, order = c(0L, 0L, 0L),
      seasonal = list(order = c(0L, 0L, 0L), period = NA),
      xreg = NULL, include.mean = TRUE,
      transform.pars = TRUE,
      fixed = NULL, init = NULL,
      method = c("CSS-ML", "ML", "CSS"), n.cond,
      SSinit = c("Gardner1980", "Rossignol2011"),
      optim.method = "BFGS",
      optim.control = list(), kappa = 1e6)
```

Arguments

Most relevant arguments

`x`	a univariate time series
`order`	A specification of the non-seasonal part of the ARIMA model: the three integer components *(p, d, q)* are the AR order, the degree of differencing, and the MA order.
`seasonal`	A specification of the seasonal part of the ARIMA model, plus the period (which defaults to `frequency(x)`). This should be a list with components `order` and `period`, but a specification of just a numeric vector of length 3 will be turned into a suitable list with the specification as the `order`.
`xreg`	Optionally, a vector or matrix of external regressors, which must have the same number of rows as `x`.
`include.mean`	Should the ARMA model include a mean/intercept term? The default is `TRUE` for undifferenced series, and it is ignored for ARIMA models with differencing.

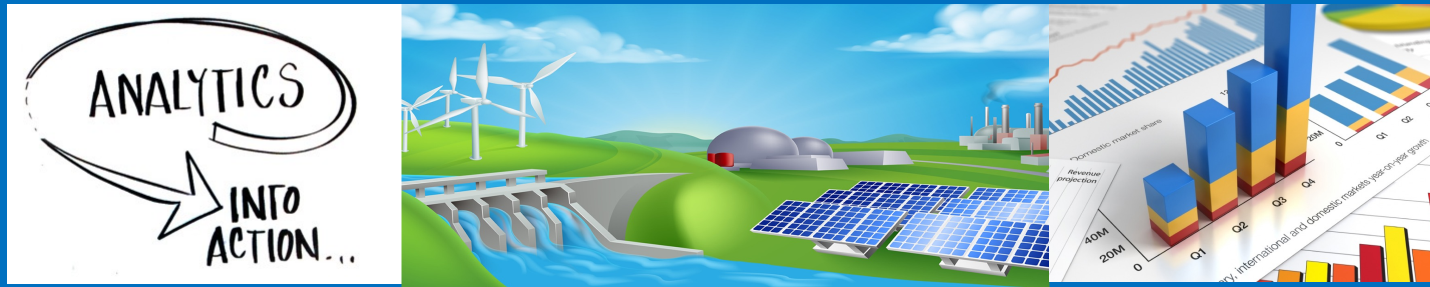
Simulate ARIMA Models in R

- arima.sim() from package “stats”

```
arima.sim(model, n, rand.gen = rnorm, innov = rand.gen(n, ...),
          n.start = NA, start.innov = rand.gen(n.start, ...),
          ...)
```

Arguments

`model`	A list with component `ar` and/or `ma` giving the AR and MA coefficients respectively. Optionally a component `order` can be used. An empty list gives an ARIMA(0, 0, 0) model, that is white noise.
`n`	length of output series, before un-differencing. A strictly positive integer.
`rand.gen`	optional: a function to generate the innovations.
`innov`	an optional times series of innovations. If not provided, `rand.gen` is used.
`n.start`	length of ‘burn-in’ period. If `NA`, the default, a reasonable value is computed.



THANK YOU !

luana.marangon.lima@duke.edu