Time Series Analysis

Lecture 1: Introduction

Tohid Ardeshiri

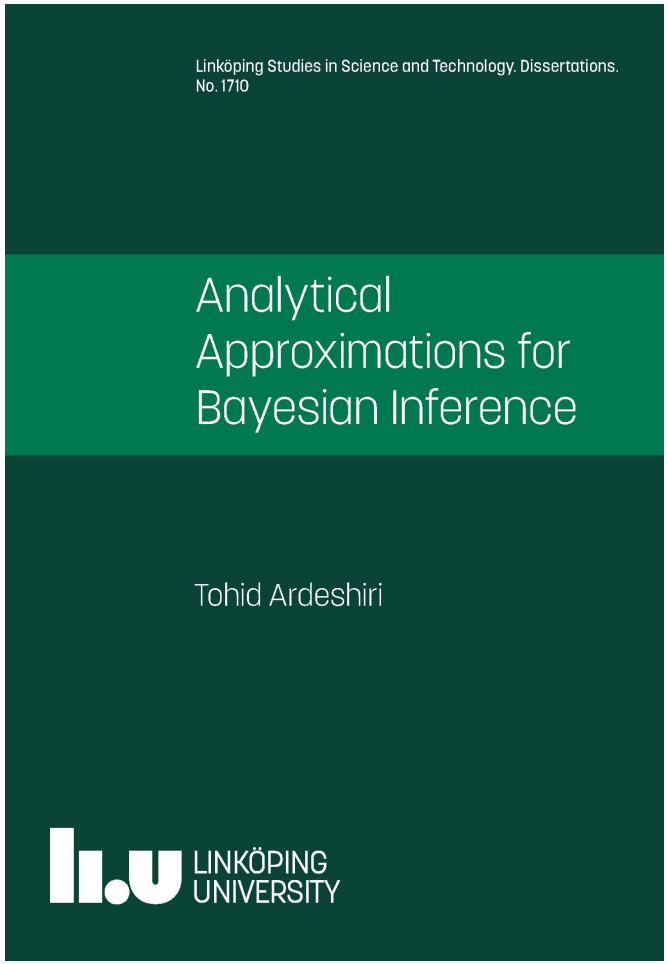
Linköping University
Division of Statistics and Machine Learning

September 2, 2019

Course teacher

Tohid Ardeshiri
PhD in 2015 in Bayesian inference

Senior Data Scientist at
Qamcom Research & Technology AB



Bayesian Inference

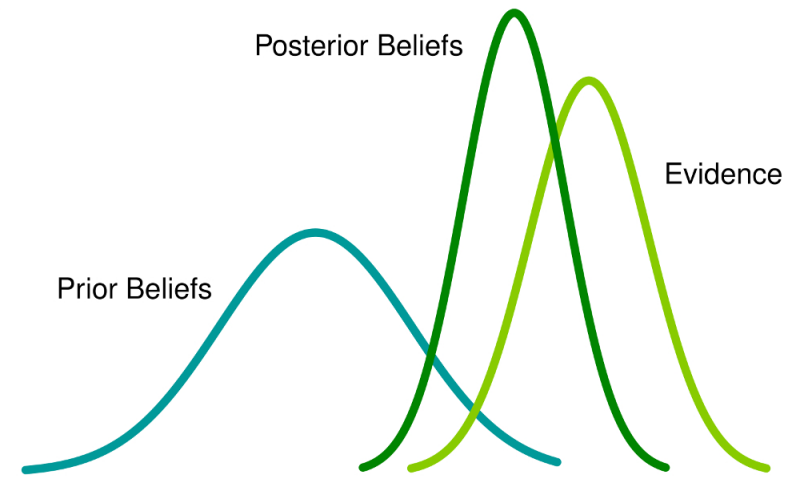
Bayesian inference is a means of combining prior beliefs with the data (evidence) to obtain posterior beliefs.

Example: Parameter learning

$$f(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})$$

Probability Calculus

$$f(\boldsymbol{\theta}, \mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})$$
$$f(\boldsymbol{\theta}, \mathbf{x}) = f(\boldsymbol{\theta}|\mathbf{x})f(\mathbf{x})$$

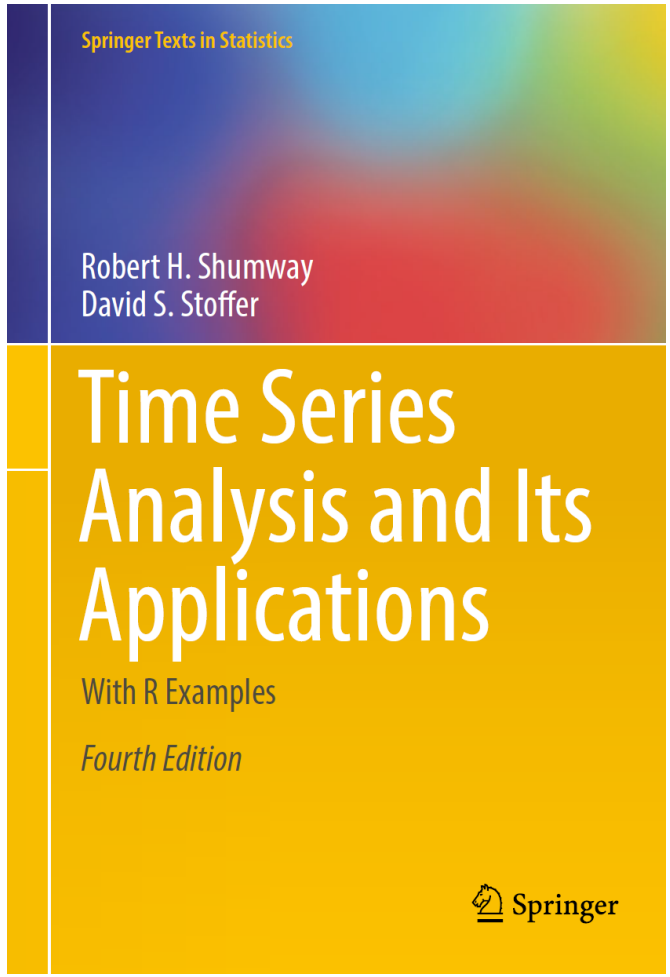


Course literature and software

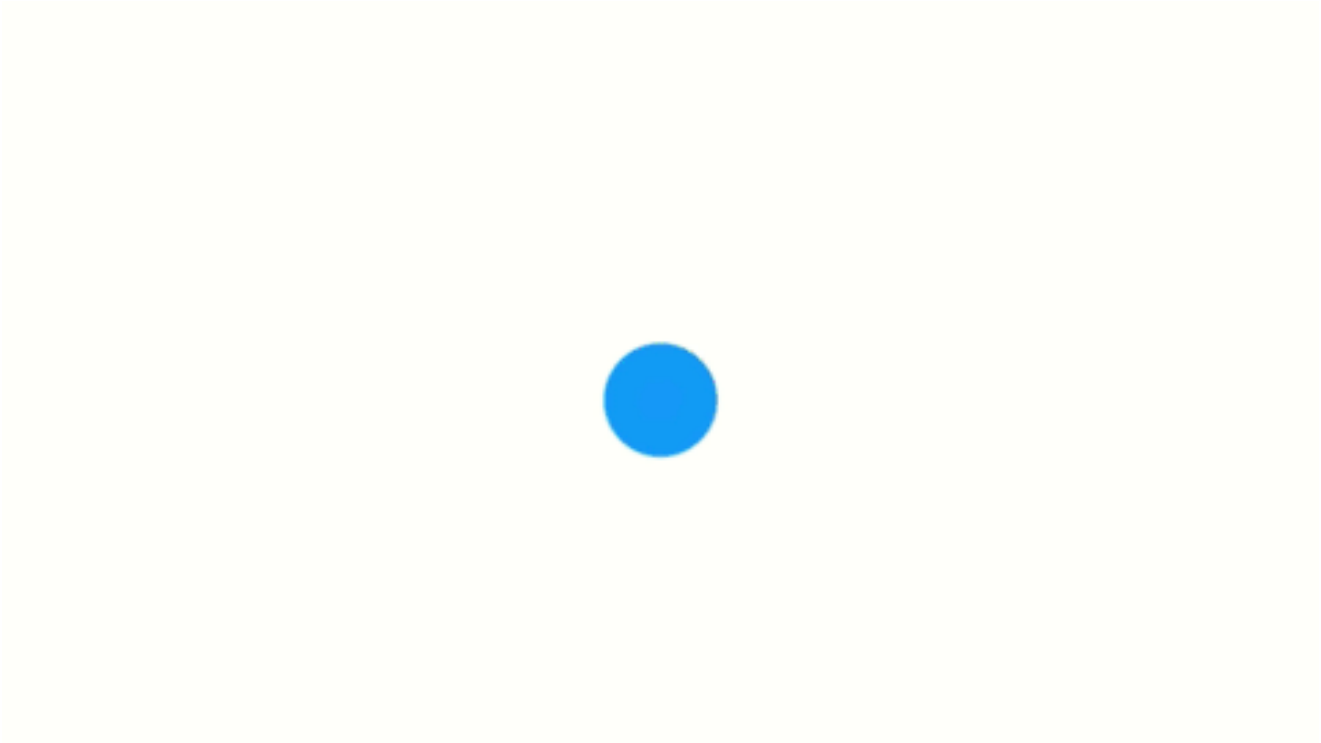
Course literature:
Time series Analysis and its Applications
Can be downloaded freely here:

https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf

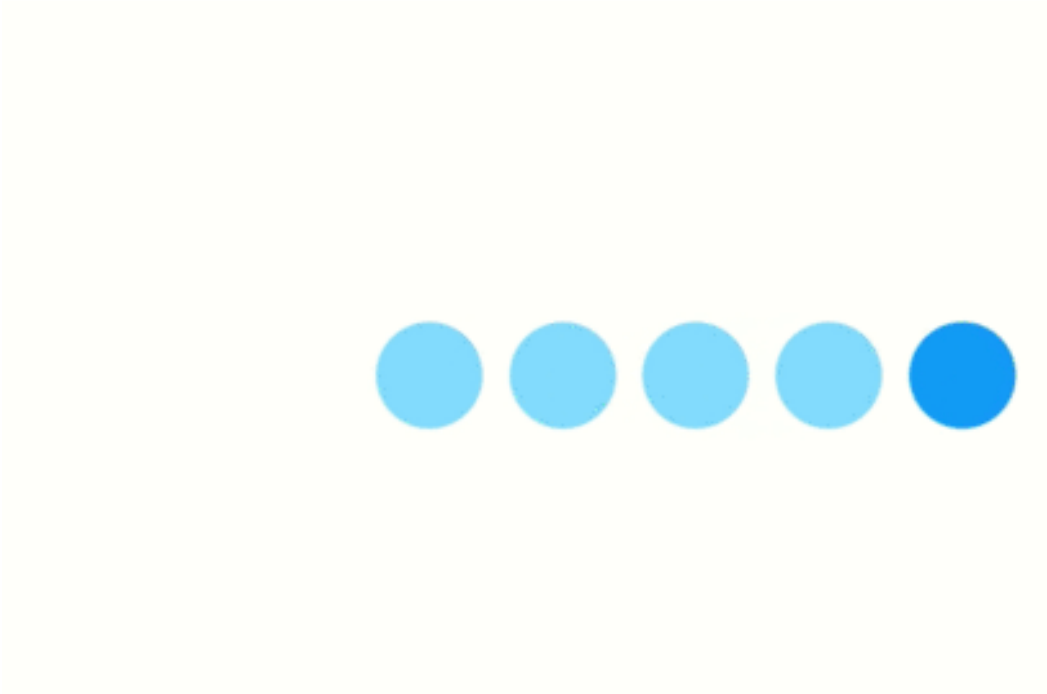
Software for computer labs is **R**:



Sequential data



Sequential data: Motion of a ball



Sequential data: A sentence

This is a sequential data type.

Sequential data: A sentence

This is a sequential data type.

This is a sequential data type .

Sequential data: A sentence or a word

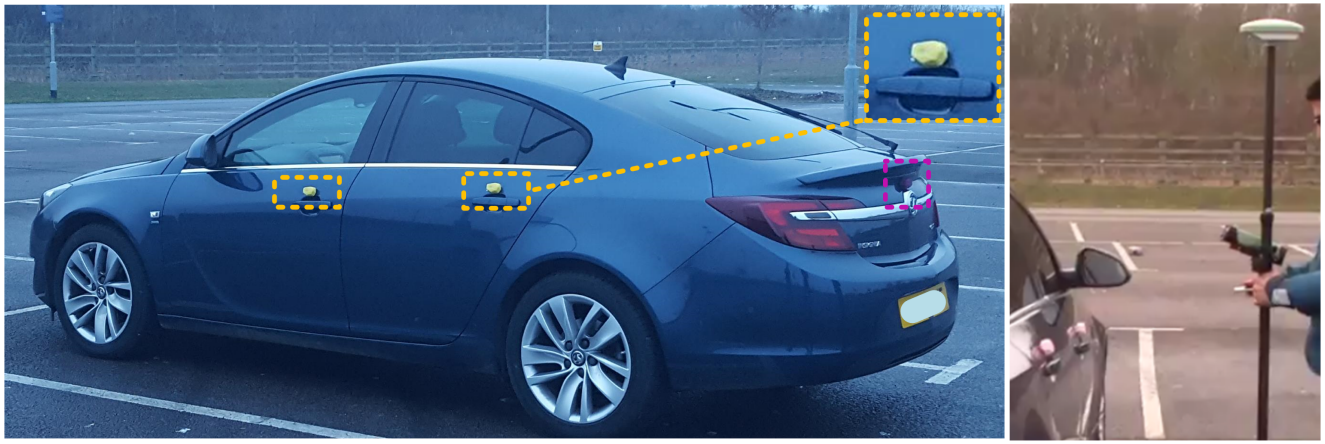
This is a sequential data type.

This is a sequential data type .

s e q u e n t i a l

A look at real data

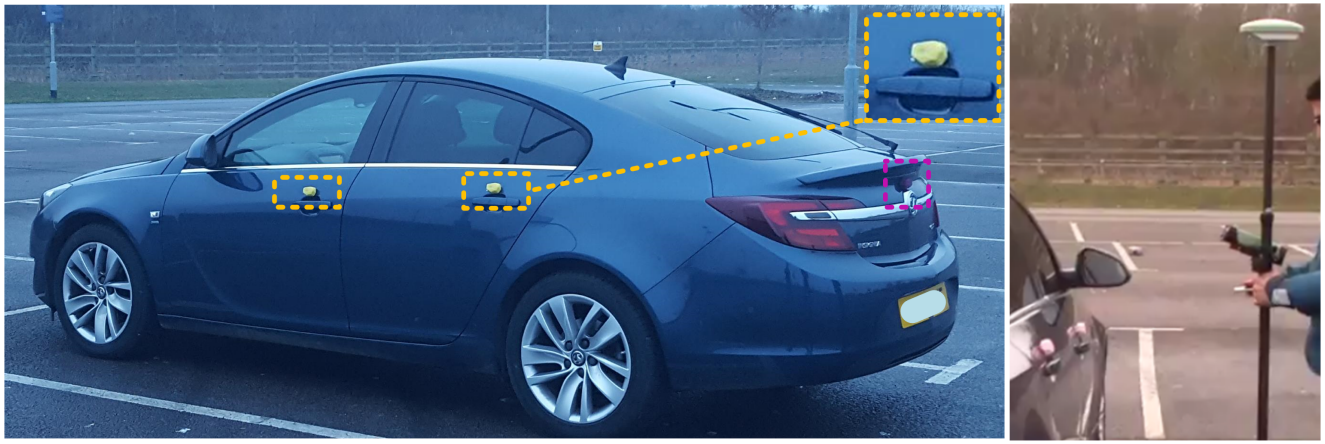
Received signal strength indicator (RSSI) is a common observation (data).



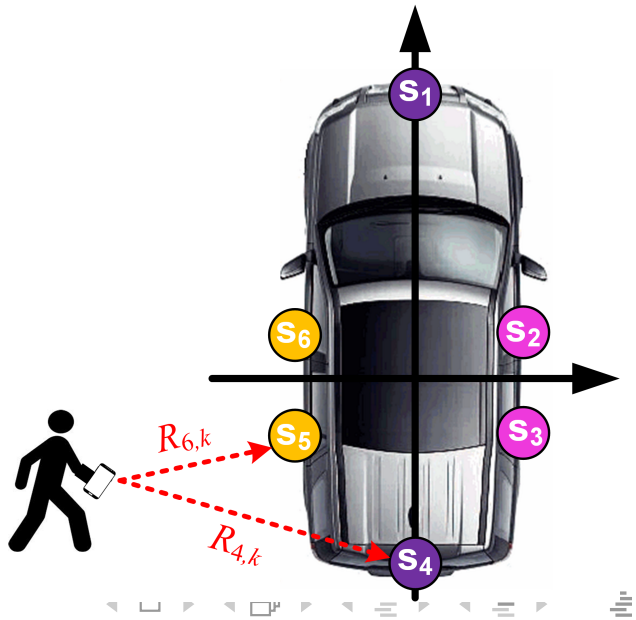
Where is the driver?

A look at real data

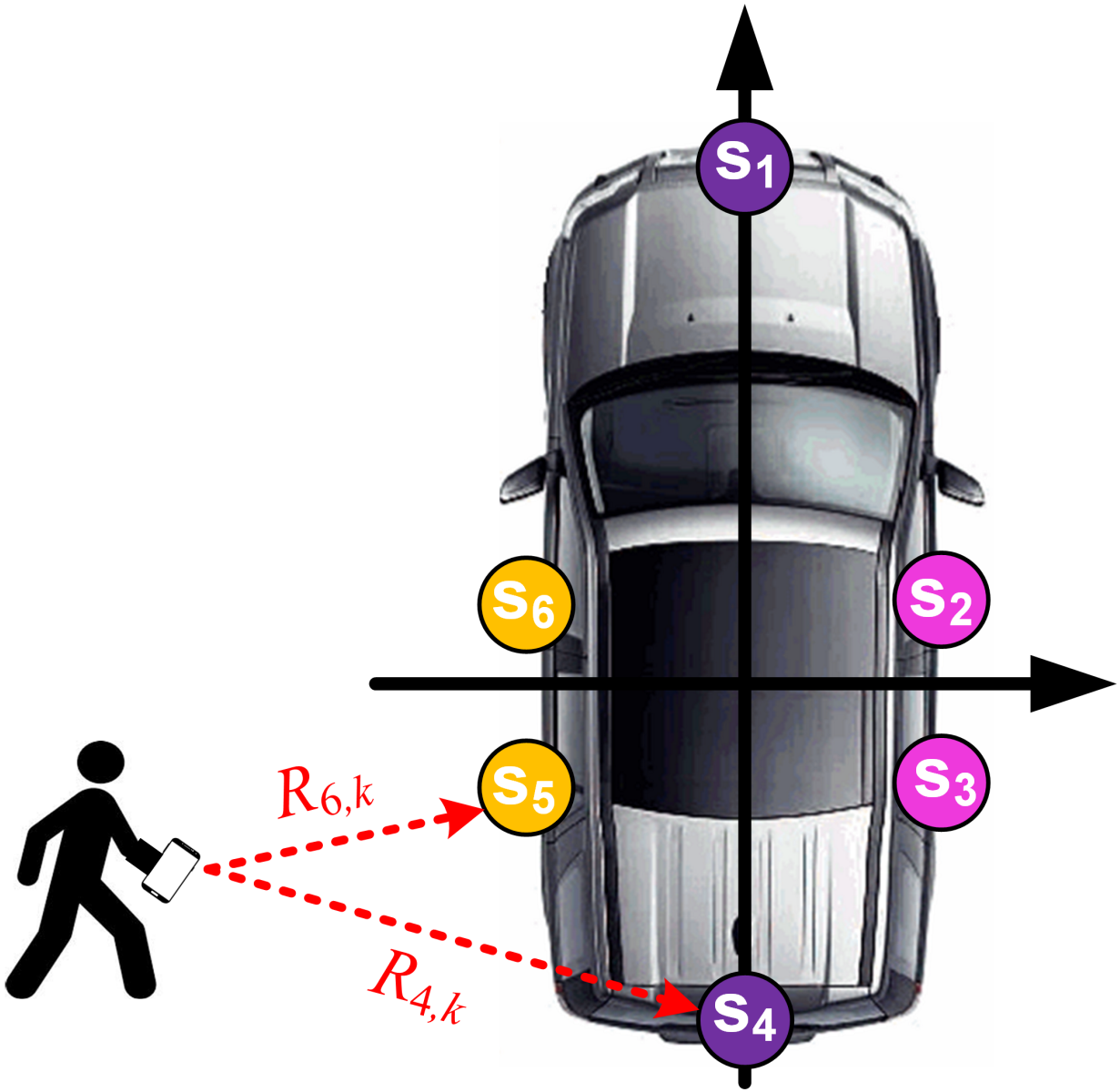
Received signal strength indicator (RSSI) is a common observation (data).



Where is the driver?



Where is the driver?

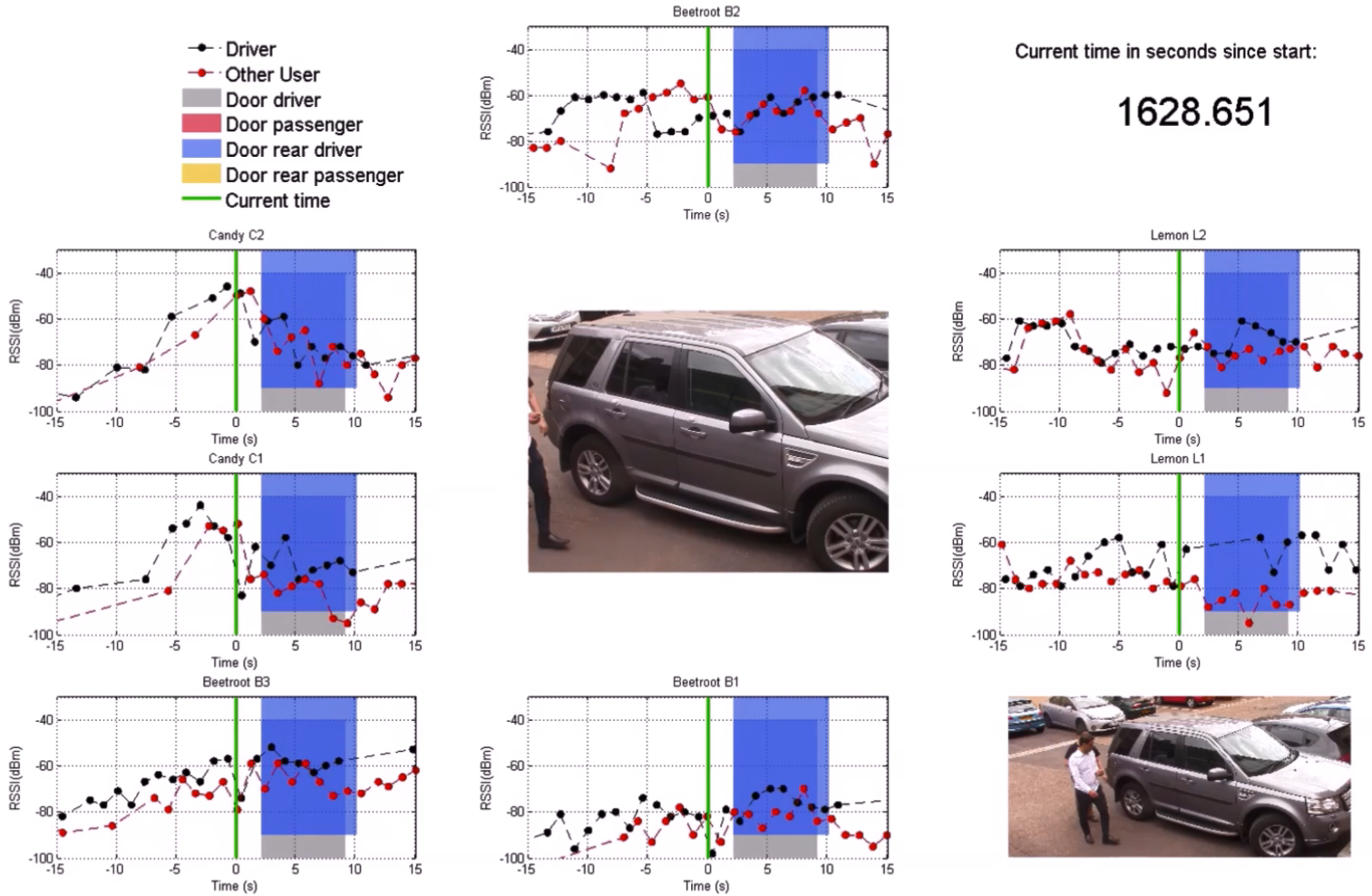


Where is the driver?

Video of data collection

Where is the driver?

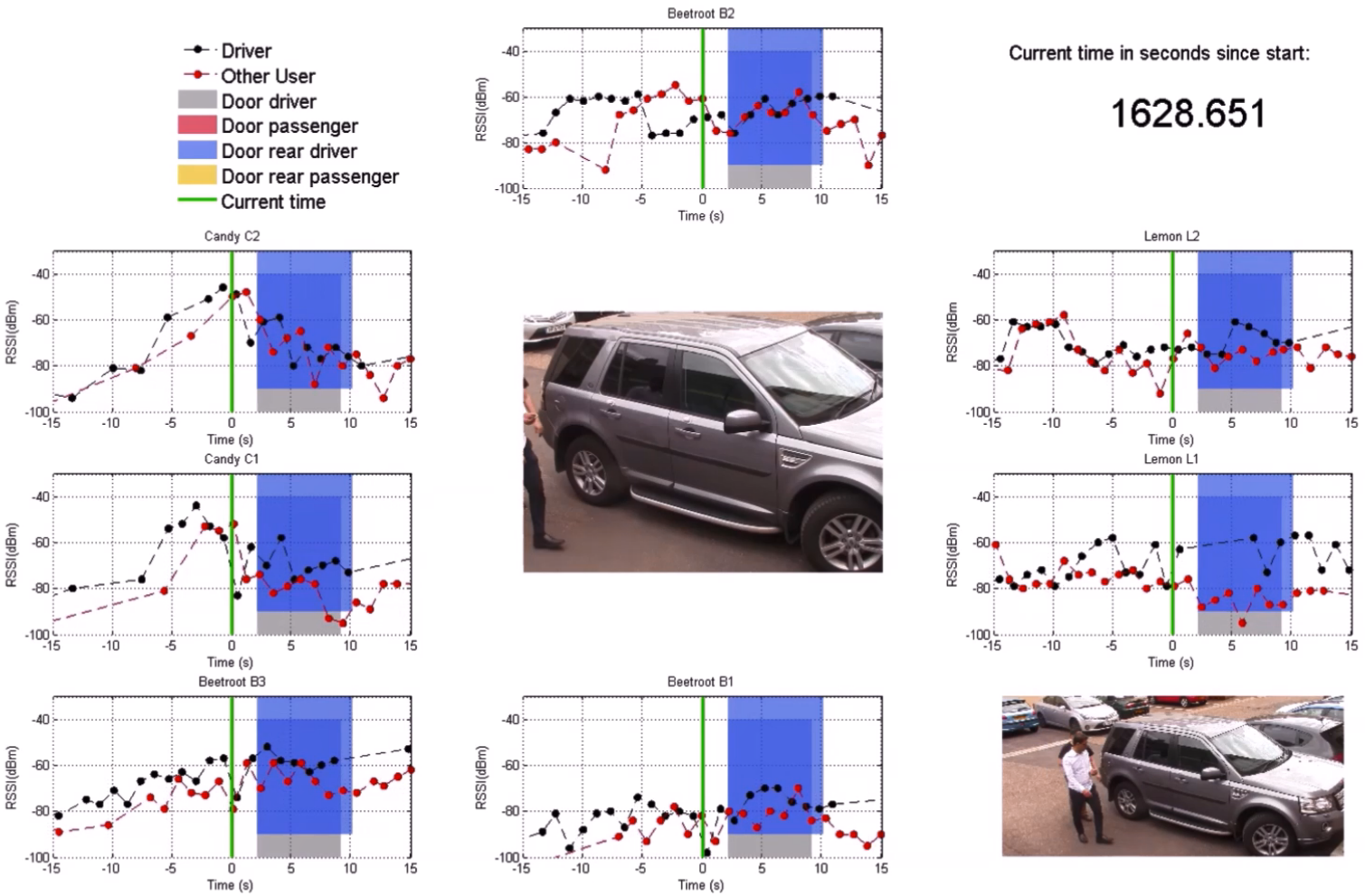
Animation of the of signals



Time Series Analysis

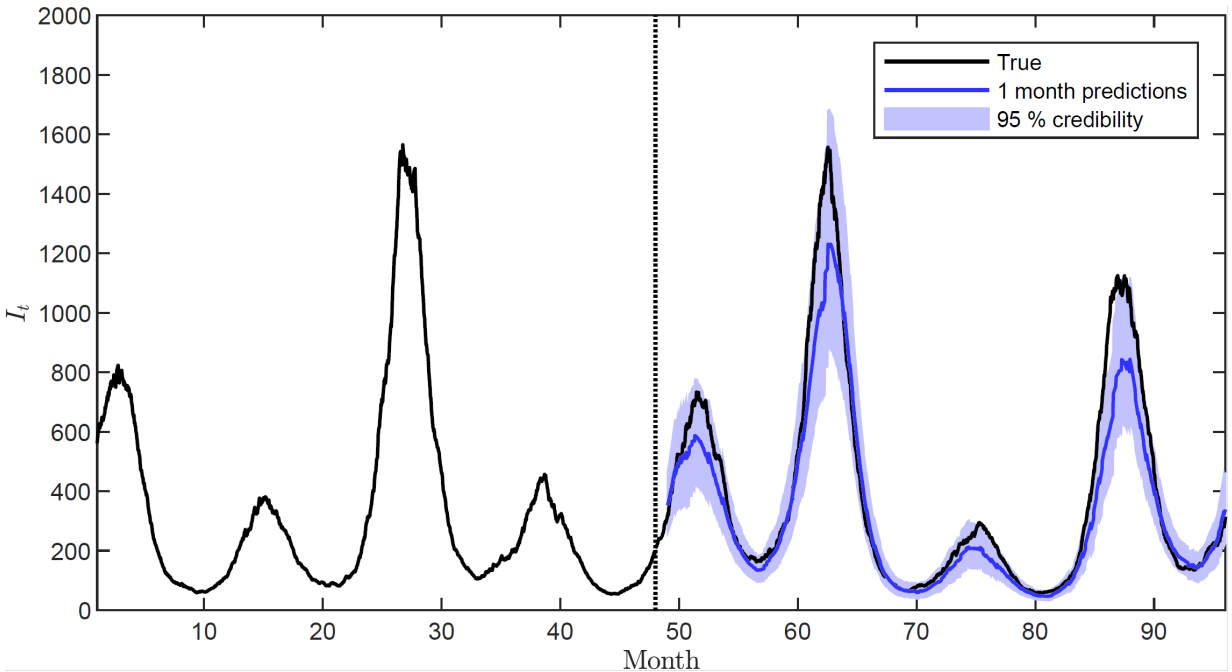
What is a Time Series?

- A sequential data where observations are collected over time
- Observations are typically correlated!



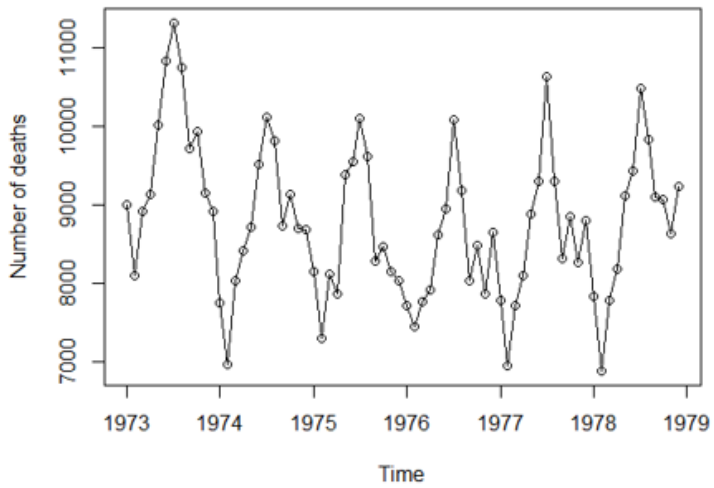
Time Series Analysis

- Understand the properties of the underlying process
- Be able to predict (forecast) possible future values
- Reason about the **uncertainties** in the predictions
  requires statistical methods!

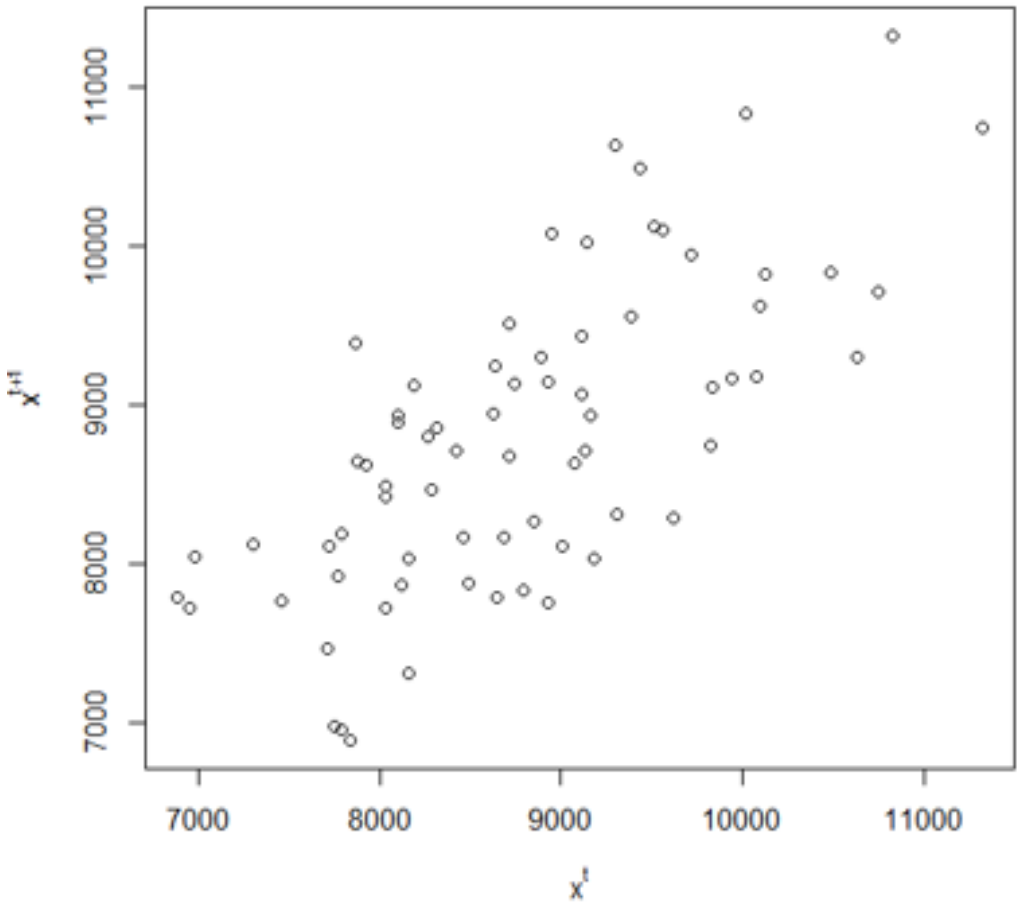


Time Series Analysis

Usual regression analysis: observations are often **iid.**
Time Series Analysis: observations are **correlated!**

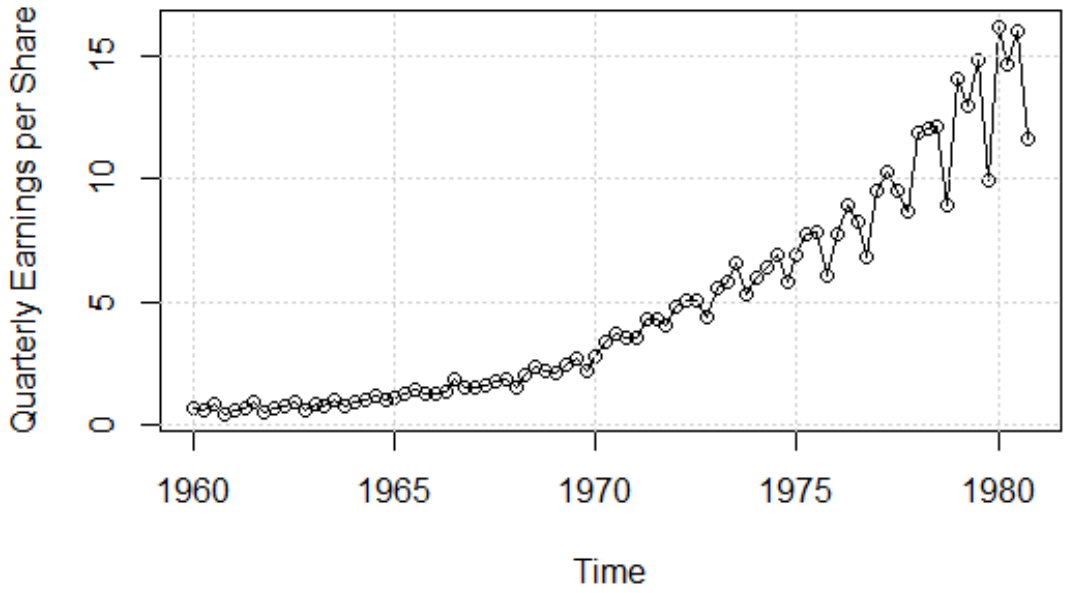


Ex) See connection between x_t and x_{t+1}

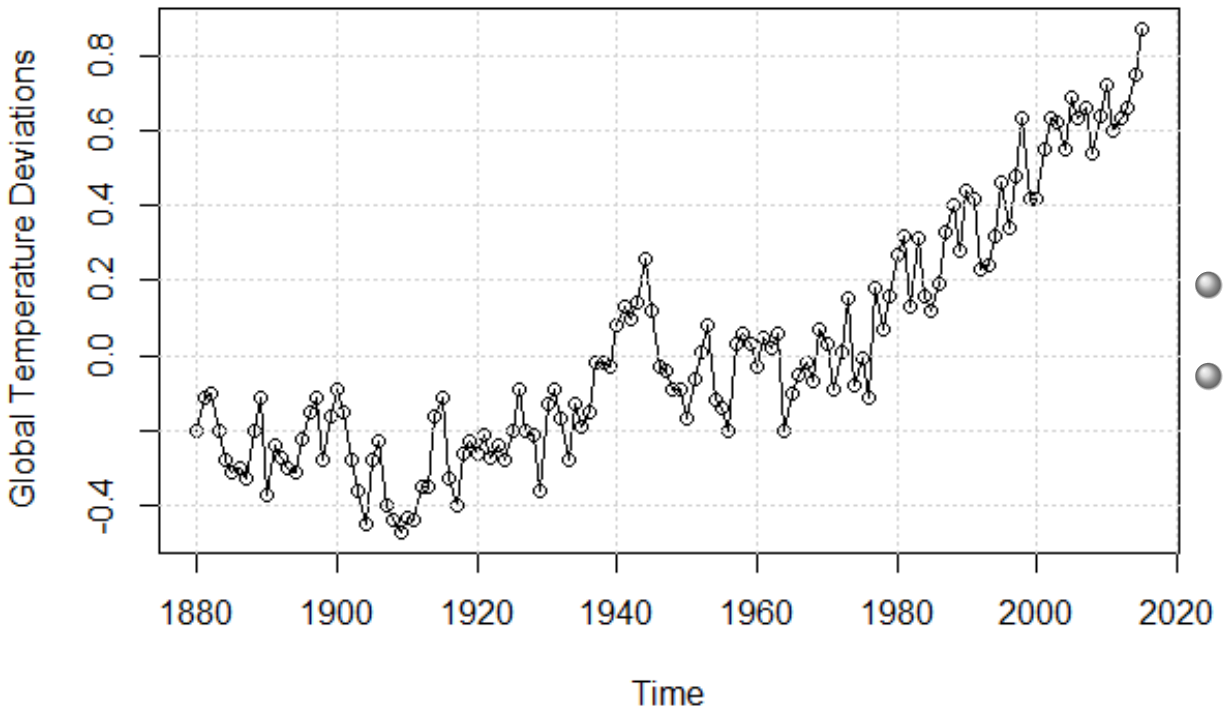


Ex 1: Johnson & Johnson quarterly earnings

- **What can you see?**
  - Trend?
    - ★ Constant
    - ★ Linear
    - ★ Other
  - Variation?
  - Seasonality?
  - Outliers?

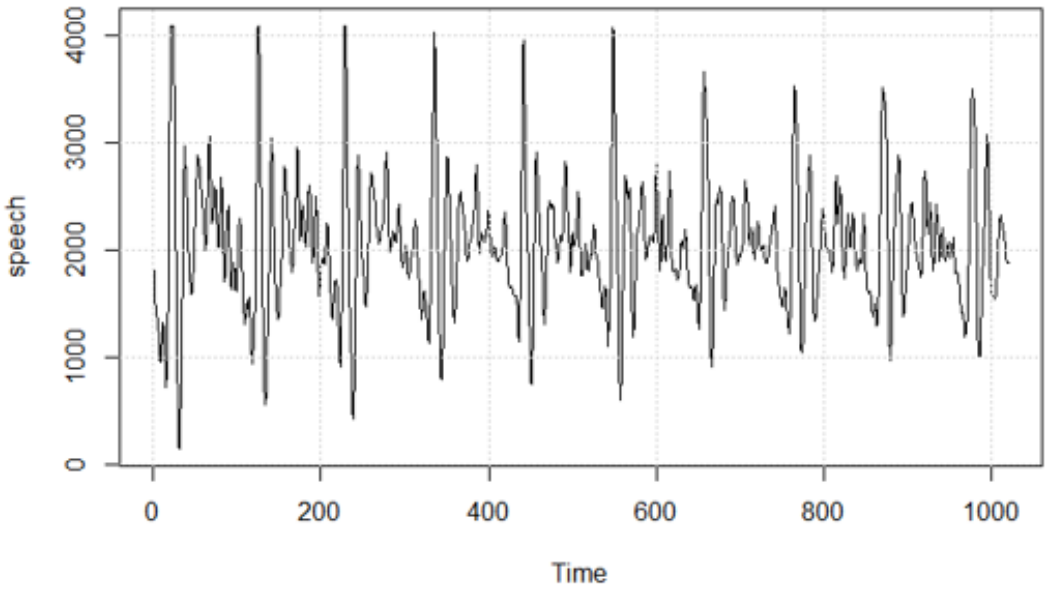


Ex 2: Global warming



- **What can you see?**
- Trend is of interest

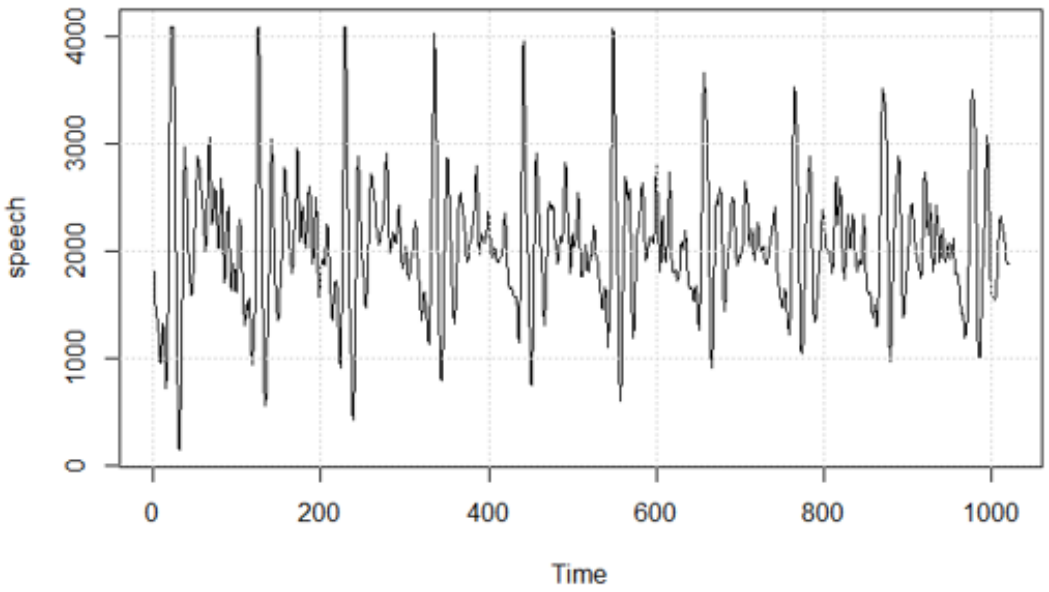
Ex 3: Speech data



- **What can you see?**

Pattern of periodicity is of interest $\rightarrow$ decompose signal into different frequencies

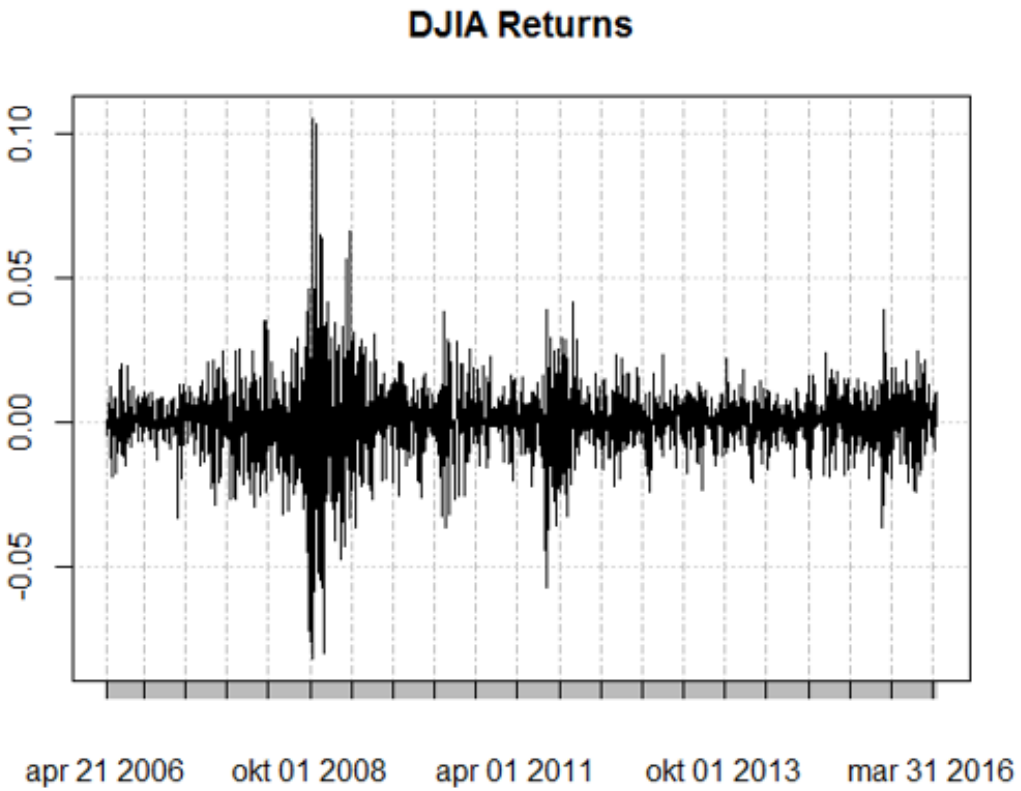
Ex 3: Speech data



- **What can you see?**

Pattern of periodicity is of interest $\rightarrow$ decompose signal into different frequencies **not covered in this course!**

Ex 4: Dow Jones Industrial Average

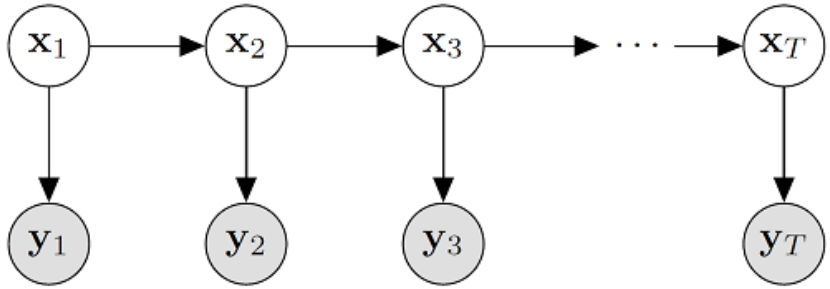


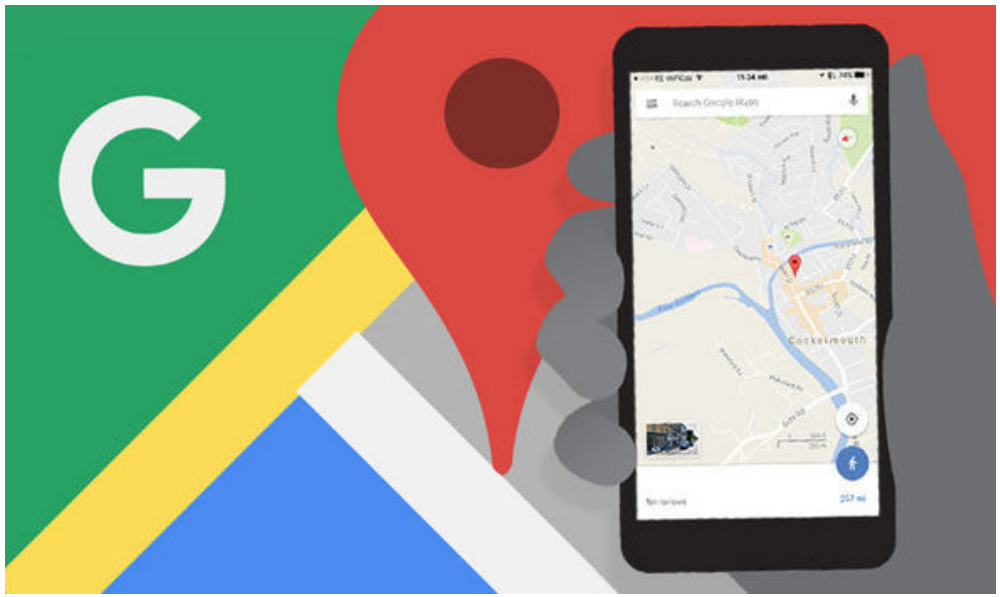
- **What can you see here?**

Pattern of periodicity is of interest $\rightarrow$ Stochastic volatility

Ex 5: Dynamical systems

Linear and Guassian state-space models for tracking objects

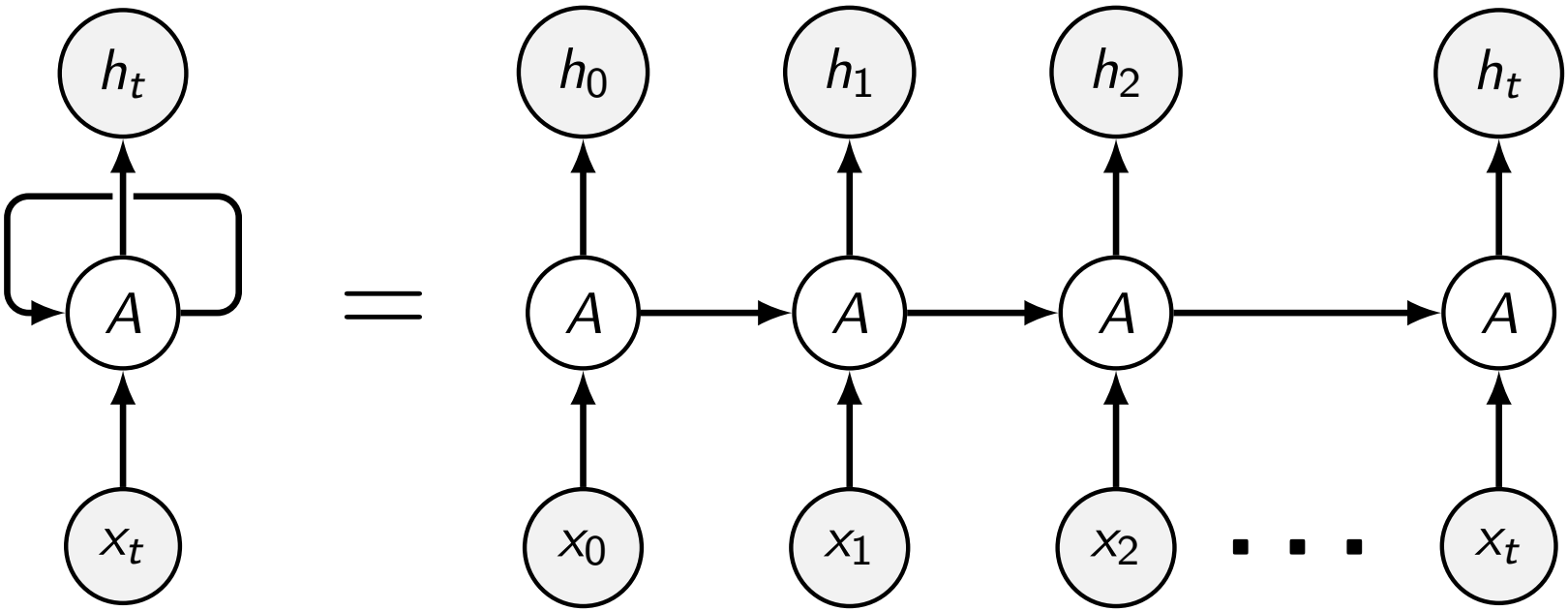




Ex 6: Recurrent neural networks

Natural Language Processing

This is a sequential data type .

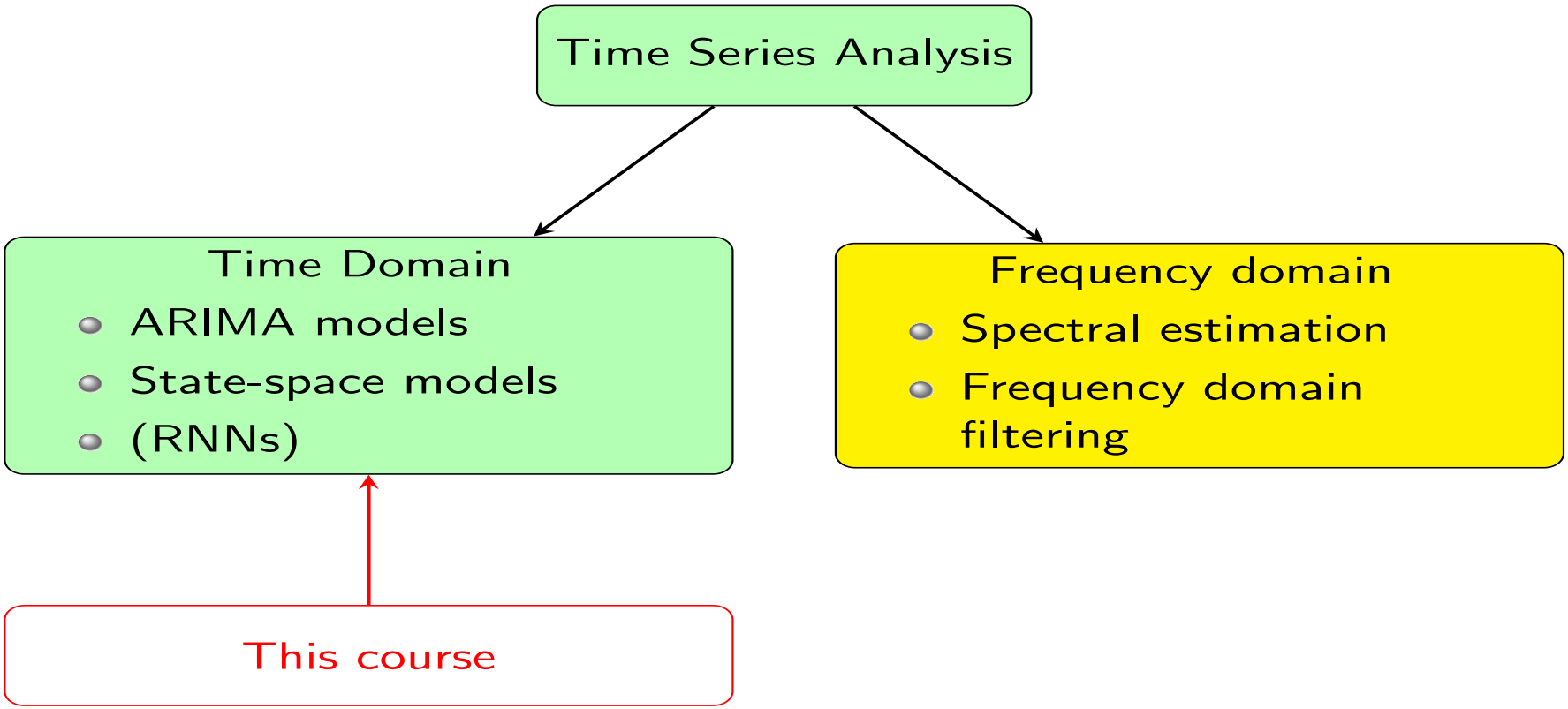


Time Series Analysis

Application areas

- Natural sciences
- Climatology
- Robotics/autonomous systems
- Social sciences
- Medicine
- Economics
- Telecommunications
- ...

The Big Picture



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

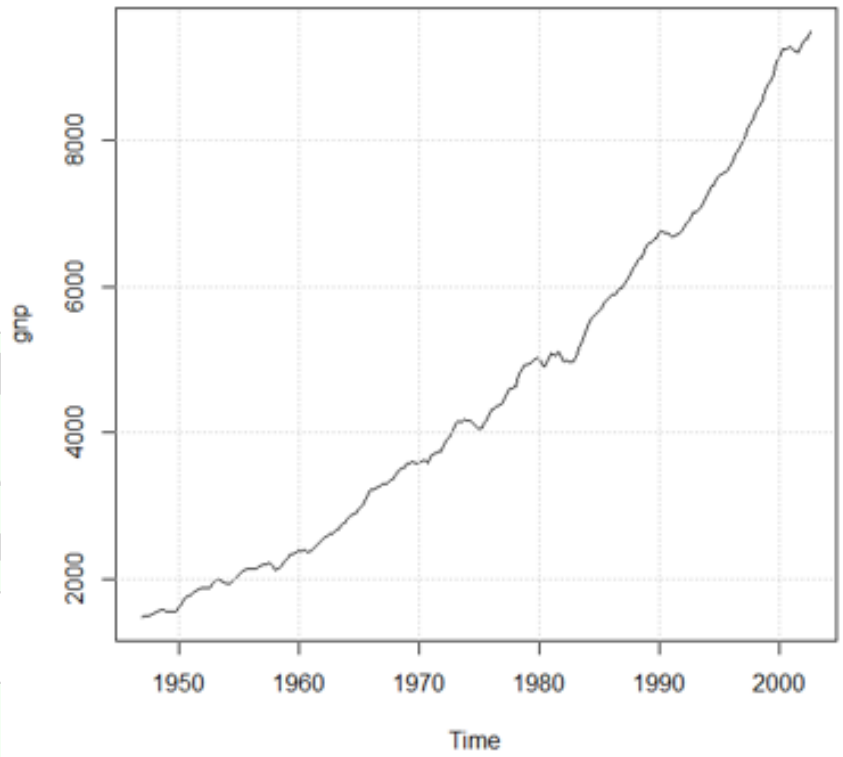
$\theta = (\theta_1, ..., \theta_q)$

Mod

- A
  ar

- Estimate $\phi's$, $\theta's$, ...

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

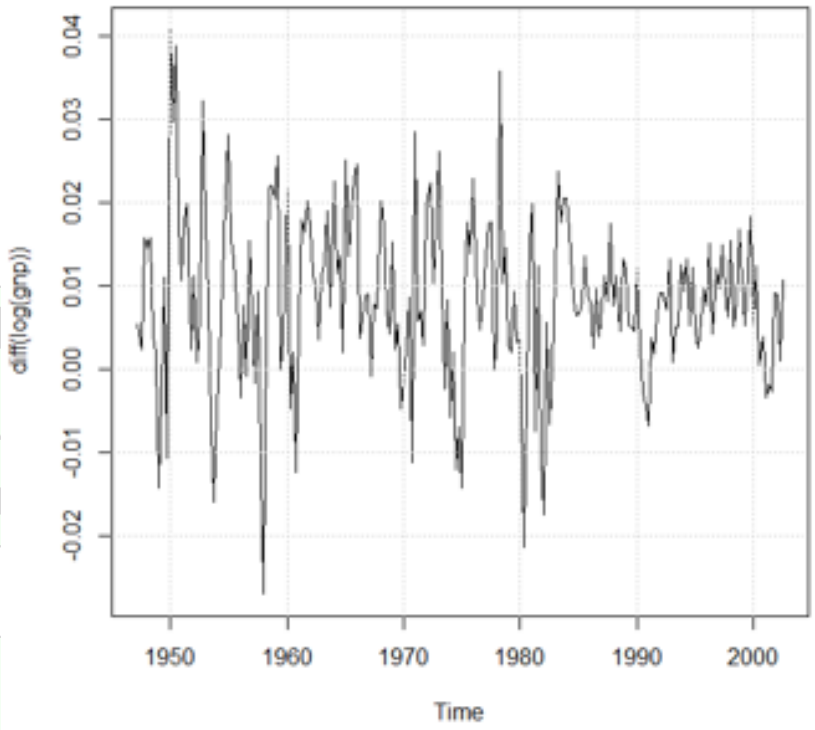
Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Prediction

$Y_t = \nabla(\log(X_t))$



Estimate $\phi's$, $\theta's$, ...

Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

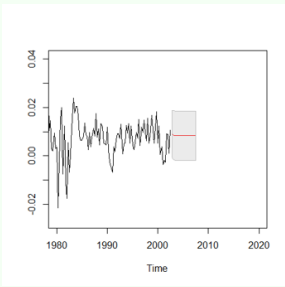
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

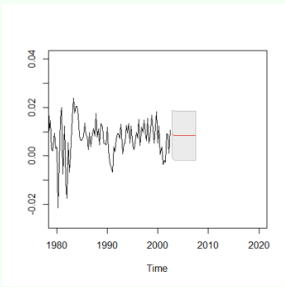
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

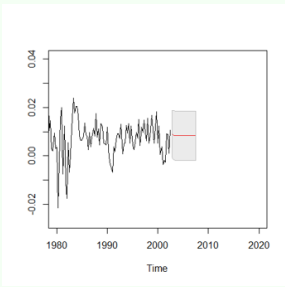
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

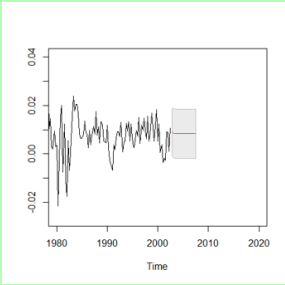
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

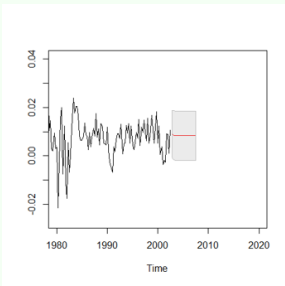
- Are there better models with other p and q or even different kind?

Revise the model

yes

no

Prediction



Course topics

- Time series regression and explorative analysis
- ARIMA models
  - AR, MA, ARMA, ARIMA, seasonal ARIMA
  - Model selection
  - Estimation
  - Forecasting
- State space models
  - Linear and Gaussian state space models
  - Kalman filtering and smoothing
- Recurrent Neural Networks (RNNs)

Course organization

- Lectures
  - Available at LISAM
- Teaching sessions
- Computer labs
  - Available at LISAM, under Submissions
  - Work in pairs
  - Send your report via LISAM
  - Deadlines
- Written assignments
  - Submissions needed - keys are given for some assignments
- Examination
  - Computer based exam
  - Submission of lab reports and written assignments

Course organization

- Software: R
  - https://www.r-project.org/
  - https://www.rstudio.com/

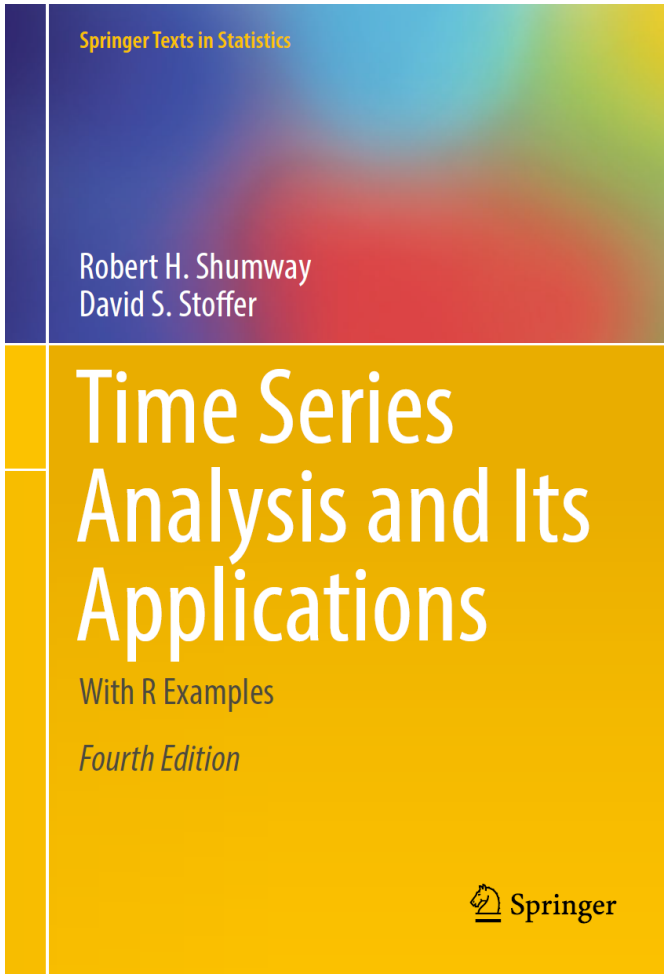
- Define your groups (2 persons) this week:
  - https://docs.google.com/spreadsheets/d/1tzG35WSDWRhHWFAOcNOLlWUzoYqdn0GhZUwvXz3HhII/edit?usp=sharing
  - **Difficult to find a group? Put your name in some cell.. I will merge you to someone**

Course organization

Course literature:
Time series Analysis and its Applications, Fourth Edition (2017), ISBN 978-3-319-52451-1
Can be downloaded freely here:

https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf

- Do not skip examples when you read!
- First 2 chapters are easy, but don't relax!



Time Series models

- Time series x_t : random variable
  - A collection of x_t = stochastic process
  - $t = 0, \pm 1, \pm 2, ...$
- (probably) Simplest series: white noise
  - w_t uncorrelated (white: all possible periodic oscillations are present at equal strength)

$$w_t \sim wn(0, \sigma_w^2)$$

  - w_t independent and identically distributed (white independent noise)

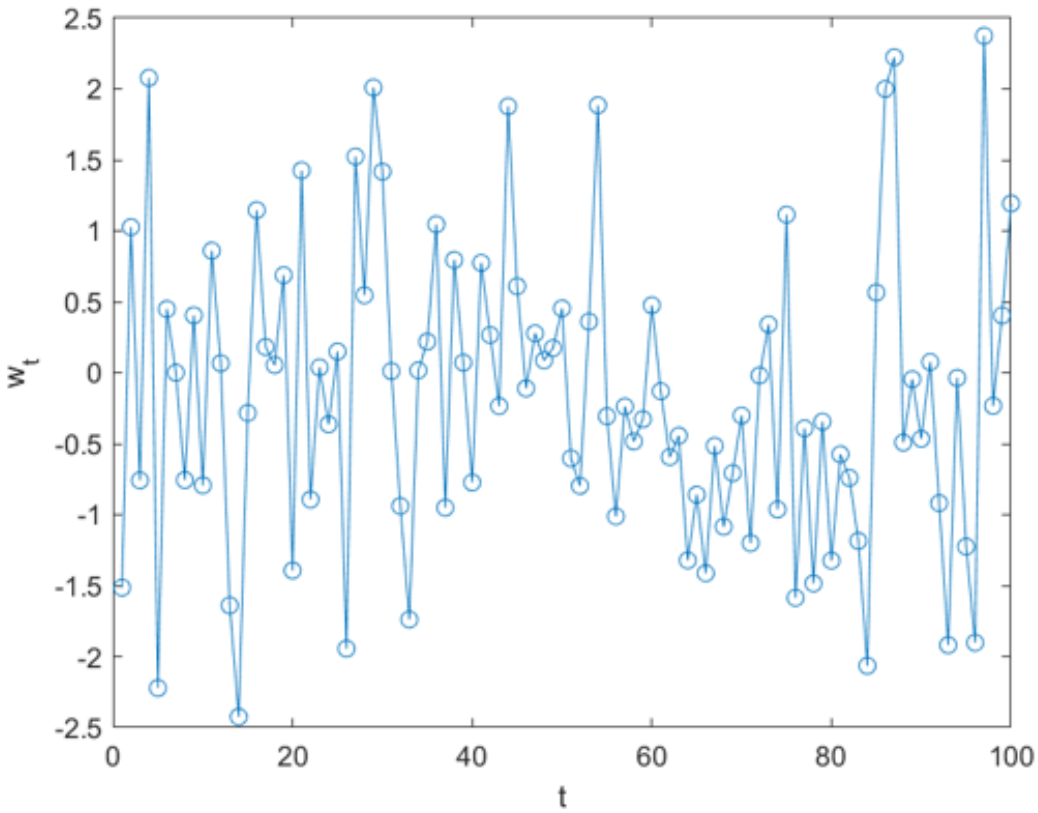
$$w_t \sim iid(0, \sigma_w^2)$$

- Reminder:

$$\text{uncorrelated} \iff E(XY) = EX.EY$$
$$\text{independent} \iff f_{X,Y}(x, y) = f_X(x).f_Y(y)$$

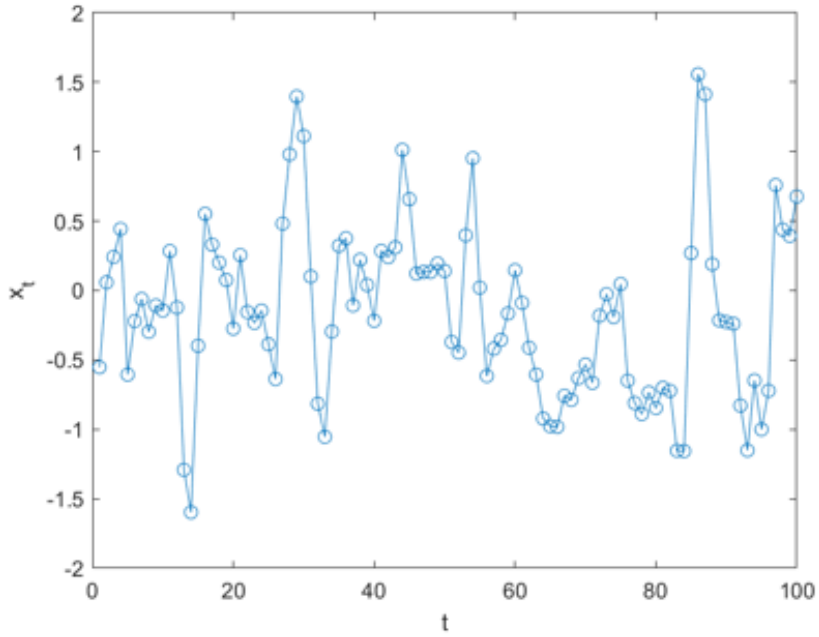
White noise

- Example: $w_t \sim iidN(0,1)$



Moving average

Example: $x_t = 0.2w_{t-1} + 0.5w_t + 0.2w_{t+1}$



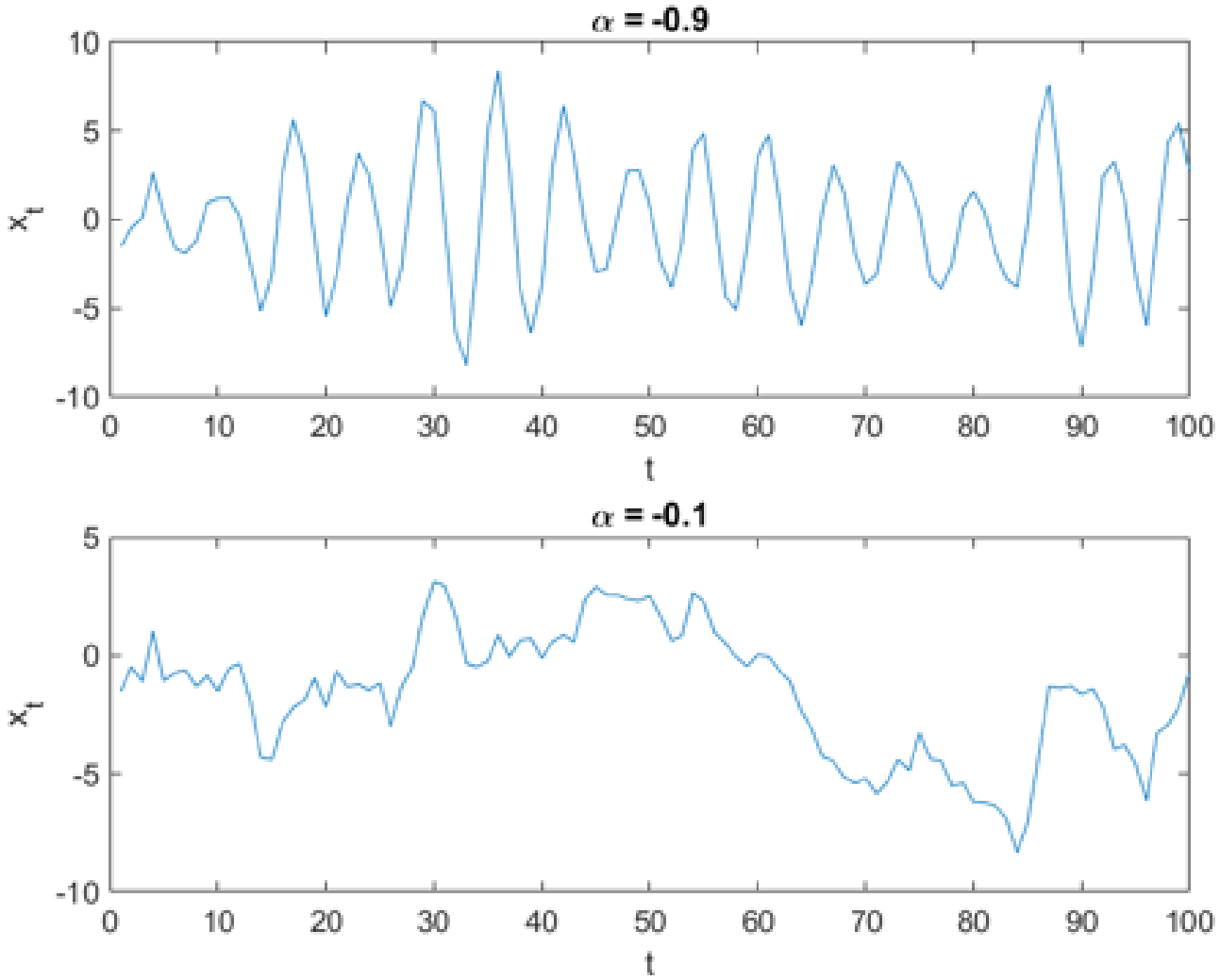
Very Interesting Fact: most stationary processes can be represented as a sum of lagged white noise:

$$x_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}$$

Autoregressive model

Example: AR(2) process (Assume $x_0 = 0, x_{-1} = 0$)

$$x_t = x_{t-1} + \alpha x_{t-2} + w_t$$



Random walk with drift

A simple model for a "drifting" time series

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

- δ is the drift
- $\delta = 0 \Rightarrow$ random walk

Note: if we assume $x_0 = 0$,

$$x_t = \delta t + \sum_{j=1}^{t} w_j$$

Random walk with drift

A simple model for a "drifting" time series

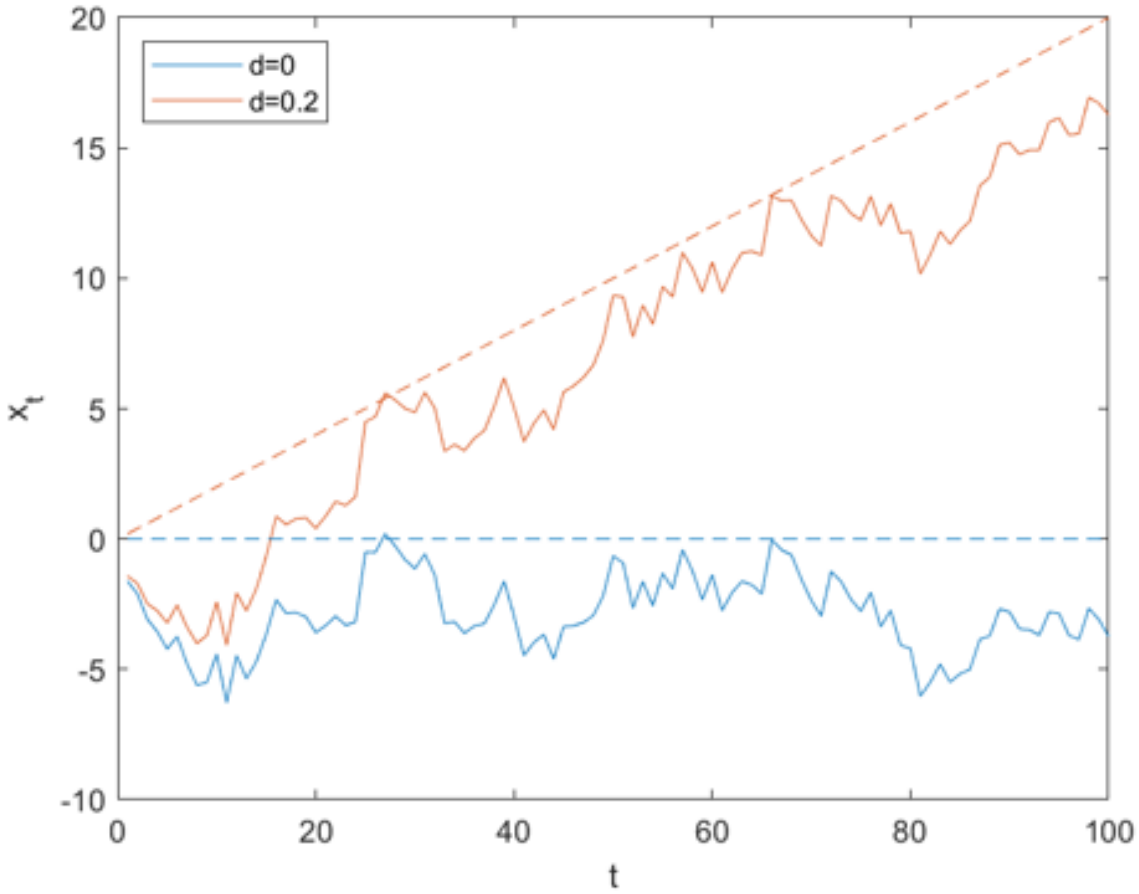
$$x_t = \delta + x_{t-1} + w_t$$

- δ is the drift
- $\delta = 0 \Rightarrow$ random walk

Note: if we assume $x_0 = 0$,

$$x_t = \delta t + \sum_{j=1}^{t} w_j$$

δ=0 and δ=0.2



Basic statistics - reminder

- Probability density function for x: $f(x)$
- Marginal density $f_i(x_i) = \int f(x) dx_1 ... dx_{i-1} dx_{i+1} ... dx_p$
- Expected (mean) value $Ex = \int x f(x) dx$
- Covariance $\text{cov}(x, y) = E\{(x - Ex)(y - Ey)\}$
- Variance $\text{var}(x) = E\{(x - Ex)^2\} = \text{cov}(x, x)$
- Relationships (a is a constant)
  - $E(x + a) = Ex + a$, $E(ax) = aEx$
  - $E(x + y) = Ex + Ey$
  - $\text{cov}(x + a, y) = \text{cov}(x, y)$
  - $\text{cov}(x + z, y) = \text{cov}(x, y) + \text{cov}(z, y)$
  - $\text{var}(ax) = a^2 \text{var}(x)$

Statistical representation of a time series

Which measures of dependence exist for time series?

- Theoretical?
- Practical?

Given time series $x_1, ...x_n$ measured at fixed $t_1, ...t_n$

- Joint pdf

$$f_{t_1,...t_n}(x_{t_1}, ..., x_{t_n})$$

- Marginal pdf

$$f_t(x_t)$$

Statistical representation of a time series

on whiteboard

Mean function at time t

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

Examples: Compute mean function for

- Moving average $x_t = 0.2w_{t-1} + 0.5w_t + 0.2w_{t+1}$
- Random walk $x_t = \delta t + \sum_{j=1}^{t} w_j$

Autocovariance and ACF

How do we measure linear dependence between two variables? $\rightarrow$ Covariance or Correlation

How do we measure linear dependence between two time-lags in a time series? In the same way!

- Autocovariance function

$$\gamma(s,t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$$

  Note $\text{var}(x_t) = \gamma(t,t)$

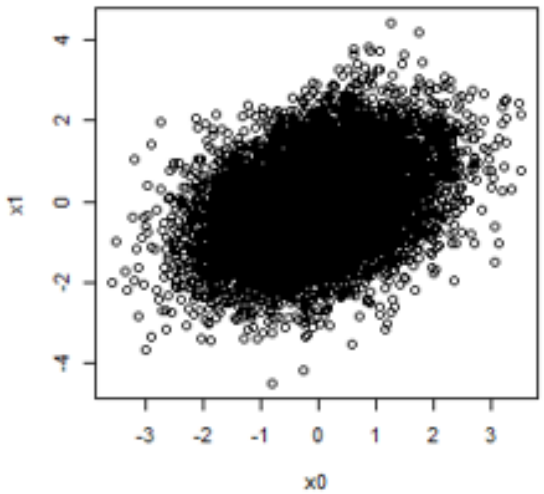
- Autocorrelation function (ACF)

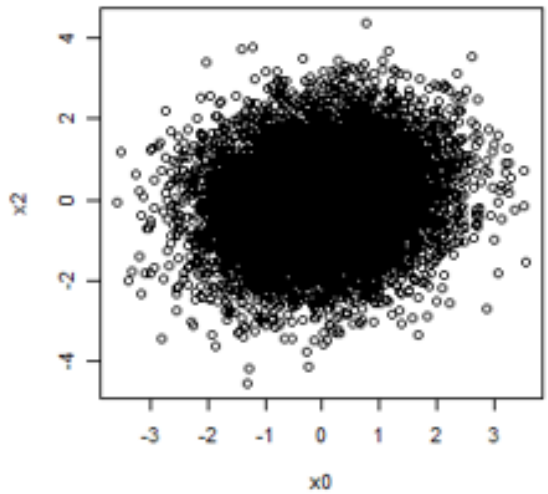
$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$$

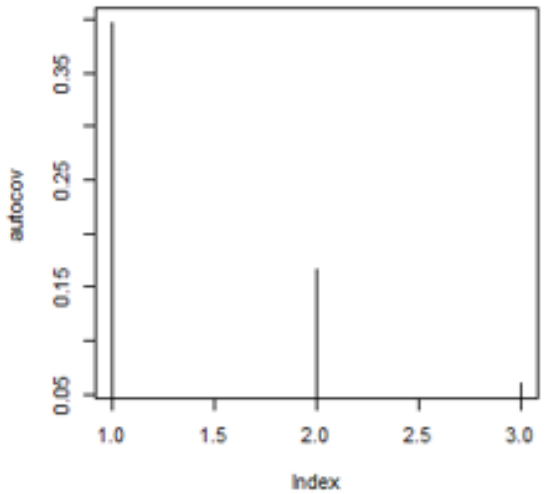
Autocovariance and ACF

Generate x_0, x_1, x_2 from $x_t = 0.4x_{t-1} + w_t$

- Consider $\gamma(0,1), \gamma(0,2)$







Autocovariance and ACF

Useful fact: If $U = \sum_{j=1}^{m} a_j x_j$ and

$$V = \sum_{k=1}^{r} b_k y_k$$

$$\operatorname{cov}(U, V) = \sum_{j=1}^{m} \sum_{k=1}^{r} a_j b_k \operatorname{cov}(x_j, y_k)$$

Examples: Autocovariance and ACF of **on whiteboard**

- White noise
- Random walk $x_t = \delta t + \sum_{j=1}^{t} w_j$
- Moving average $x_t = 0.2w_{t-1} + 0.5w_t + 0.2w_{t+1}$

Home reading

- Shumway and Stoffer, chapters 1.1-1.3
- TS functions in R: ts, plot.ts, acf, ts.intersect, filter, ts.plot

Time Series Analysis

Lecture 2: Exploratory analysis and Time Series Regression

Tohid Ardeshiri

Linköping University
Division of Statistics and Machine Learning

September 4, 2019

Summary of Lecture 1

- Time series
  - White noise
  - Random walk
  - Moving average filter

- Autocovariance and autocorrelation functions:

$$\gamma(s,t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$$

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

Autocovariance and ACF

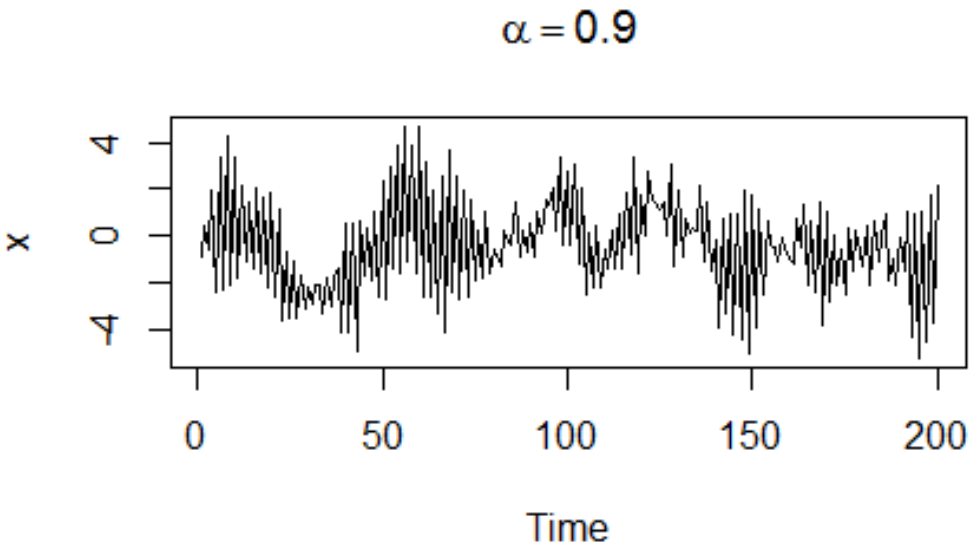
Examples: Autocovariance and ACF of **on whiteboard**

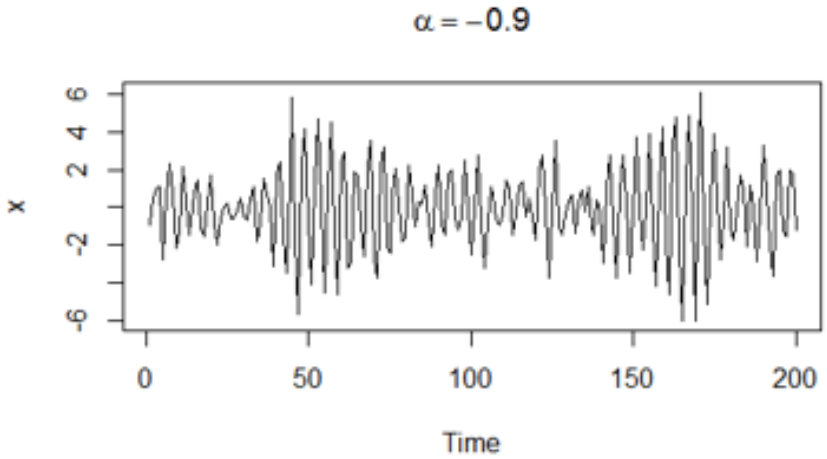
- White noise ✓
- Random walk $x_t = \delta t + \sum_{j=1}^{t} w_j$
- Moving average $x_t = 0.2w_{t-1} + 0.5w_t + 0.2w_{t+1}$

Autocovariance

- Intuition:

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





- when $x_0 = 0$ and $w_t \sim wn(0, 1)$:

$$\text{cov}(x_t, x_{t-1}) = \phi$$

Autocovariance (read at home)

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

Mean function:

$Ex_t = \phi Ex_{t-1} + Ew_t = \phi Ex_{t-1} = \phi(\phi Ex_{t-2}) = \cdots = \phi^t Ex_0$

for $Ex_0 = 0$, $Ex_t = 0$ for all t.

Variance $\text{var}(x_t)$ when $Ex_0 = 0$ and w_t is uncorrelated with x_0 for all t:

$\text{var}(x_t) = E\{(x_t - 0)^2\} = E\{\phi^2 x_{t-1}^2 + 2\phi x_{t-1} w_t + w_t^2\} =$

$\phi^2 \text{var}(x_{t-1}) + 2\phi \,\text{cov}(x_{t-1}, w_t) + \text{var}(w_t) = \phi^2 \text{var}(x_{t-1}) + \text{var}(w_t) =$

$\phi^2 \text{var}(x_{t-1}) + \sigma_w^2 = \phi^2(\phi^2 \text{var}(x_{t-2}) + \sigma_w^2) + \sigma_w^2 =$

$\phi^{2t} \text{var}(x_0) + \sigma_w^2 \sum_{k=0}^{t-1}(\phi^{2k}) = \phi^{2t} \text{var}(x_0) + \frac{\sigma_w^2(1-\phi^{2t})}{1-\phi^2}$

When $\text{var}(x_0) = \frac{\sigma_w^2}{1-\phi^2}$ then $\text{var}(x_t) = \frac{\sigma_w^2}{1-\phi^2}$ and time independent.

Autocovariance (read at home)

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

$$x_t = \phi(\phi x_{t-2} + w_{t-1}) + w_t = \cdots = \phi^h x_{t-h} + \sum_{j=0}^{h-1} \phi^j w_{t-j}$$

$$\gamma(x_t, x_{t-h}) = \text{cov}(x_t, x_{t-h}) = E(x_t x_{t-h}) =$$
$$E\{(\phi^h x_{t-h} + \sum_{j=0}^{h-1} \phi^j w_{t-j}) x_{t-h}\} = \phi^h \,\text{var}(x_{t-h}) = \frac{\phi^h \sigma_w^2}{1-\phi^2}$$

Hence,

$$\gamma(h) = \frac{\phi^h \sigma_w^2}{1-\phi^2}$$

Also,

$$\rho(h) = \phi^h$$

Stationarity

Fact: sometimes $\rho(s,t)$ pends on lag $|s-t|$ only

Time series is strictly stationary if distributions of $\{x_{t1}, \ldots x_{tn}\}$ and $\{x_{t1+h}, \ldots x_{tn+h}\}$ are identical for any $\{t_1, \ldots t_n\}$ and all lags $h = 0, \pm 1, \pm 2, \ldots$

$$P(x_{t1} \leq c_1, \ldots x_{tn} \leq c_n) = P(x_{t1+h} \leq c_1, \ldots x_{tn+h} \leq c_n)$$

Note: This means

- Mean function $\mu_t = Ex_t = \text{const.}$
- Autocovariance $\gamma(t, t+h) =$ function only of lag h

Stationarity

Strict stationarity is often too strong!

- Time series x_t is weakly stationary (stationary) if
  - $Ex_t = const$
  - $\gamma(s,t) = \gamma(|s-t|)$
  - $\text{var}(x_t) < \infty$
- $\gamma(t, t+h) = \gamma(|t+h-t|) = \gamma(h)$
  - Autocovariance depends on lag only!
- Autocovariance for stationary process $\gamma(h) = \text{cov}(x_t, x_{t+h})$
- ACF for stationary process $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$

Stationarity

Properties of stationary process:

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

$$|\gamma(h)| \leq \gamma(0) \qquad \rho(h) \leq 1, \rho(0) = 1$$

Reflect: Are these processes stationary?

- White noise
- Moving average, $x_t = 0.2w_{t-1} + 0.5w_t + 0.2w_{t+1}$
- Random walk, $x_t = \delta t + \sum_{j=1}^{t} w_j$

Sample autocovariance and ACF

Dependence measures for samples?

- Idea: replace mean and covariance with sample estimates

If x_t is stationary,

- Sample mean

$$Ex \approx \bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

- Sample autocovariance function

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

Sample autocovariance and ACF

Example: n=6, h=2

		X1	X2	X3	X4	X5	X6
X1	X2	X3	X4	X5	x6		

Sample autocorrelation function (sample ACF)

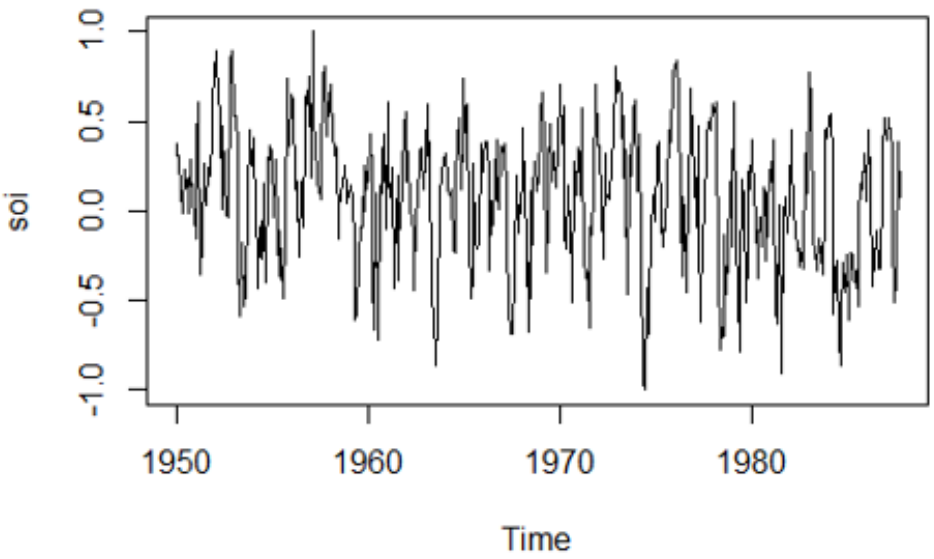
$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

Sample ACF

In R: acf()

Example: southern oscillation index (SOI)

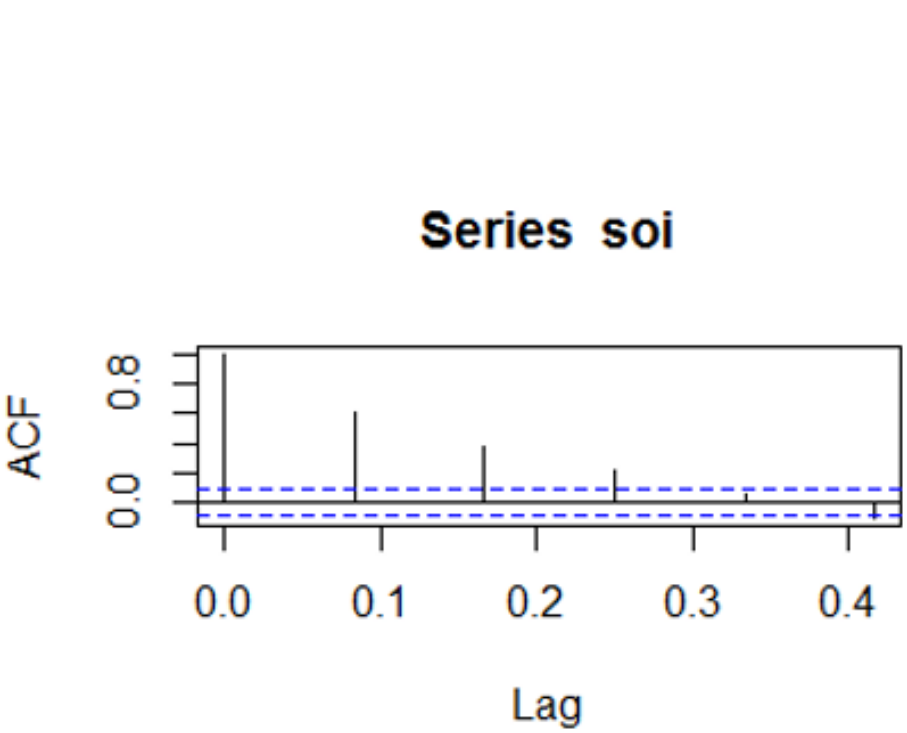
- rho=acf(soi, 5, type="correlation", plot=T)

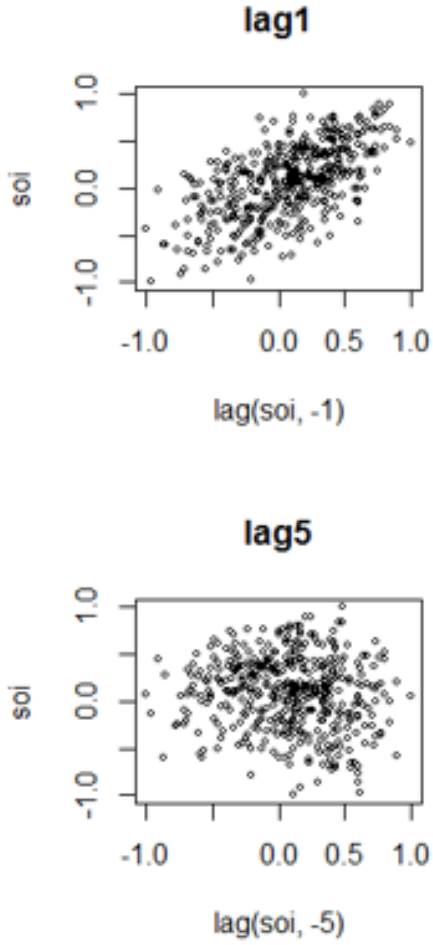



```
> print(rho)

Autocorrelations of series 'soi', by lag

0.0000 0.0833 0.1667 0.2500 0.3333 0.4167
 1.000  0.604  0.374  0.214  0.050 -0.107
```

Why is sample ACF '1' for h=0?

Sample ACF


Series soi
ACF
0.0
0.8
0.0
0.1
0.2
0.3
0.4
Lag


lag1
soi
1.0
0.0
-1.0
-1.0
0.0
0.5
1.0
lag(soi, -1)
lag5
soi
1.0
0.0
-1.0
-1.0
0.0
0.5
1.0
lag(soi, -5)

Sample ACF

What are these blue lines?

Theorem: Under weak conditions, if x_t is white noise and $n \to \infty$ then $\hat{\rho}(h)$ is approximately $N(0, \frac{1}{n})$

Consequence: If some $|\hat{\rho}(h)| > \frac{2}{\sqrt{n}}$ then the time series is not a white noise (with approximately 95 % confidence).

Typical modeling strategy:

- Fit a model
- Compute residuals
- Check ACF within $\pm\frac{2}{\sqrt{n}}$

Sample ACF vs theoretical

- Moving average $x_t = 0.2w_{t-1} + 0.5w_t + 0.2w_{t+1}$
  - $$ACF\gamma(h) = \begin{cases} 1 & h = 0 \\ 0.61 & h = 1 \\ 0.12 & h = 2 \\ 0 & other \end{cases}$$
  - n=10

```
Autocorrelations of series 'y1', by lag

    0     1     2     3     4     5
1.000 0.236 -0.399 -0.187 -0.008 -0.118
```

  - n=1000

```
Autocorrelations of series 'y1', by lag

    0     1     2     3     4     5
1.000 0.609 0.129 -0.007 0.001 0.044
```

Vector-valued time series

If $x_t = (x_{t1}, x_{t2}, \cdots, x_{tp})'$ is stationary,

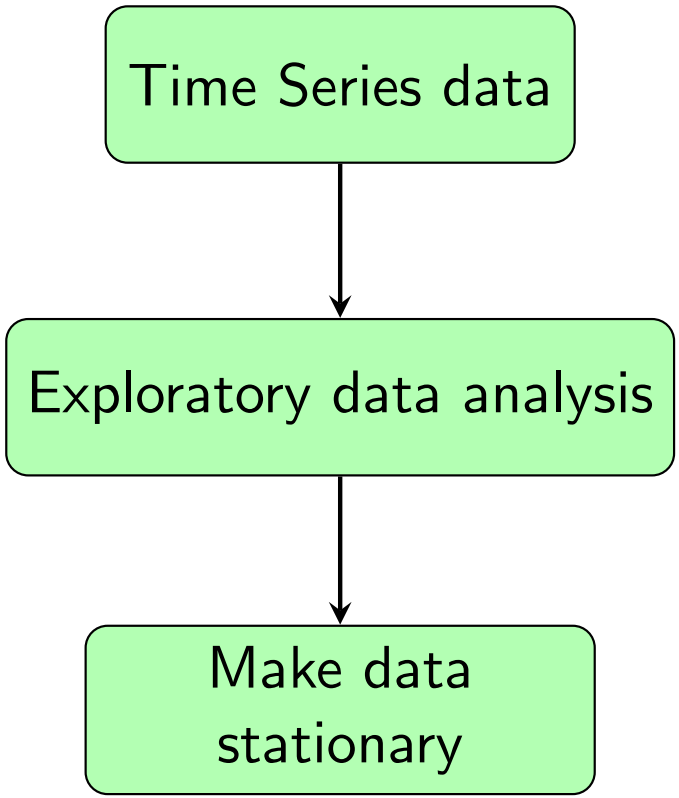
- mean vector is $\mu = E(x_t)$ and sample mean is its approximation

$$\mu = E(x_t) \approx \bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

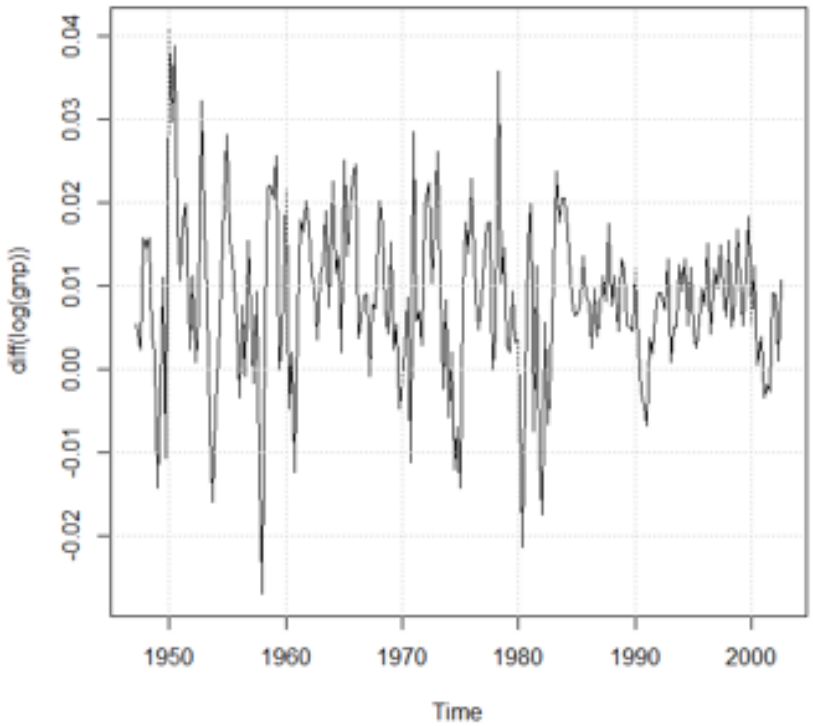
- Autocovariance function is $\Gamma(h) = E[(x_{t+h} - \mu)(x_t - \mu)']$ and sample autocovariance matrix

$$\hat{\Gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})'$$

Recap: time domain modeling



$$Y_t = \nabla(\log(X_t))$$



Stationarity

- Why do we need stationarity?
  - Sample ACF becomes consistent
  - ARIMA models require stationarity

- Tools
  - Detrending (trend removal)
  - Differencing
  - Transformations

whiteboard

- Introduce linear regression/least squares
- Trend removal, simple drift

Trend removal by regression

Regressing on covariates

Given x_t (dependent series) and $z_{t1}, ..., z_{t2}$ (independent series) we model

$$x_t = \beta_0 + \beta_1 z_{t1} + ... + \beta_q z_{tq} + w_t$$

where w_t is assumed white noise.

Note: w_t is seldom white noise in practice, used as a tool for detrending!

Trend removal by regression

Still a linear regression in β

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad Z = \begin{pmatrix} 1 & z_{11} & \cdots & z_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & \cdots & z_{nq} \end{pmatrix}$$

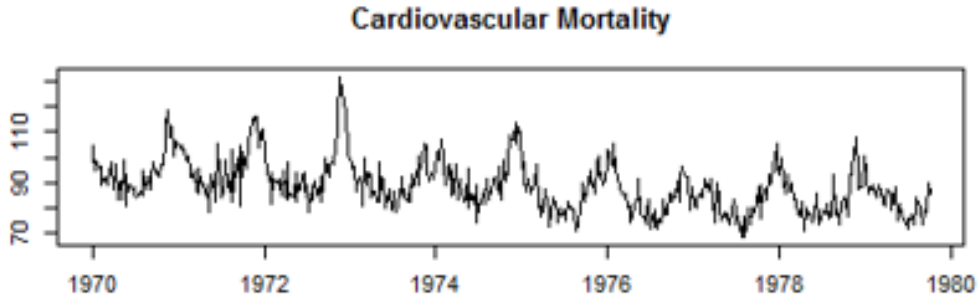
Least squares estimate is computed as

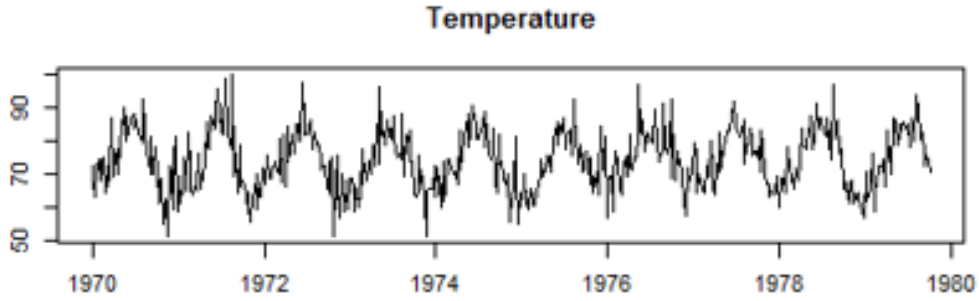
$$\hat{\beta} = (Z^T Z)^{-1} Z^T X$$

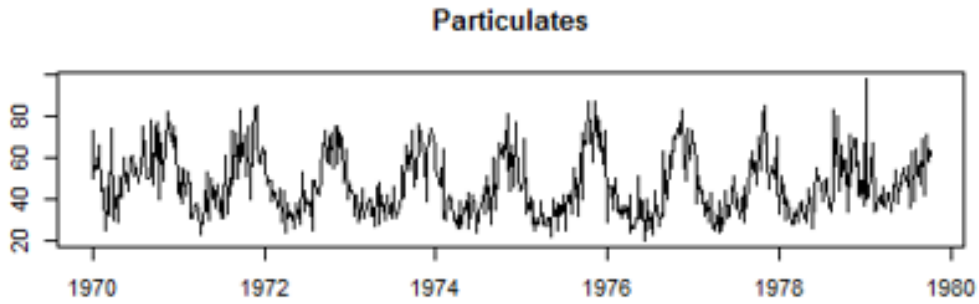
Trend removal

Example: Mortality

- x_t: Cardiovascular mortality
- z_{t1}: Temp (centered)
- z_{t2}: Temp (centered, squared)
- z_{t3}: Time
- z_{t4}: Levels of particles

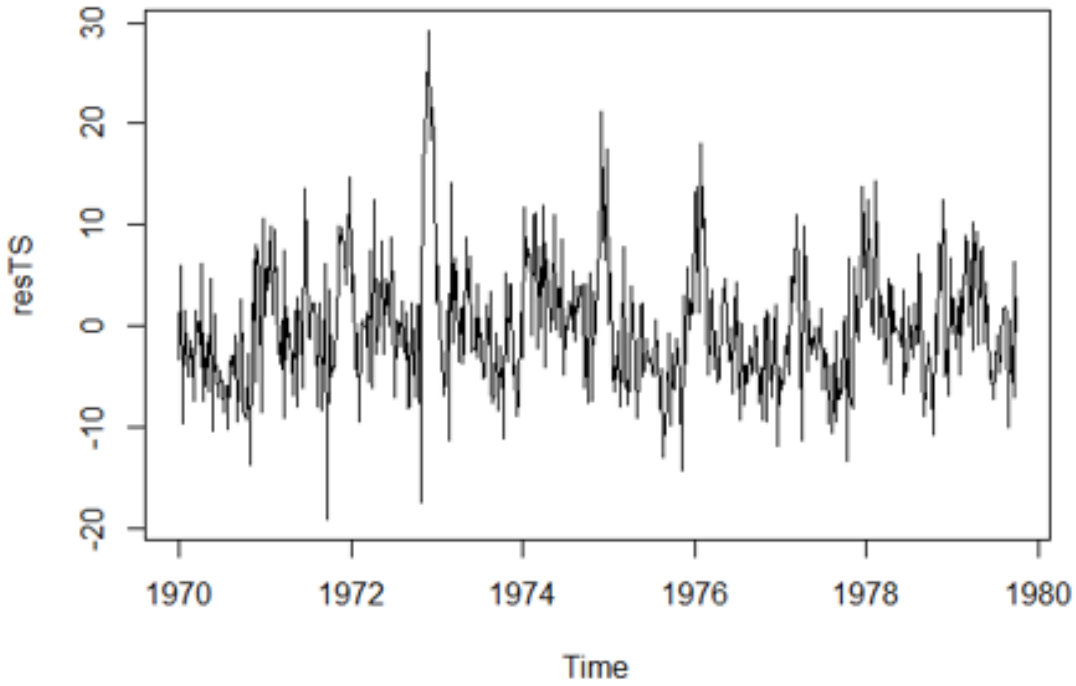






Trend removal

- Residuals
  - Stationary?
  - Independent?
  - Some additional modeling of the residuals (ARIMA) can be done



Differencing

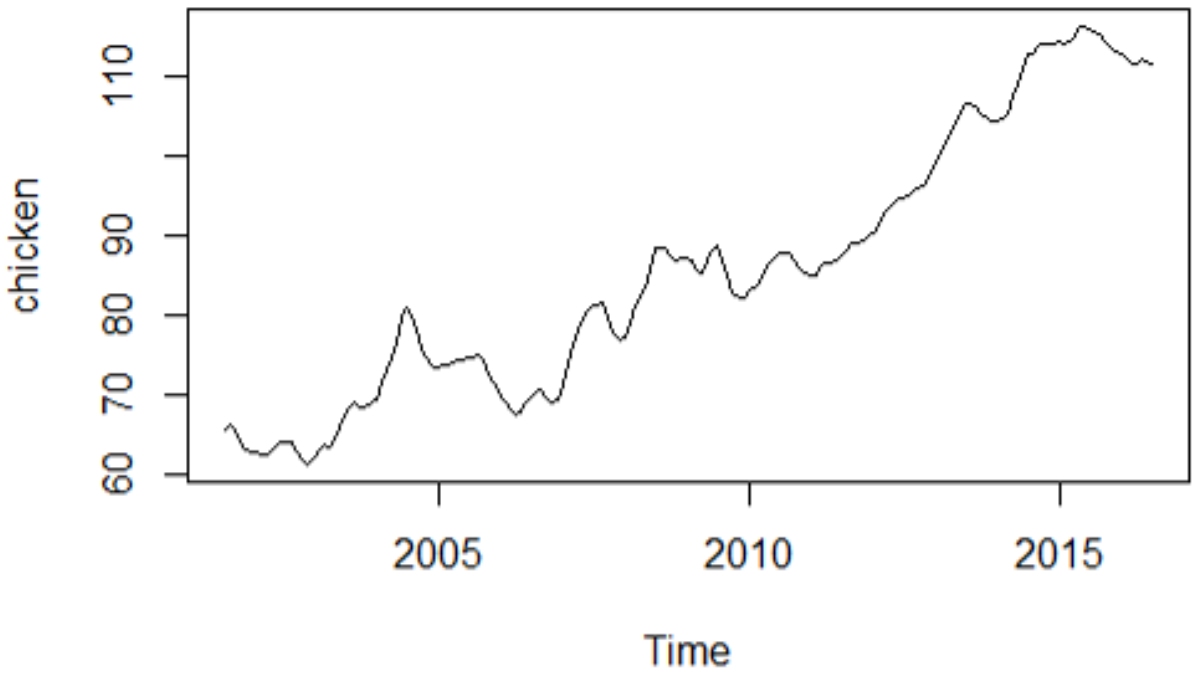
Assume $x_t = \mu_t + y_t$, y_t stationary

Differencing gives $z_t = \nabla x_t = x_t - x_{t-1}$

- Property 1: If $\mu_t = \alpha_0 + \alpha_1 t$ then z_t is stationary
- Property 2: If μ_t is random walk with a drift then z_t is stationary

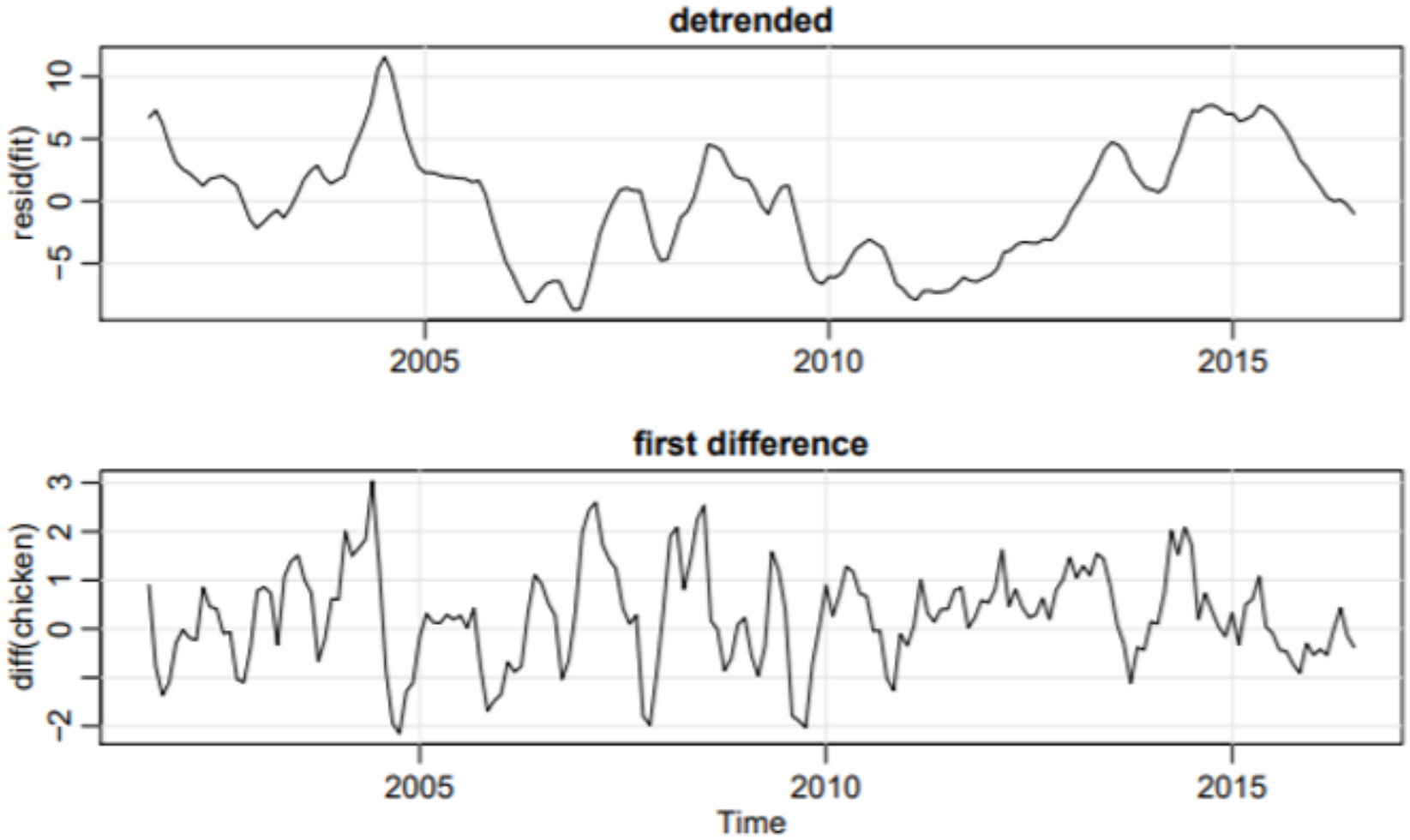
Example:

Chicken prices

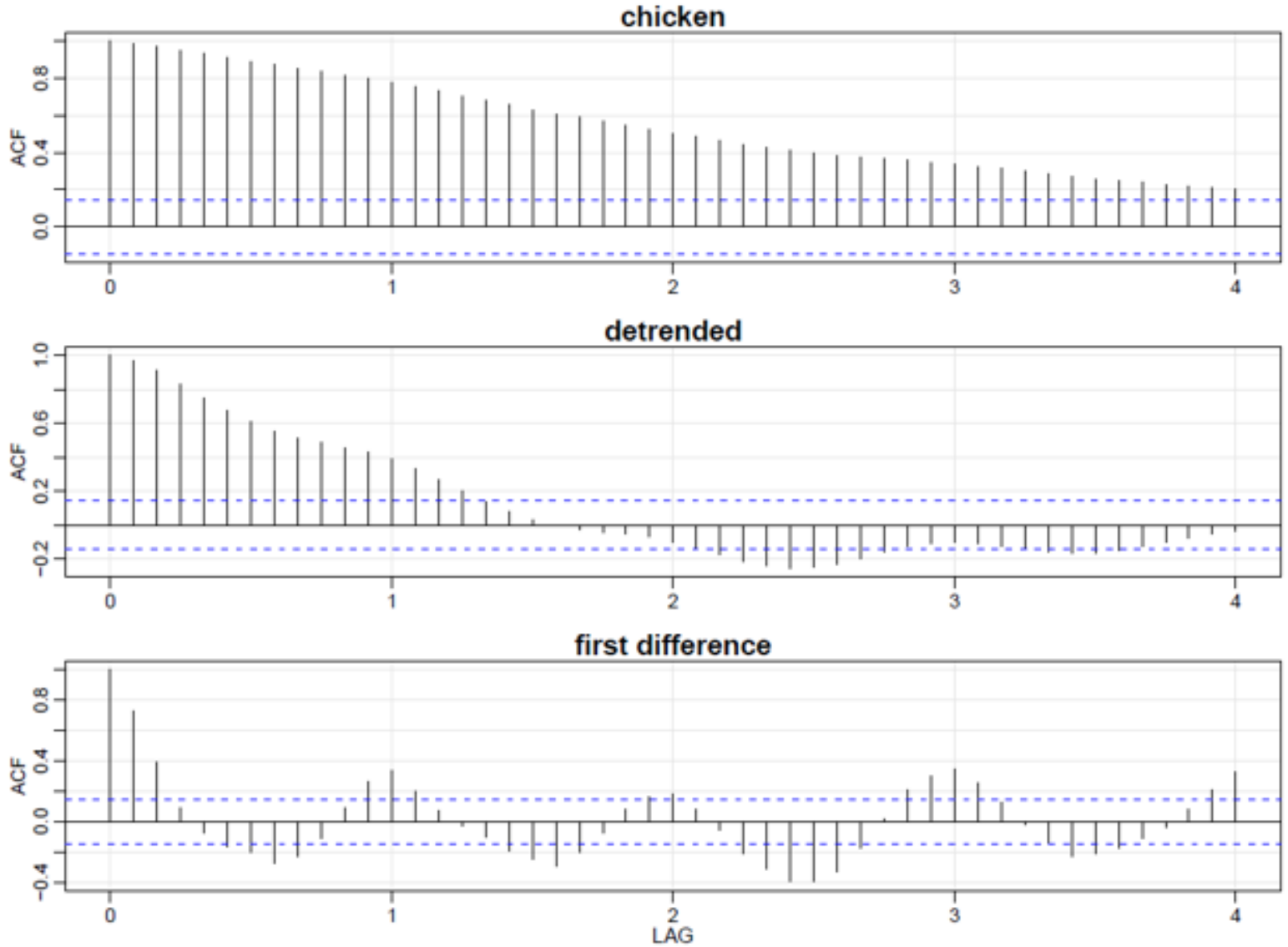


Differencing

Which looks **most random**? Other differences?



Differencing



Detrending vs differencing

- Differencing is more flexible than linear detrending
- Differencing does not require model estimation
- If trend is complex, detrending with a flexible (machine learning) model can be better
- Differencing does not give us the trend

Backshift operator

- Backshift operator $Bx_t = x_{t-1}$, Powers $B^k x_t = x_{t-k}$
- Forward-shift operator $B^{-1}x_t = x_{t+1}$
- Note $BB^{-1}x_t = x_t$ (i.e. $BB^{-1} = 1$)
- Differencing $\nabla x_t = (1 - B)x_t$
- Differences of order d: $\nabla^d = (1 - B)^d$
- Property: Operators can be manipulated as polynomials
- Example Check that $\nabla^2 x_t = x_t - 2x_{t-1} + x_{t-2}$
- Property: Differencing of order p can remove polynomial trend of order p

Transformations

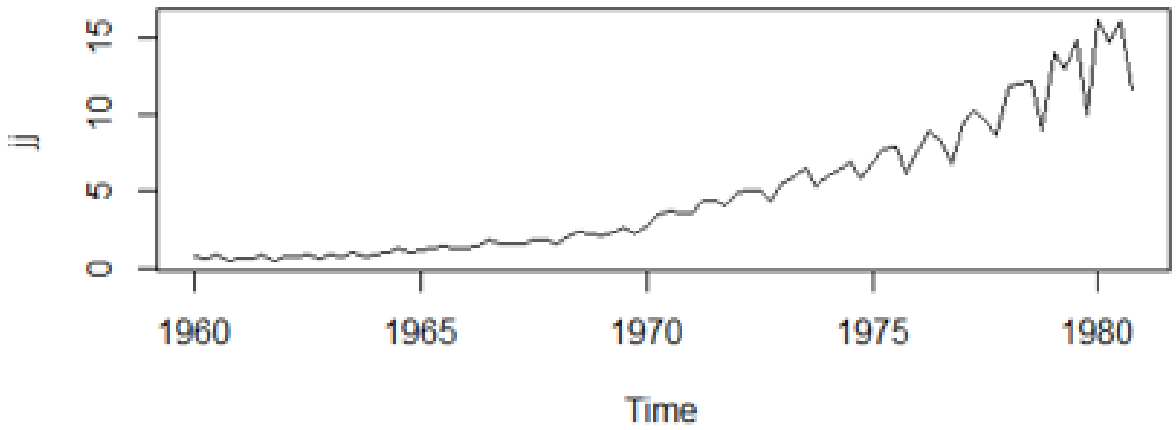
- Often used to stabilize variance
  - If for ex.var$(x_t) \neq$ var(x_s) then time series is non-stationary $\cdots$
- Sometimes makes data more similar to normal distr.

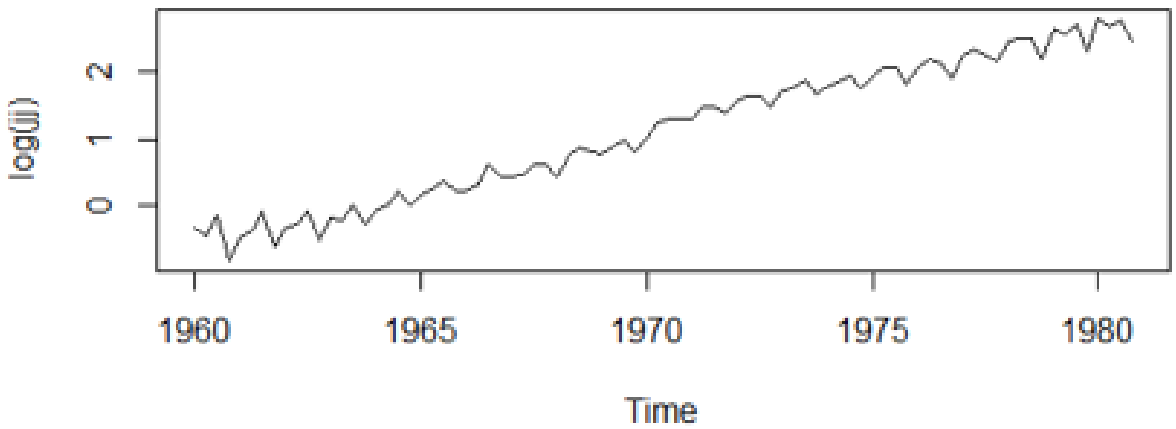
- Common transforms:
  - $z_t = \log(x_t)$
  - Power transformation

$$z_t = \begin{cases} \frac{(x_t^\lambda - 1)}{\lambda} & \lambda \neq 0 \\ \log(x_t) & \lambda = 0 \end{cases}$$

Transformations

- Johnson & Johnson quarterly earnings

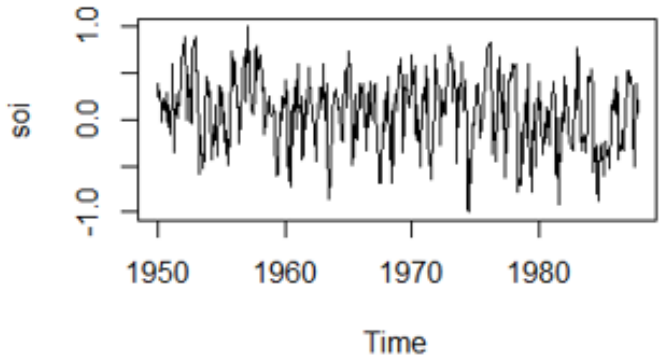




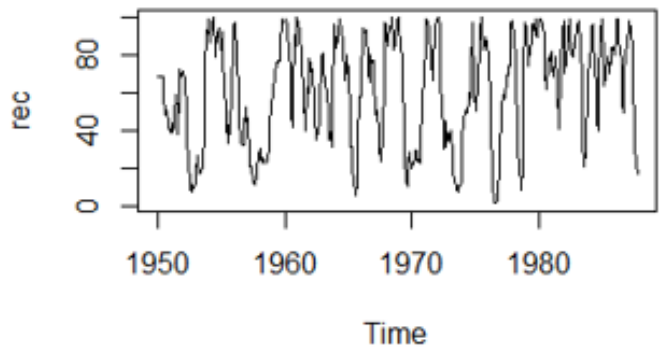
Scatterplots

- Plot x_t vs z_{t_i} or z_{t_i} vs z_{t_j}
- Exploratory tool: indicates which relationship to model

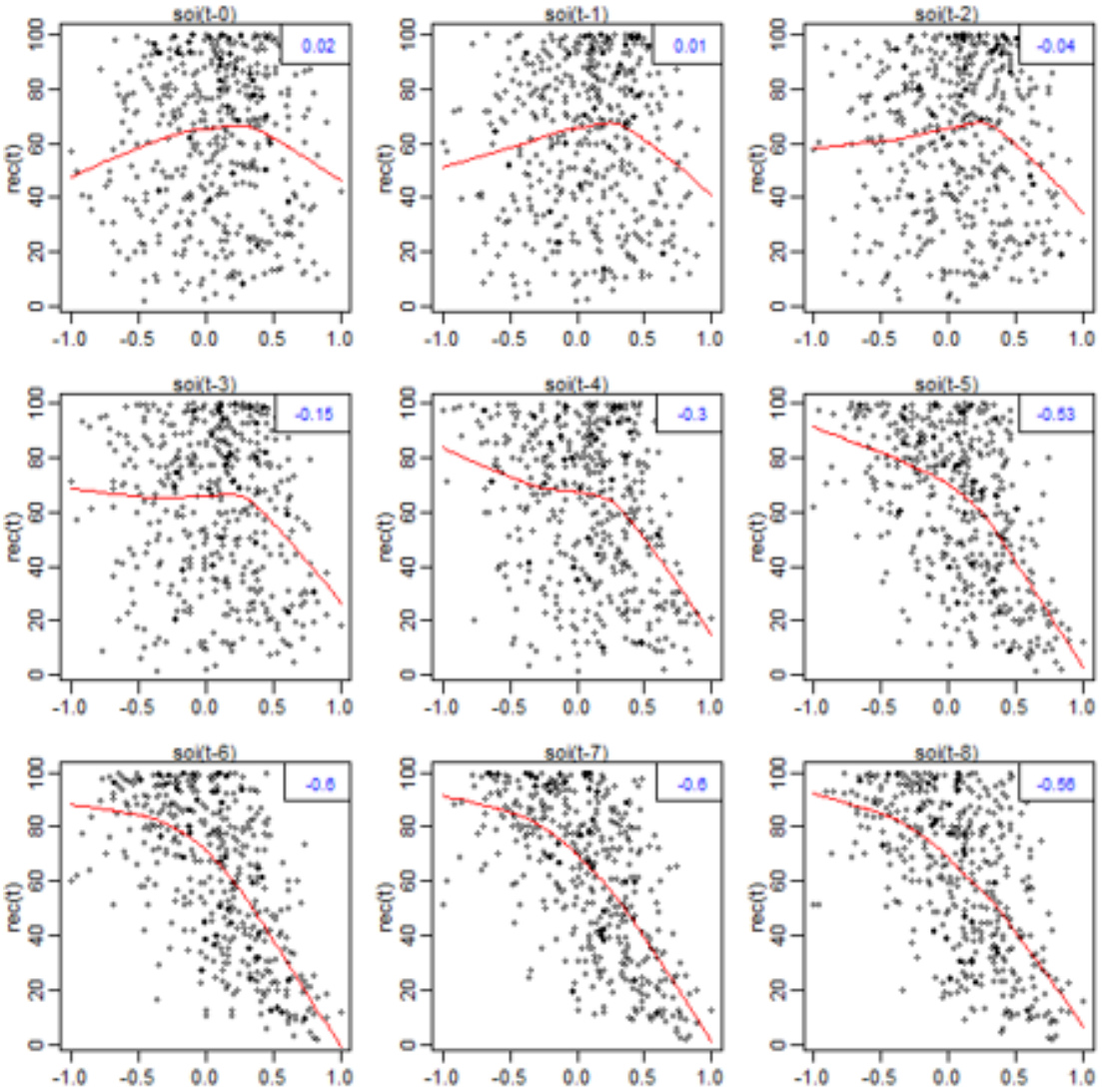
$$x_t = f(z_{t_1}, z_{t_2}, \cdots, z_{t_q}) + w_t$$



- Example: SOI and Recruitment



Scatterplots



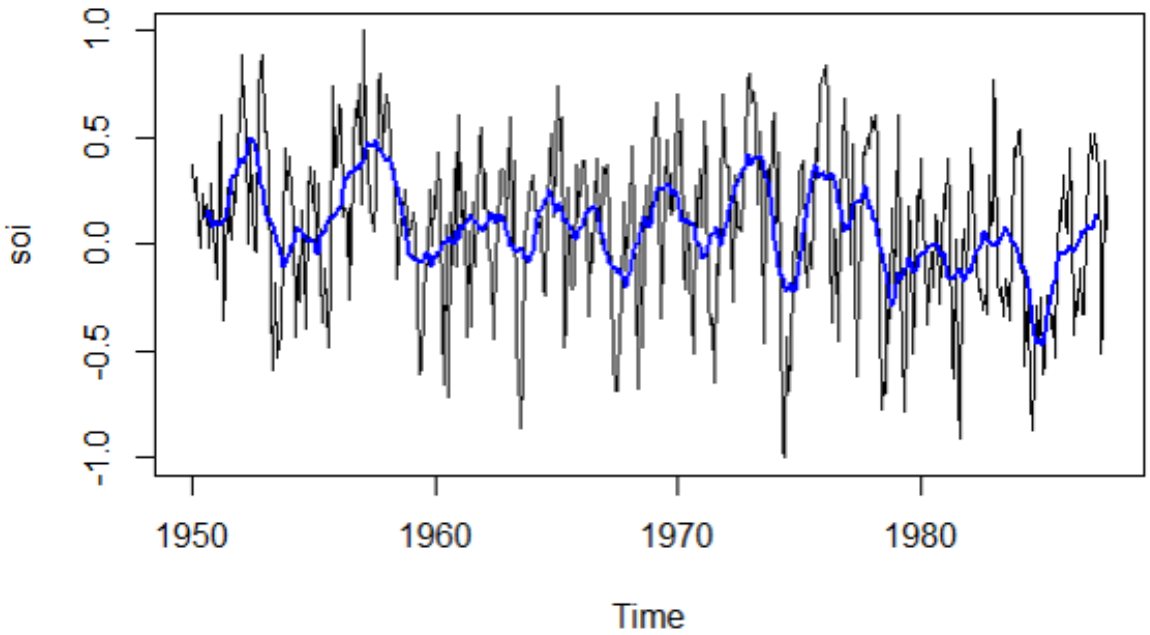
- Which relationships are nonlinear?
- Conclusion: include dummy variables $I(soi(t-j) > 0)$ in the linear model

Smoothing

- Moving average smoother

$$m_t = \sum_{j=-k}^{j=k} a_j x_{t-j}$$

- Where $\sum_{j=-k}^{j=k} a_j = 1$ and $a_j = a_{-j} \geq 0$,
- Example: SOI data Disadvantage?



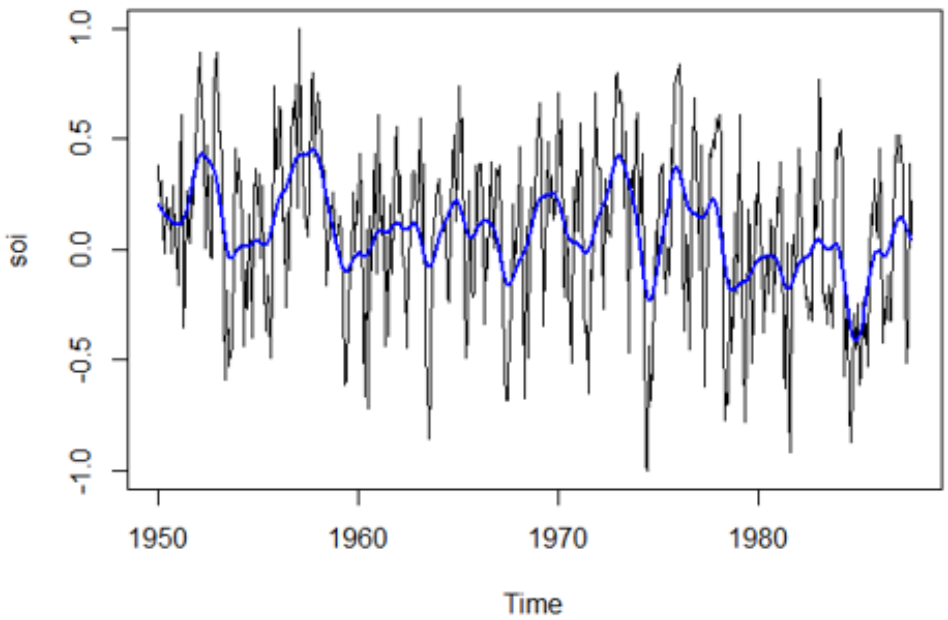
Smoothing

More flexible models?

- Splines
- Kernel smoothers
- Gaussian Process
- Neural networks
- ...

Welcome to ML courses!!

Example: kernel smoothers



Home reading

- Shumway and Stoffer, sections 1.4-1.6 and chapter 2
- TS functions: lag, ksmooth, lm, diff, lag1.plot, lag2.plot

Time Series Analysis

Lecture 3: Introduction to ARIMA

Tohid Ardeshiri

Linköping University
Division of Statistics and Machine Learning

September 6, 2019

Recap

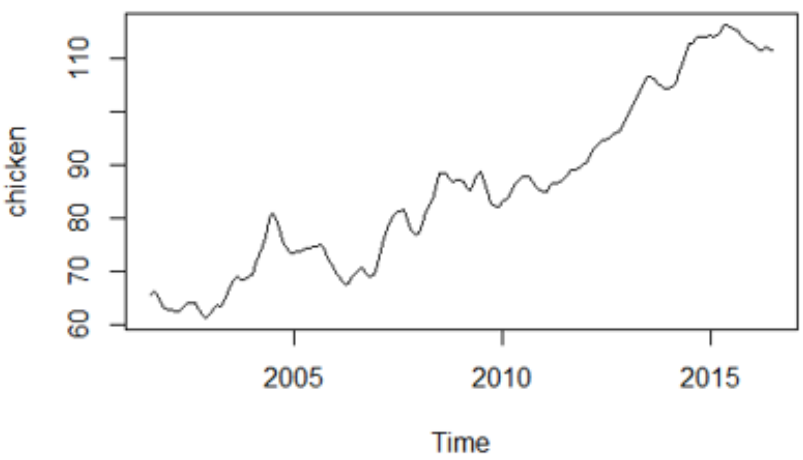
How to make data stationary?

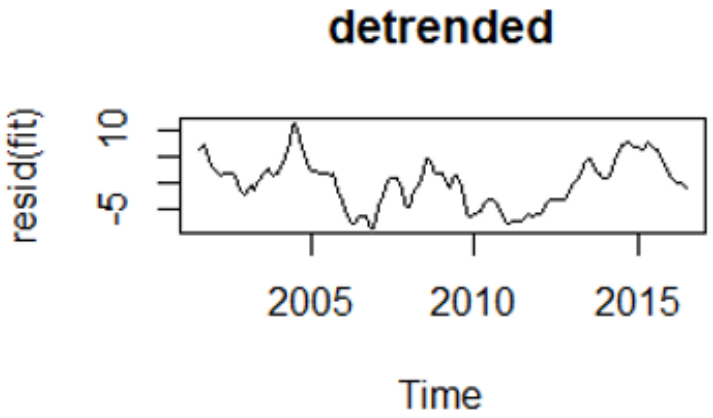
- Transformations (log, other)
- Detrending
  - Differencing
  - Linear regression
  - Kernel smoother
  - ...

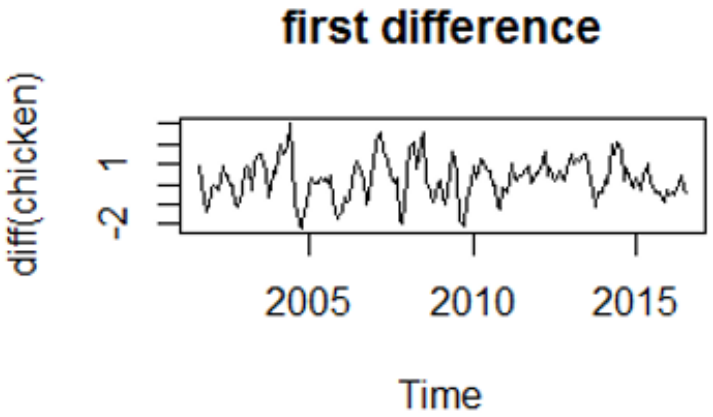
How shall we model the data after detrending and transformations (residuals)? $\rightarrow$?**ARIMA models!**

ARIMA models

- Why ARIMA models?
  - Removing trend is not sufficient







Moving average models

- **Moving average model of order q, MA(q)**

$$x_t = w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}$$
$$= \sum_{j=0}^{q} \theta_j w_{t-j}$$

  - $w_t \sim wn(0, \sigma_w^2)$
  - $\theta_1, ...\theta_q$ constants, $\theta_q \neq 0$ and $\theta_0 = 1$

- **Moving average operator**

$$\theta(B) = \sum_{j=0}^{q} \theta_j B^j$$

- MA(q): $x_t = \theta(B) w_t$

Linear process

x_t is a **linear process** if

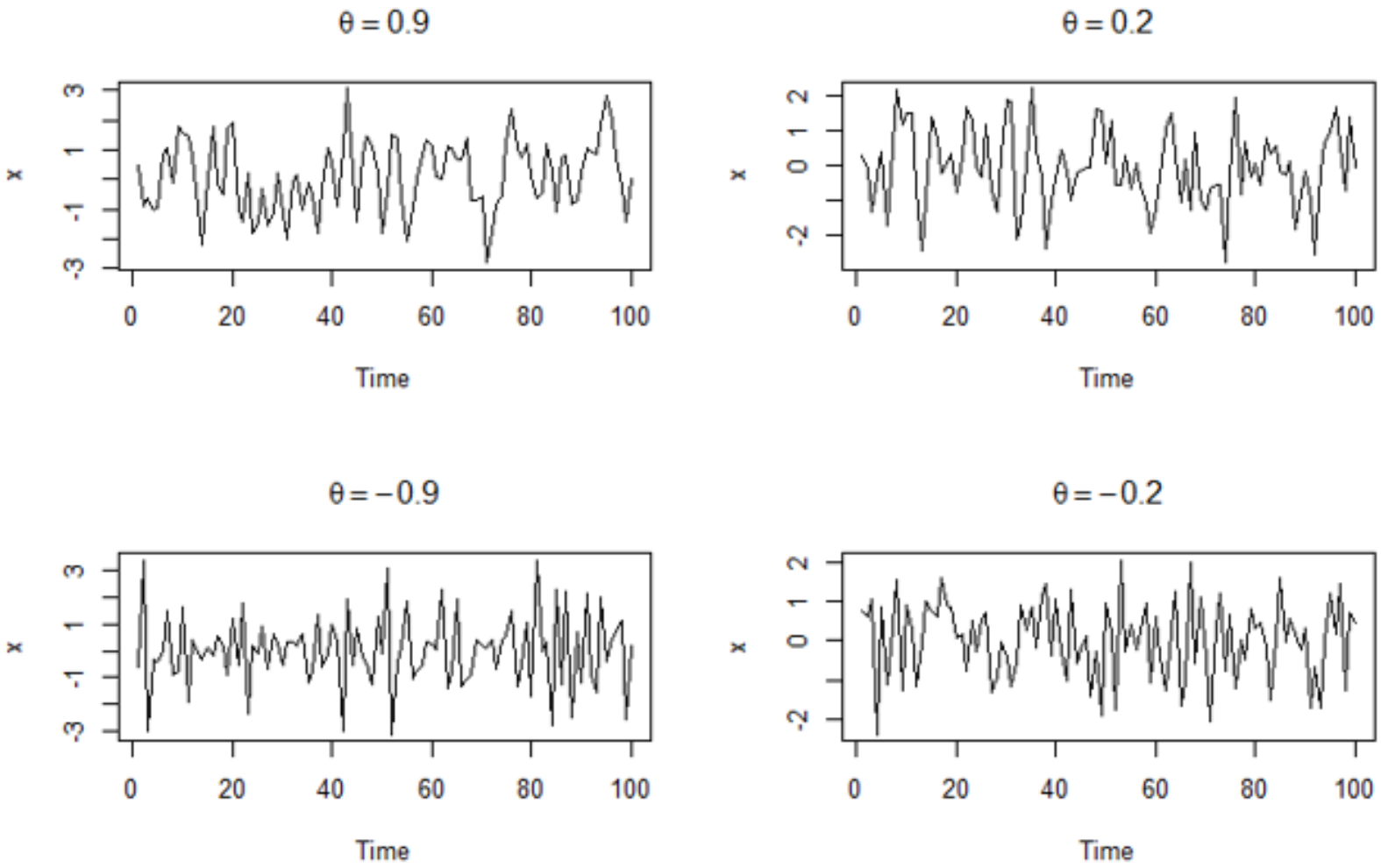
$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}$$

Property: It can be shown that

$$\gamma_x(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j$$

Example: MA(1)

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



Example: MA(1)

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

- Autocovariance and ACF

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

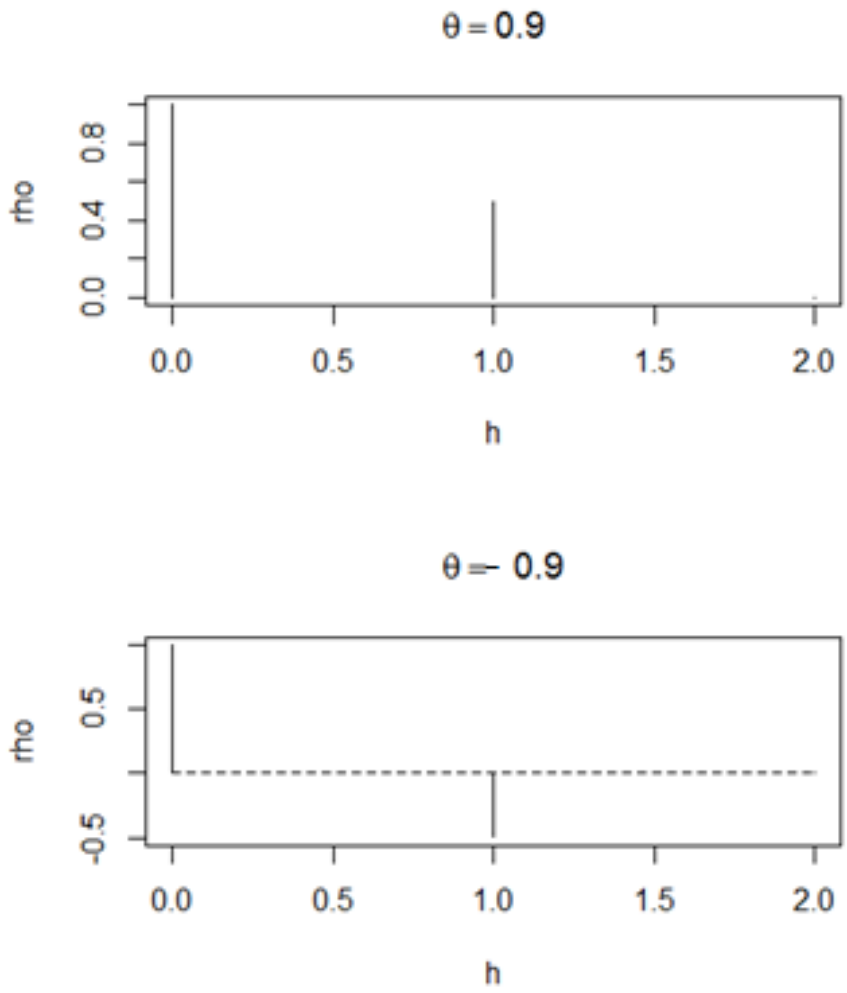
$$\rho(h) = \begin{cases} \frac{\theta}{1+\theta^2} & h = 1 \\ 0 & h > 1 \end{cases}$$

Note: $\rho(0) = 1$ is often not written as it is trivial.

- Process is stationary

Example: MA(1)

- Note: $\rho(0) = 1$ is often not shown $\rightarrow$ only 1 bar



AR models

- **Autoregressive model of order p, $AR(p)$**

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t$$

  - x_t is stationary if x_0 is sampled from the stationary distribution
  - $w_t \sim wn(0, \sigma_w^2)$
  - $\phi_1, ...\phi_p$ constants, $\phi_p \neq 0$
  - $Ex_t = 0$

- Note: if $Ex_t = \mu \neq 0$, model $x'_t = x_t - \mu$

AR models

Another form

- **Autoregressive operator**

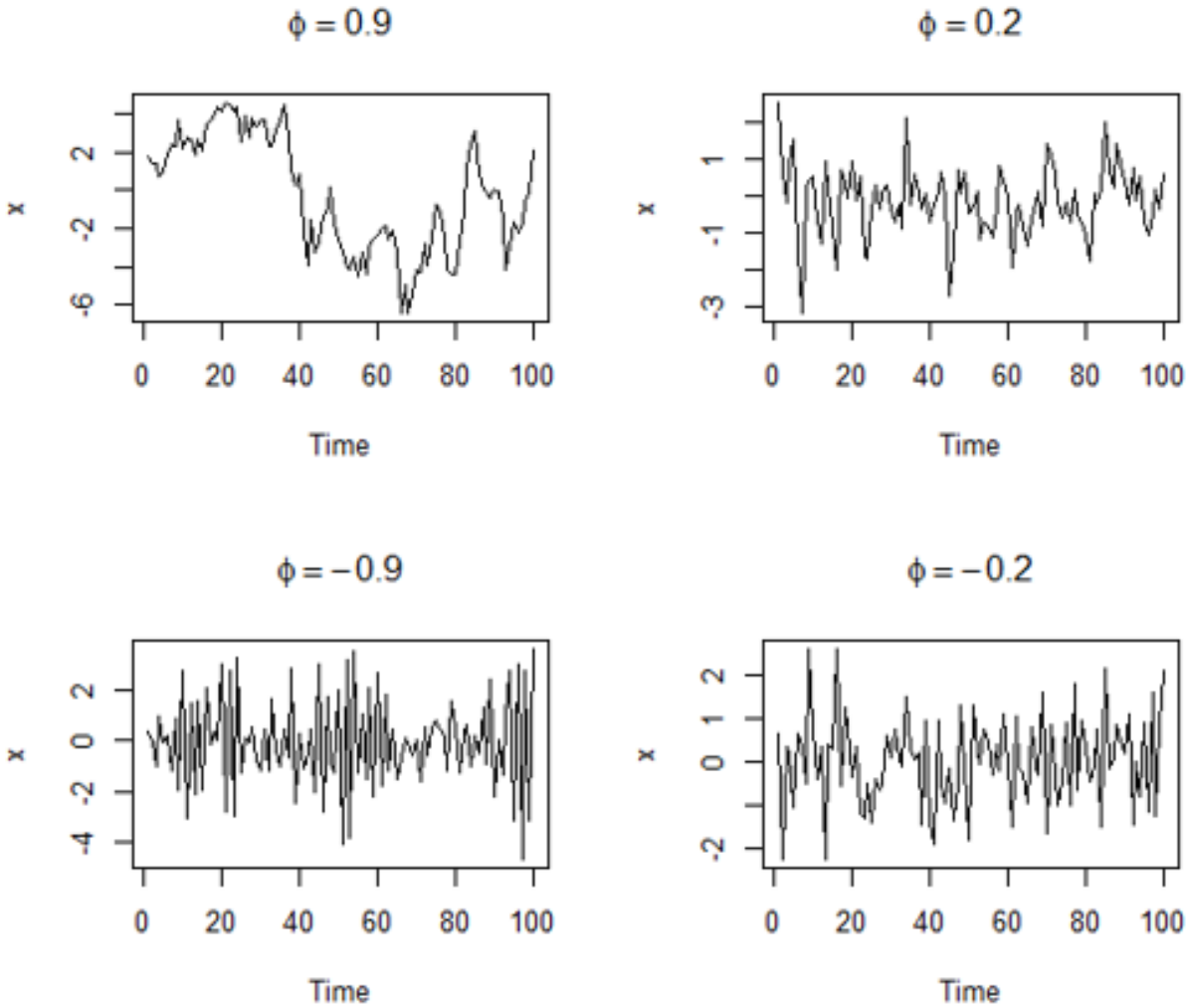
$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$$

- AR(p) model

$$\boxed{\phi(B)x_t = w_t}$$

Example: AR(1)

- How do these plots differ? $x_t = \phi x_{t-1} + w_t$



Ar(1) (read at home)

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

Mean function:

$Ex_t = \phi Ex_{t-1} + Ew_t = \phi Ex_{t-1} = \phi(\phi Ex_{t-2}) = \cdots = \phi^t Ex_0$

for $Ex_0 = 0$, $Ex_t = 0$ for all t.

Variance $\text{var}(x_t)$ when $Ex_0 = 0$ and w_t is uncorrelated with x_0 for all t:

$\text{var}(x_t) = E\{(x_t - 0)^2\} = E\{\phi^2 x_{t-1}^2 + 2\phi x_{t-1} w_t + w_t^2\} =$

$\phi^2 \text{var}(x_{t-1}) + 2\phi \,\text{cov}(x_{t-1}, w_t) + \text{var}(w_t) = \phi^2 \text{var}(x_{t-1}) + \text{var}(w_t) =$

$\phi^2 \text{var}(x_{t-1}) + \sigma_w^2 = \phi^2(\phi^2 \text{var}(x_{t-2}) + \sigma_w^2) + \sigma_w^2 =$

$\phi^{2t} \text{var}(x_0) + \sigma_w^2 \sum_{k=0}^{t-1}(\phi^{2k}) = \phi^{2t} \text{var}(x_0) + \frac{\sigma_w^2(1-\phi^{2t})}{1-\phi^2}$

When $\text{var}(x_0) = \frac{\sigma_w^2}{1-\phi^2}$ then $\text{var}(x_t) = \frac{\sigma_w^2}{1-\phi^2}$ and time independent.

A(1) (read at home)

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

$$x_t = \phi(\phi x_{t-2} + w_{t-1}) + w_t = \cdots = \phi^h x_{t-h} + \sum_{j=0}^{h-1} \phi^j w_{t-j}$$

$$\gamma(x_t, x_{t-h}) = \text{cov}(x_t, x_{t-h}) = E(x_t x_{t-h}) =$$
$$E\{(\phi^h x_{t-h} + \sum_{j=0}^{h-1} \phi^j w_{t-j}) x_{t-h}\} = \phi^h \,\text{var}(x_{t-h}) = \frac{\phi^h \sigma_w^2}{1-\phi^2}$$

Hence,

$$\gamma(h) = \frac{\phi^h \sigma_w^2}{1-\phi^2}$$

Also,

$$\rho(h) = \phi^h$$

whiteboard

- **Property:** If $|\phi| < 1$ and sup $\text{var}(x_t) < \infty$

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}$$

- Show it by
  - Substitution
  - Taylor expansion
  - Coefficient matching
- Autocovarince and ACF

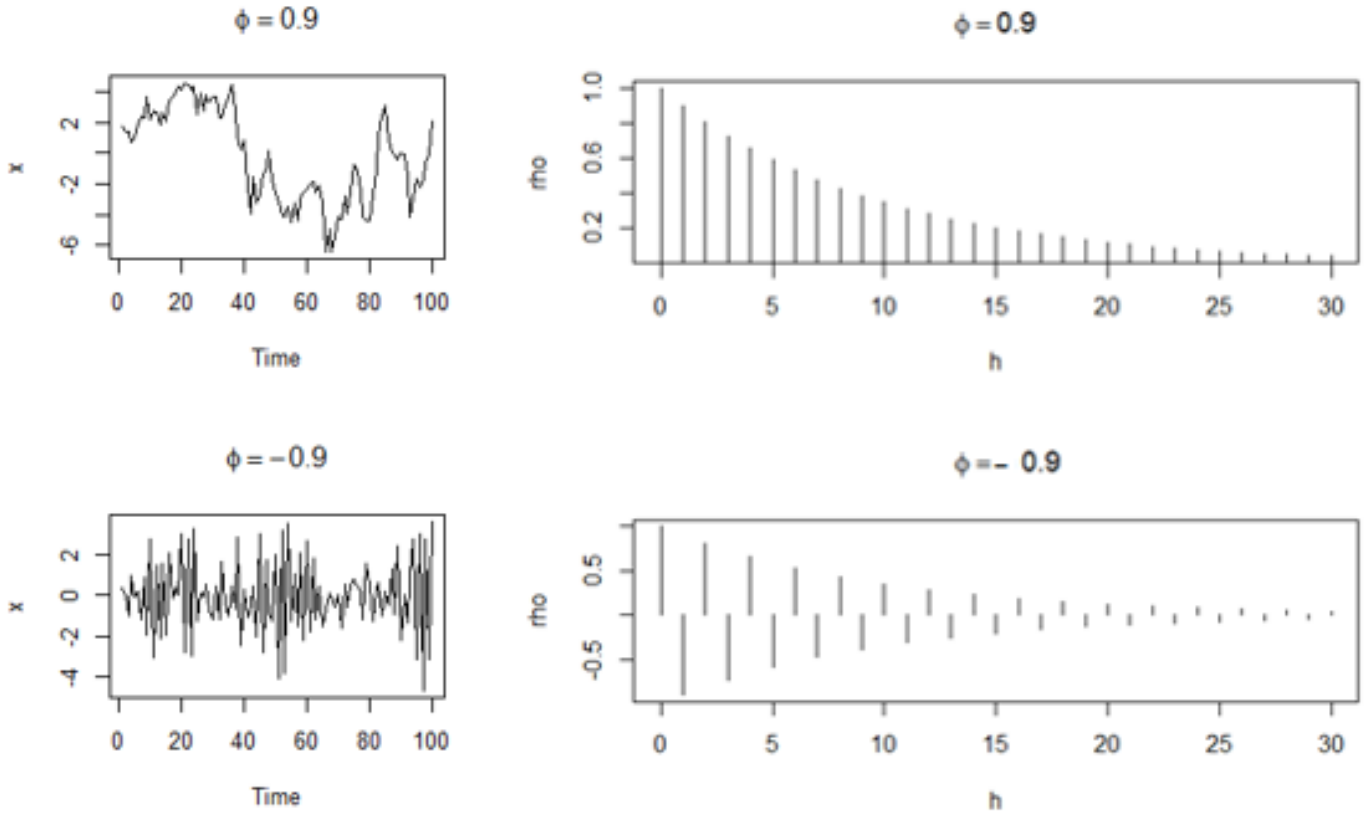
$$\gamma(h) = \frac{\sigma_w^2 \phi^h}{1-\phi^2} \quad \rho(h) = \phi^h$$

for $h \geq 0$.

Example: AR(1)

Autocovarince and ACF (for $h \geq 0$)

$$\gamma(h) = \frac{\sigma_w^2 \phi^h}{1 - \phi^2} \quad \rho(h) = \phi^h$$



Explosive AR models

- Explosive =series become arbitrarily large in magnitude
- AR(1): What if $|\phi| > 1$?
  - $x_t = \phi^p x_{t-p} + \sum_{j=0}^{p-1} \phi^p w_{t-j} \rightarrow$ grows exponentially
  - Stationary? Check variance

- Can we make it stationary?
  $x_t = \phi^{-1} x_{t+1} - \phi^{-1} w_{t+1} = \phi' x_{t+1} + w'_t$
  - Stationary, but dependent on the future
  - $w'_t \sim N(0, \phi^{-2}\sigma_w^2)$
  - $x_t = -\sum_{j=1}^{\infty} \phi_{-j} w_{t+j}$

Causal process

A stationary process is **causal** if it is only dependent on the past values of the process

Def: A linear process is nonexplosive and causal if it can be written as a one-sided sum:

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

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

MA equivalence

Whiteboard

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

Note: MA(1) gives equivalent models for $\theta = s$ and $\theta = \frac{1}{s}$
Probabilistic expressions equivalent: ACF identical
$\rightarrow$ we can not distinguish between these models

Invertibility of MA

Def: An MA process is **invertible** if it has a causal AR representation,

$$w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}$$

Example: MA(1) with $\theta = 1/5$ is invertible, $\theta = 5$ not.

ARMA models

- **Autoregressive moving average ARMA(p,q)**

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}$$

  - $\phi_p \neq 0, \theta_q \neq 0$
  - Is stationary
  - $Ex_t = 0$

- p-autoregressive order, q-moving average order
- Alternative form

$$\boxed{\phi(B)x_t = \theta(B)w_t}$$

- Note: $x_t = \phi^{-1}(B)\theta(B)w_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}$
  - But series might be non-convergent

Parameter redundancy

Note: we can multiply both sides with $\eta(B)$

$$\eta(B)\phi(B)x_t = \eta(B)\theta(B)w_t$$

- The resulting model looks different (higher orders)
- Underlying model is actually the same

Example: $x_t = w_t$, white noise. Let $\eta(B) = 1 - 0.5B$.
We get

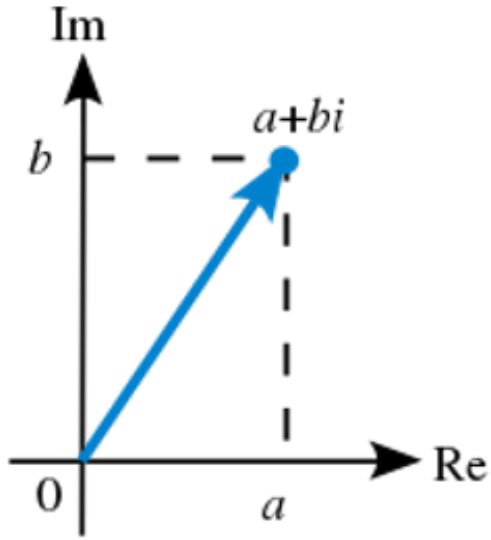
$$x_t - 0.5x_{t-1} = w_t - 0.5w_{t-1}$$

Looks like ARMA(1,1)!

Reminder: complex numbers

- Imaginary unit $i^2 = -1$
- Complex number z=a+ib
- Conjugate $\bar{z} = a - ib$

- Absolute value $|z|^2 = z\bar{z} = a^2 + b^2$
- Trigonometic form
  $z = r(\cos(\theta) + i\sin(\theta))$
- Eulers formula $e^{i\theta} = \cos(\theta) + i\sin(\theta))$



- Therefore

$$\cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2} \qquad \sin(\theta) = \frac{e^{i\theta} - e^{-i\theta}}{2i}$$

Reminder: polynomials

- Any polynomial $P_r(x)$ of degree r can be written as

$$P_r(x) = a(x - z_1)...(x - z_r)$$

- where z_i are roots (real or complex)
- If z_i is a root, $\bar{z}_i$ is also a root

Causal ARMA

Def: Linear process is **causal** and **nonexplosive** if

- $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$ (depends on the past only)
- $\sum_{j=0}^{\infty} |\psi_j| < \infty$
- We set $\psi_0 = 1$ by convention.

Property: ARMA(p,q) is **causal** iff roots $\phi(z') = 0$ are outside unit circle, i.e. $|z'| > 1$

$$\boxed{\phi(B)x_t = \theta(B)w_t}$$

Causal ARMA

Example: Is the ARMA process below causal?

$$x_t = 0.4x_{t-1} + 0.3x_{t-2} + 0.2x_{t-3} + w_t - 0.1w_{t-1}$$
$$\Rightarrow \phi(B) = 1 - 0.4B - 0.3B^2 - 0.2B^3$$

```
> z=c(1, -0.4,-0.3,-0.2)
> polyroot(z)
[1]  1.060419-0.000000i -1.280210+1.753904i -1.280210-1.753904i
>
```

Invertible ARMA

Def: ARMA(p,q) is **invertible** if

- $w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}$ (depends on the past only)
- $\sum_{j=0}^{\infty} |\pi_j| < \infty$

Property: ARMA(p,q) is **invertible** iff roots $\theta(z') = 0$ are outside unit circle, i.e. $|z'| > 1$

$$\boxed{\phi(B)x_t = \theta(B)w_t}$$

Coefficient matching **whiteboard**

- $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} \rightarrow x_t = \psi(B) w_t$
- $w_t = \sum_{j=0}^{\infty} \pi_j w_{t-j} \rightarrow w_t = \pi(B) x_t$
- How to find coefficients in ψ and $\pi \rightarrow$ **coefficient matching**

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

- Example: $x_t = 0.4x_{t-1} + 0.45x_{t-2} + w_t + w_{t-1} + 0.25w_{t-2}$

```
> ARMAtoMA(ar=.9,ma=0.5, 6)
[1] 1.400000 1.260000 1.134000 1.020600 0.918540 0.826686
```

Home reading

- Shumway and Stoffer, section 3.1
- R code: arima.sim, arima, polyroot, ARMAtoMA, ARMAacf
  - Check carefully arima() docs to see how ar and ma coefficients are specified in the software

Time Series Analysis

Lecture 4: ARIMA models-1, Estimation

Tohid Ardeshiri

Linköping University
Division of Statistics and Machine Learning

September 13, 2019

White noise

Simplest and most random time series: white noise

- w_t uncorrelated $E(w_t w_{t-h}) = 0$ for all $h \neq 0$

$$w_t \sim wn(0, \sigma_w^2)$$

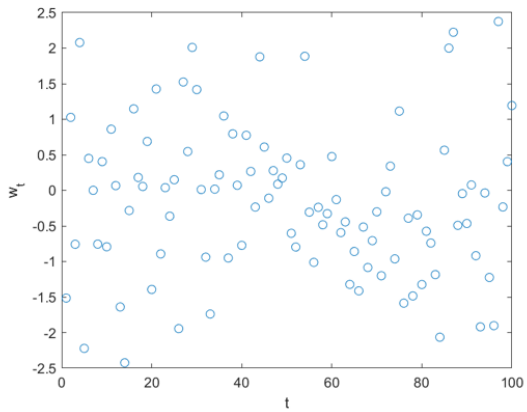
- w_t white independent noise: independent and identically distributed

  independence: $f(w_t, w_{t-h}) = f(w_t) f(w_{t-h})$

$$w_t \sim iid(0, \sigma_w^2)$$

- w_t white normal noise: independent and identically normal distributed

$$w_t \sim iidN(0, \sigma_w^2)$$



Autocovariance and ACF

- Autocovariance function

$$\gamma(s,t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$$

  Note $\text{var}(x_t) = \gamma(t,t)$

- Autocorrelation function (ACF)

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

Useful fact: If $U = \sum_{j=1}^{m} a_j x_j$ and

$$V = \sum_{k=1}^{r} b_k y_k$$

$$\text{cov}(U,V) = \sum_{j=1}^{m}\sum_{k=1}^{r} a_j b_k \,\text{cov}(x_j, y_k)$$

Stationarity

- Time series x_t is weakly stationary (stationary) if
  - $Ex_t = const$
  - $\gamma(s,t) = \gamma(|s-t|)$
  - $\text{var}(x_t) < \infty$
- $\gamma(t, t+h) = \gamma(|t+h-t|) = \gamma(h)$
  - Autocovariance depends on lag only!
- Autocovariance for stationary process $\gamma(h) = \text{cov}(x_t, x_{t+h})$
- ACF for stationary process $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$

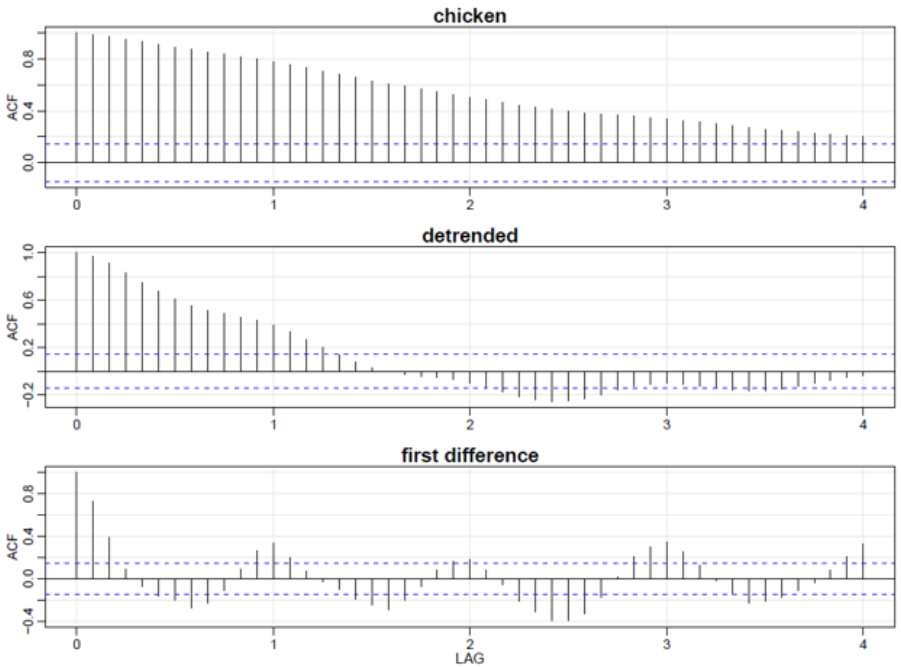
Sample ACF

Theorem: Under weak conditions, if x_t is white noise and $n \to \infty$ then $\hat{\rho}(h)$ is approximately $N(0, \frac{1}{n})$

Consequence: If some $|\hat{\rho}(h)| > \frac{2}{\sqrt{n}}$ then the time series is not a white noise (with approximately 95 % confidence).

Typical modeling strategy:

- Propose a model
- Fit a model
- Compute residuals
- Check ACF within $\pm\frac{2}{\sqrt{n}}$



Moving average models

- **Moving average model of order q, MA(q)**

$$1\ x_t = 1\ w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}$$

$$x_t = \sum_{j=0}^{q} \theta_j w_{t-j}$$

  - $w_t \sim wn(0, \sigma_w^2)$
  - $\theta_1, \ldots \theta_q$ constants, $\theta_q \neq 0$ and $\theta_0 = 1$

- **Moving average operator**

$$\theta(B) = \sum_{j=0}^{q} \theta_j B^j$$

- **MA(q):**

$$\boxed{x_t = \theta(B) w_t}$$

Autoregressive models

- **Autoregressive model of order p, $AR(p)$**

$$\mathbf{1}\ x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + \mathbf{1}\ w_t$$

$$x_t - \sum_{j=1}^{p} \phi_j x_{t-j} = w_t$$

  - x_t is stationary if x_0 is sampled from the stationary distribution
  - $w_t \sim wn(0, \sigma_w^2)$
  - $\phi_1, \ldots \phi_p$ constants, $\phi_p \neq 0$
  - $Ex_t = 0$ if $Ex_0 = 0$

- **Autoregressive operator**

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$$

- **AR(p) model**

$$\boxed{\phi(B) x_t = w_t}$$

ARMA models

- **Autoregressive moving average ARMA(p,q)**

$$\mathbf{1}\ x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + \mathbf{1}\ w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}$$

  - $\phi_p \neq 0, \theta_q \neq 0$
  - Is stationary
  - $Ex_t = 0$ if $Ex_0 = 0$

- p-autoregressive order, q-moving average order
- Alternative form

$$\boxed{\phi(B)x_t = \theta(B)w_t}$$

- Criteria for **causality** and **invertibility**
  - Check roots of the characteristic polynomials $\phi(\cdot)$ and $\theta(\cdot)$

Property: ARMA(p,q) is **causal** iff **ALL** roots $\phi(z') = 0$ are outside unit circle, i.e. $|z'| > 1$

Property: ARMA(p,q) is **invertible** iff **ALL** roots $\theta(z') = 0$ are outside unit circle, i.e. $|z'| > 1$

Linear process

For a **linear process** x_t: $x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j} = \mu + \psi(B) w_t$
where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$,

$$\gamma_x(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h} \psi_j$$

Note: $x_t = \phi^{-1}(B)\theta(B) w_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}$ But series might be non-convergent

- Coefficient matching **whiteboard**
- How to find coefficients in $\psi(B) \rightarrow$ **coefficient matching**
- Example: $x_t = 0.4x_{t-1} + 0.45x_{t-2} + w_t + w_{t-1} + 0.25w_{t-2}$

```
> ARMAtoMA(ar=.9,ma=0.5, 6)
[1] 1.400000 1.260000 1.134000 1.020600 0.918540 0.826686
```

Differencing

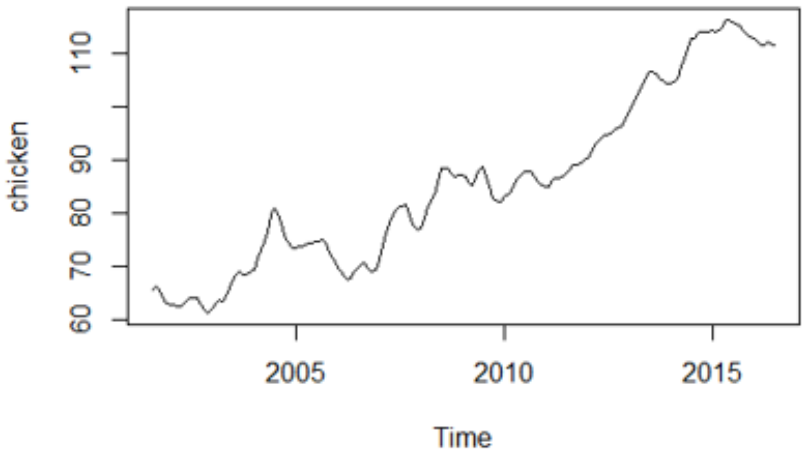
Assume $x_t = \mu_t + y_t$ and y_t stationary

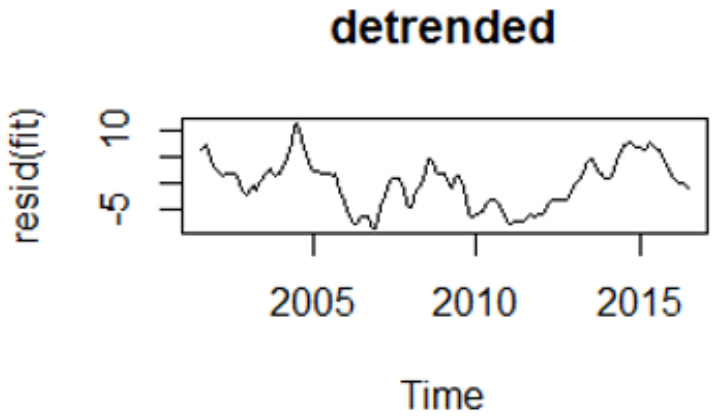
Differencing gives

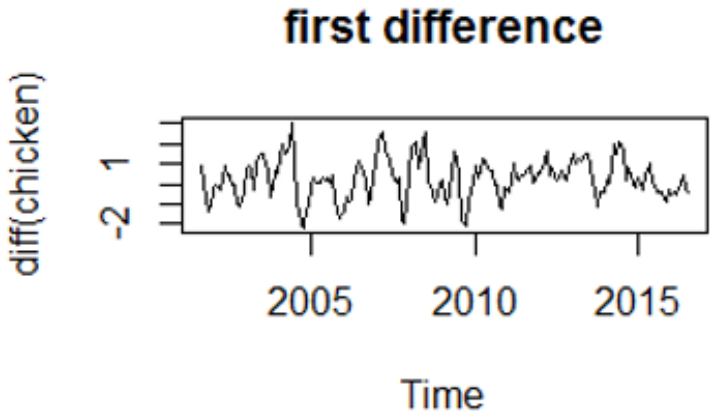
$z_t = \nabla x_t = x_t - x_{t-1}$

Also,

- $\nabla x_t = (1 - B)x_t$
- $\nabla^d = (1 - B)^d$

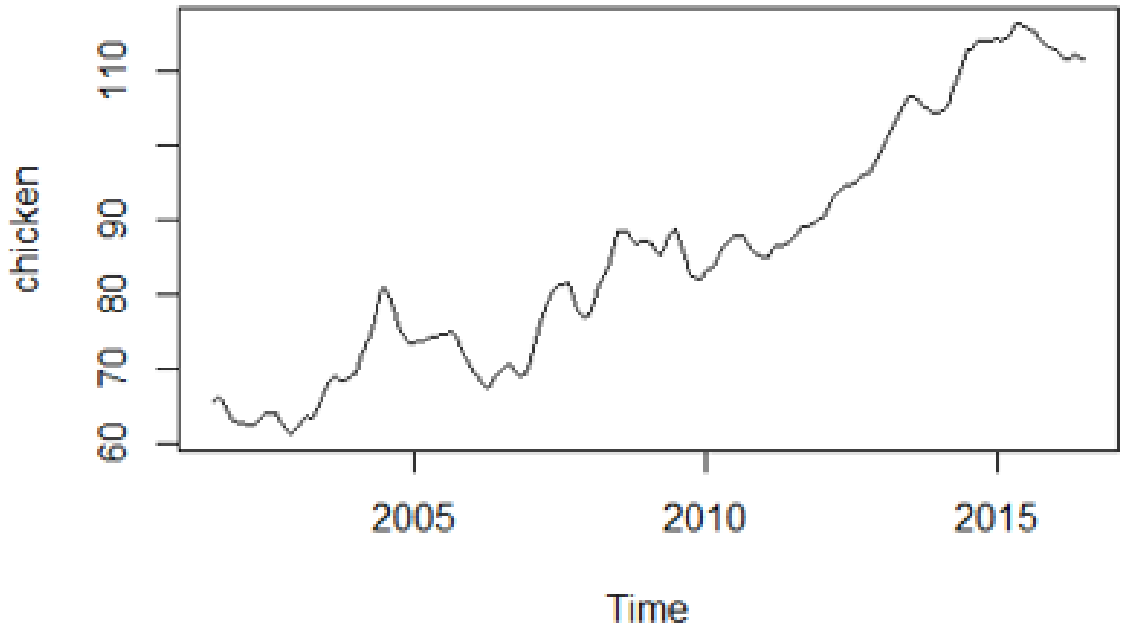






ARIMA models

- ARMA for stationary models
  - What if not stationary?



ARIMA models

- Differencing helps (lecture 2)
  - $\nabla x_t = x_t - x_{t-1}$ removes linear trend and random walk
  - $\nabla^d x_t$ removes polynomial of order d and some stochastic trends
  - $\rightarrow$ differencing is important modeling instrument!

- **Def:** x_t is **ARIMA(p,d,q)** if $\nabla^d x_t$ is ARMA(p,q), i.e.

$$\phi(B)(1-B)^d x_t = \theta(B) w_t$$

- For nonzero mean $E(\nabla^d x_t) = \mu$,

$$\phi(B)(1-B)^d x_t = \theta(B) w_t + \delta$$

$$\delta = \mu(1 - \phi_1 - ... - \phi_p)$$

ARIMA models

- Notation: p=0 $\rightarrow$ IMA(d,q), q=0 $\rightarrow$ ARI(p,d)
- Estimation: Differenciate + fit ARMA
- Forecasting:
  - Transform data $y_t = \nabla^d x_t$ and forecast ARMA(p,q)
  - Solve $(1-B)^d x_t^n = y_t^n$

Estimation

Consider **ARIMA(p,d,q)**

$$\phi(B)(1-B)^d x_t = \theta(B) w_t$$

- What are the unknowns?
  - Orders p, d and q
  - Parameters $\phi_1, \cdots, \phi_p$ and $\theta_1, \cdots, \theta_q$
  - variance σ_w^2 where $w_t \sim N(0, \sigma_w^2)$
- How to estimate these?
- Assumption: Let us assume for now that we know p, d and q
  - **Maximum likelihood (ML) estimate**
  - **Least squares**

Maximum likelihood estimation: reminder

Let $x \sim f(x|\alpha)$

- Likelihood of α given observations $x_1, \cdots, x_t$ is

$$L(\alpha) = f(x_1, \cdots, x_t|\alpha)$$

- Maximum likelihood: Optimal α

$$\widehat{\alpha} = \arg\max_{\alpha} L(\alpha)$$

- Independent observations: $x_i \overset{iid}{\sim} f(x_i|\alpha)$
- $L(\alpha) = \prod_i f(x_i|\alpha)$
- Negative log-likelihood $l(\alpha) = -\sum_i \log(f(x_i|\alpha))$
- Maximum likelihood α can be obtained from negative log-likelihood

$$\max_{\alpha} L(\alpha) = \min_{\alpha} l(\alpha)$$

Maximum likelihood estimation: reminder

Time series data are NOT independent

- Likelihood of α given observations $x_1, \cdots, x_t$ is

$$L(\alpha) = f(x_1, \cdots, x_t|\alpha)$$

- Maximum likelihood: Optimal α

$$\widehat{\alpha} = \arg\max_{\alpha} L(\alpha)$$

- Dependent data (time series): chain rule

$$L(\alpha) = f(x_1|\alpha)f(x_2|\alpha, x_1)f(x_3|\alpha, x_2, x_1)...$$

- Negative log-likelihood $l(\alpha) = -\sum_i \log(f(x_i|\alpha, x_{i-1}, \cdots))$
- Maximum likelihood: Optimal α

$$\max_{\alpha} L(\alpha) = \min_{\alpha} l(\alpha)$$

Maximum likelihood estimation: reminder

- Normal distributions: if $x_i \sim N(\mu, \sigma^2)$, iid.

$$L(\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}}$$

- Maximum likelihood

$$\widehat{\mu} = \bar{x}$$
$$\widehat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$$

Whiteboard

- For ARMA models, assume normality of w_t!
- Negative log-likelihood

$$l(\mu, \phi, \sigma_w^2) = \frac{S(\mu, \phi)}{2\sigma_w^2} + \frac{n}{2}\log(2\pi\sigma_w^2) - \frac{1}{2}\log(1-\phi^2)$$

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

- How to find optimum?
  - For σ^2 explicit

$$\widehat{\sigma}_w^2 = \frac{1}{n}S(\widehat{\mu}, \widehat{\phi})$$

  - Otherwise numerical optimization (unconstrained optimization)

Optimization methods

- Examples:
  - Steepest descent
  - Newtons Methods
  - Gauss-Newton methods
  - (least squares)
  - ...

Least squares

- Unconditional least squares
  - Estimate by numerical methods or sometimes analytically

$$\min_{\mu,\phi} S(\mu, \phi)$$

- Conditional least squares: assume x_1 given (constant)

$$min \sum_{i=1}^{t} w_i^2$$

- For AR(1), $\sum_{i=1}^{t} w_i^2 = S_c(\mu, \phi)$

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

- Note: Minimize by doing regression $Y = x_t, X = \text{lag}(x_t)$

Home reading

- Shumway and Stoffer, parts of sections 3.5, 3.6, 3.7
- R code: arima.sim, arima, polyroot, ARMAtoMA, ARMAacf

Time Series Analysis

Lecture 5: ARIMA models-2
Estimation, PACF, Forecasting

Tohid Ardeshiri

Linköping University
Division of Statistics and Machine Learning

September 16, 2019

Maximum likelihood estimation: reminder

Time series data are NOT independent

- Likelihood of α given observations $x_1, \cdots, x_t$ is

$$L(\alpha) = f(x_1, \cdots, x_t|\alpha)$$

- Maximum likelihood: Optimal α

$$\widehat{\alpha} = \arg\max_{\alpha} L(\alpha)$$

- Dependent data (time series): chain rule

$$L(\alpha) = f(x_1|\alpha)f(x_2|\alpha, x_1)f(x_3|\alpha, x_2, x_1)...$$

- Negative log-likelihood $l(\alpha) = -\sum_i \log(f(x_i|\alpha, x_{i-1}, \cdots))$
- Maximum likelihood: Optimal α

$$\max_{\alpha} L(\alpha) = \min_{\alpha} l(\alpha)$$

Maximum likelihood estimation: reminder

- Normal distributions: if $x_i \sim N(\mu, \sigma^2)$, iid.

$$L(\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}}$$

- Maximum likelihood

$$\widehat{\mu} = \bar{x}$$
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

Whiteboard

- For ARMA models, assume normality of w_t!
- Negative log-likelihood

$$l(\mu, \phi, \sigma_w^2) = \frac{S(\mu, \phi)}{2\sigma_w^2} + \frac{n}{2}\log(2\pi\sigma_w^2) - \frac{1}{2}\log(1-\phi^2)$$

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

- How to find optimum?
  - For σ^2 explicit

$$\widehat{\sigma}_w^2 = \frac{1}{n}S(\widehat{\mu}, \widehat{\phi})$$

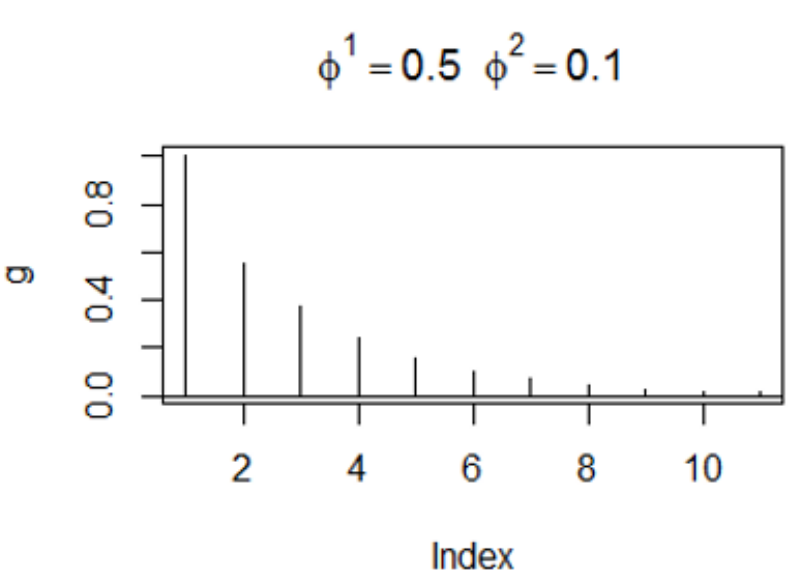
  - Otherwise numerical optimization (unconstrained optimization)

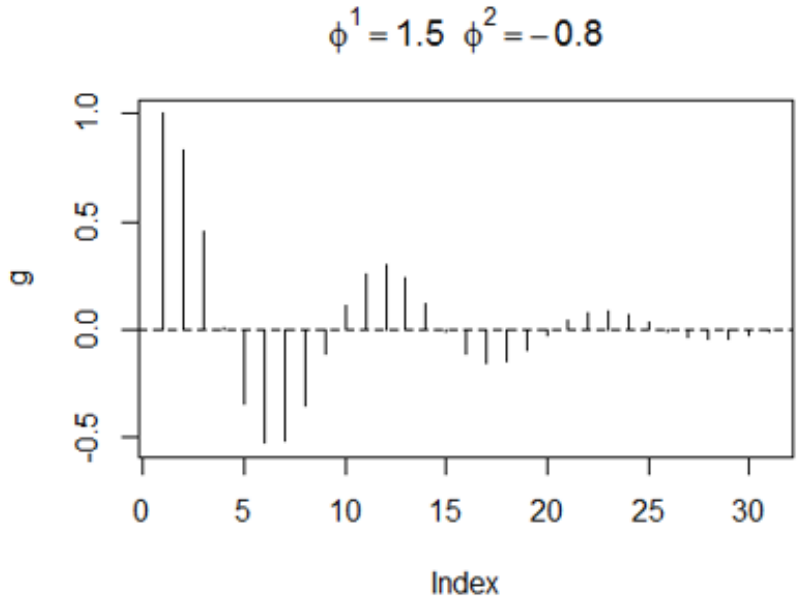
ARMA

- **Autoregressive moving average ARMA(p, q)**
  $x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}$
  - $\phi_p \neq 0, \theta_q \neq 0$
  - Is stationary
  - $Ex_t = 0$
- ACF for AR(1), MA(1), MA(2)

  $\rightarrow$ how to compute ACF for a general ARMA?

ACF for AR(2)


$\phi^1 = 0.5$ $\phi^2 = 0.1$
g
0.0
0.4
0.8
2
4
6
8
10
Index


$\phi^1 = 1.5$ $\phi^2 = -0.8$
g
-0.5
0.0
0.5
1.0
0
5
10
15
20
25
30
Index

ACF for AR(p), MA(p)

- AR(p): using difference equations
- MA(q): using difference equations

$$\rho(h) = \begin{cases} \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1+\theta^2+\ldots+\theta_q^2} & 0 \leq h \leq q \\ 0 & h > q \end{cases}$$

ACF for ARMA(p,q)

- ARMA(p,q):

$$\phi(B)x_t = \theta(B)w_t$$

- Causal ARMA: $x_t = \phi^{-1}(B)\theta(B)w_t = \psi(B)w_t = \sum_{j=0}^{\infty}\psi_j w_{t-j}$
  - How to find ψ_j in practice? Coefficient matching
- Theorem: ACF of ARMA(p,q) can be found by solving general homogeneous equations:

$$\gamma(h) - \phi_1\gamma(h-1) - ... - \phi_p\gamma(h-p) = 0, \qquad h \geq max(p, q+1)$$

  - Initial conditions

$$\gamma(h) - \phi_1\gamma(h-1) - ... - \phi_p\gamma(h-p) = \sigma_w^2 \sum_{j=h}^{q} \theta_j\psi_{j-h}, 0 \leq h < \max(p, q+1)$$

ACF for ARMA(1,1)

- Show for ARMA(1,1)

$$\rho(h) = \frac{(1+\theta\phi)(\phi+\theta)}{1+2\theta\phi+\theta^2}\phi^{h-1}, h \geq 1$$

- Note: pattern similar to AR(1) $\rightarrow$ hard to differentiate
- Note: ACF is 0 for $h > q$ from MA(q) $\rightarrow$ MA(q) is identifiable from ACF
- How to differentiate between AR(p)? ARMA(p)?

Partial correlation

A Gaussian intuition:

- Conditional density: $f(x, y|z) = \frac{f(x,y,z)}{f(z)}$
- if x, y and z were jointly normal then

$$f(x, y|z) = N\left(\begin{bmatrix} x \\ y \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

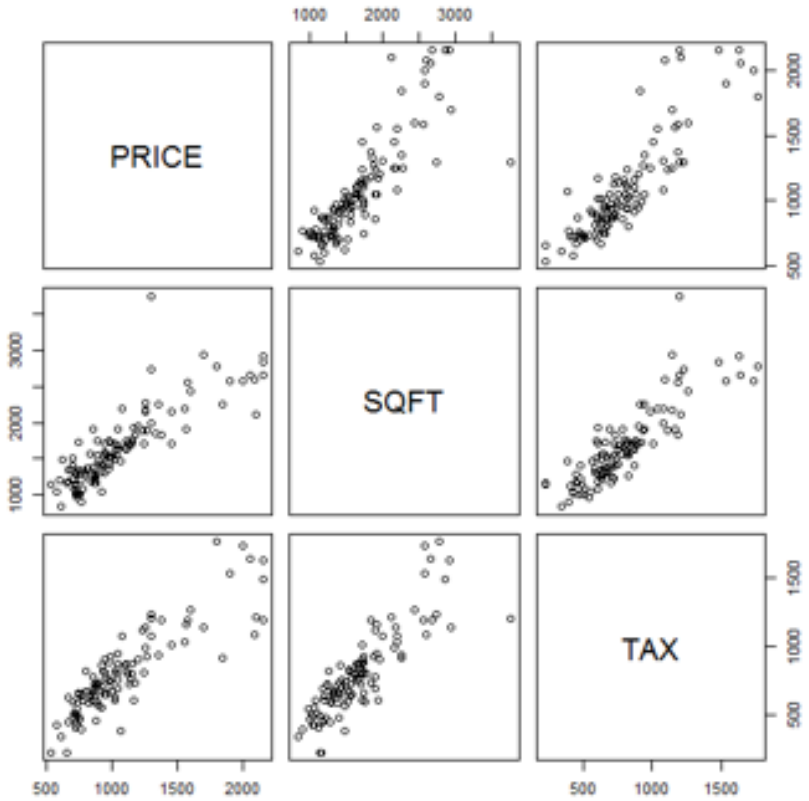
- Also,

$$\rho_{xy|z} = \frac{\text{cov}(x, y|z)}{\sqrt{\text{var}(x|z)\,\text{var}(y|z)}} = \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}$$

- **What if $\Sigma_{12} = 0$?**

Partial autocorrelation

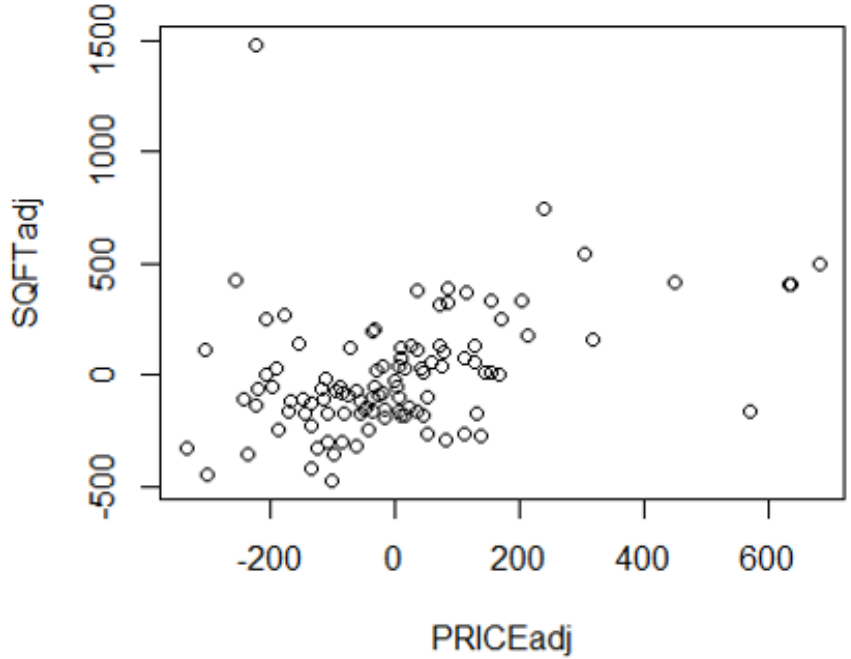
- Example: Albuquerque home prices
  - What if we remove information stored in TAX from PRICE and SQFT?



Partial autocorrelation

- $\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 z$
- $\hat{x} = \hat{\beta}_0 + \hat{\beta}_1 z$
- $x' = x - \hat{x}$
- $y' = y - \hat{y}$

- **Partial autocorrelation**



  - If we know α, β and z, we can reduce connection between x and y

```
> corr(cbind(PRICEadj,SQFTadj))
[1] 0.3675204
```

PACF

- Partial autocorrelation function (PACF) for a stationary process

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

$$\phi_{hh} = \text{corr}(x'_{t+h}, x''_t), \quad h > 1$$

  - where $x'_{t+h} = x_{t+h} - \sum_{j=1}^{h-1} \hat{\beta}_j x_{t+h-j}$
  - and $x''_t = x_t - \sum_{j=1}^{h-1} \hat{\beta}_j x_{t+j}$
  - Note: coefficients in x''_{t+h} and x'_{t+h} are the same (stationarity)
- Example: AR(1) $\phi_{11} = \phi, \phi_{22} = 0$

PACF for AR(p)

$$x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + w_t$$

- It can be shown:
  - $\phi_{pp} = \phi_p$
  - $\hat{\beta}_1 = \phi_1, ...\hat{\beta}_p = \phi_p, \hat{\beta}_{p+1} = 0, ..., \hat{\beta}_h = 0$ for $h > p$

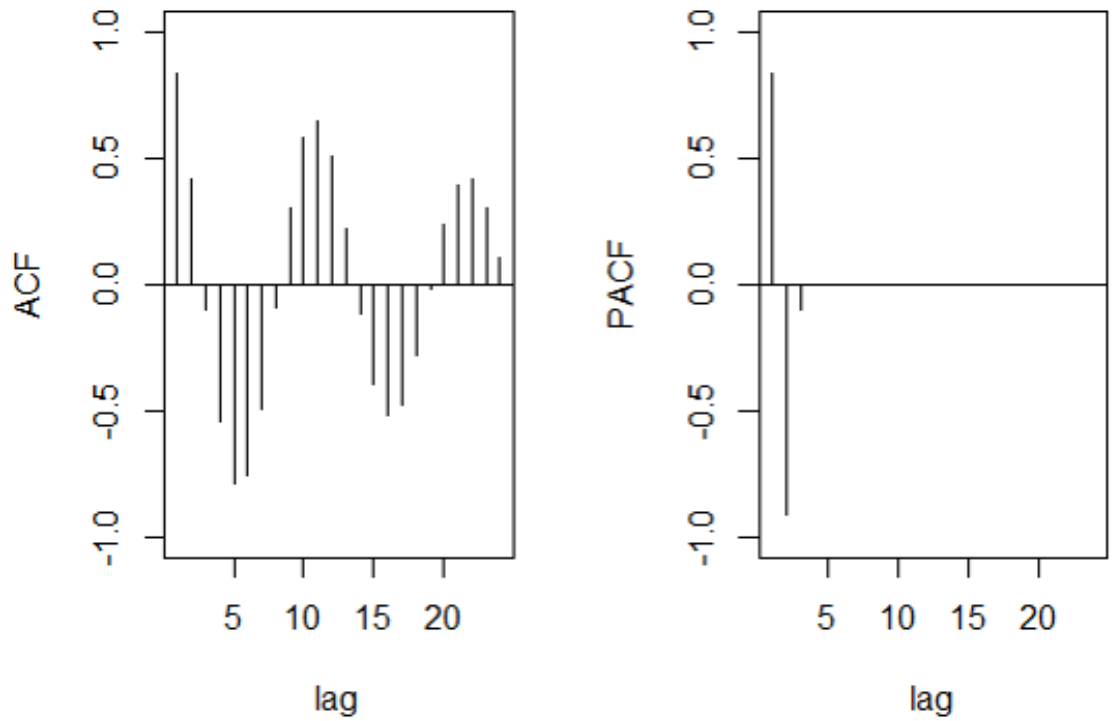
- It means

$$\phi_{hh} = \text{cov}(x_{t+h} - \sum_{j=1}^{p} \phi_j x_{t+h-j}, x_- \sum_{j=1}^{p} \phi_j x_{t+j})$$

$$= \text{cov}(w_{t+h}, x_t - \sum_{j=1}^{p} \phi_j x_{t+j}) = 0, \quad \text{when h} > p$$

PACF for AR(p)

- Example: AR(3) $\phi_1 = 1.5$, $\phi_2 = -0.75, \phi_3 = -0.1$

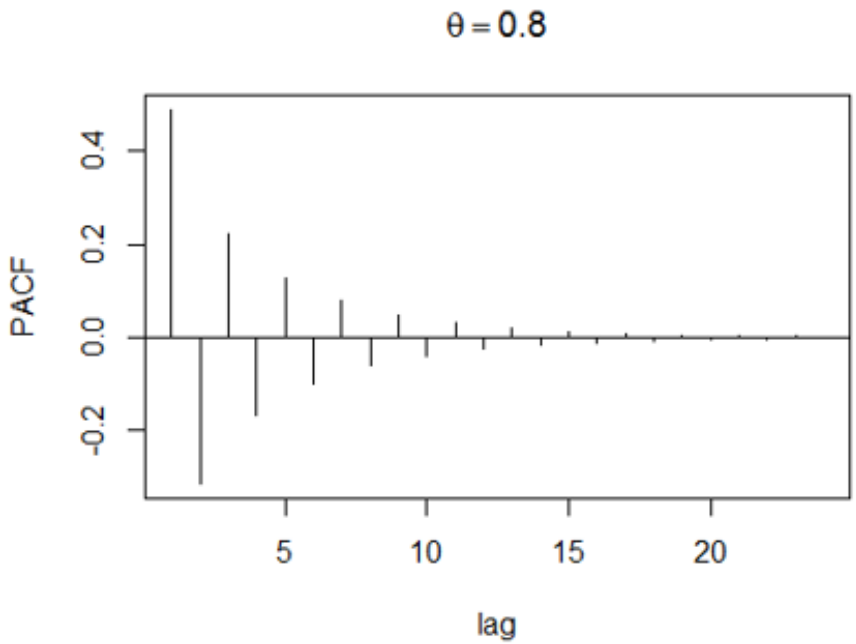


PACF for MA(1)

- If invertible,

$$\phi_{hh} = -\frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2h+2}}, \quad h \geq 1$$

Decreases exponentially with h



ACF and PACF

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

How to differentiate between ARMA(p,q)?

Empirical ACF (EACF)

Idea:

- ARMA(p,q): $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + \sum_{j=1}^{q} \theta_j w_{t-j} + w_t$
- If we can estimate $\phi_j \rightarrow x'_t = x_t - \sum_{j=1}^{p} \phi_j x_{t-j}$ is linear function in $w_t, ..., w_{t-q}$
- If we run regression x'_t against $w_t ... w_{t-j}$:
  - Residuals are white noise, $j \geq q \rightarrow$ ACFs not significant
    - ⋆ Some of the coefficients will be 0
  - Residuals are not white noise, $j < q \rightarrow$ ACFs significant
  - Note: $w_t s$ substituted by lagged residuals from a series of regressions
- If $x'_t = x_t - \sum_{j=1}^{k} \phi_j x_{t-j}, k < p \rightarrow$ white noise will never be achieved $\rightarrow$ ACFs are not zero

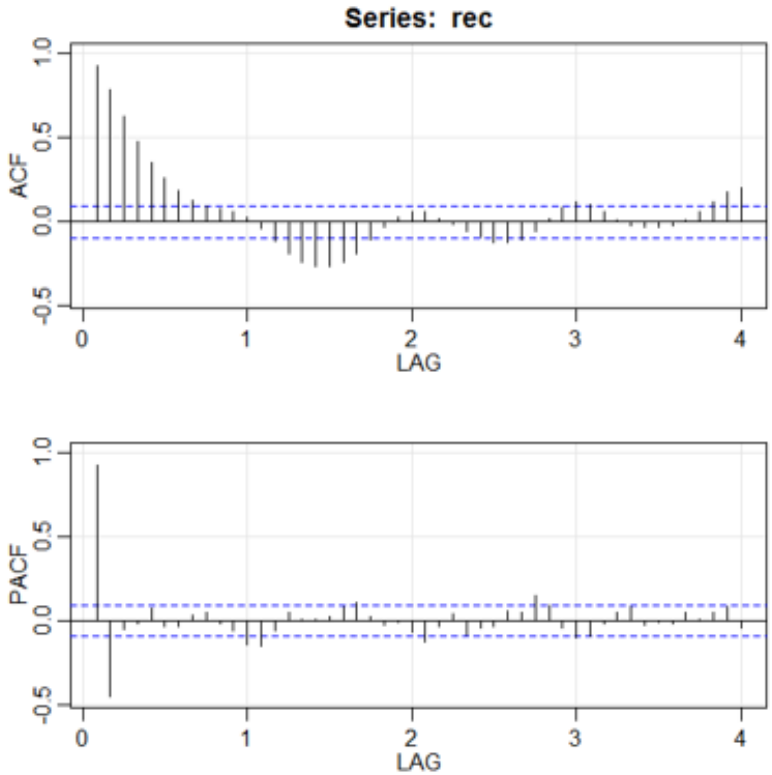
Empirical ACF (EACF)

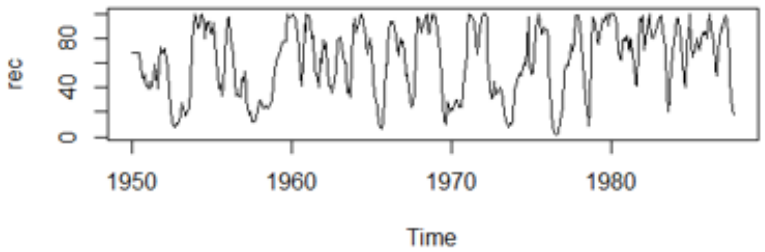
- $k > p$ General result: ACFs are 0 for $j > q + (k - p)$
  - Example: ARMA(0,1)
- General conclusion for AR,MA =(k,j):
  - This is theoretical one! $\rightarrow$ not exactly the same for the samples

AR/MA	0	1	2	...	...	...	...
0	X	X	X	X	X	X	X
1	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X
...	X	X	X	X	X	X	X
...	X	X	X	X	X	X	X
...	X	X	0	0	0	0	0
...	X	X	X	0	0	0	0
...	X	X	X	X	0	0	0
...	X	X	X	X	X	0	0

ARMA orders

- Recruitment series





Conclusion?

ARMA orders

- EACF

```
> TSA::eacf(rec)
AR/MA
     0 1 2 3 4 5 6 7 8 9 10 11 12 13
   0 x x x x x x x x o o o o o x
   1 x x x o o o o o o o o o o o
   2 o o x x o o o o o o o o o o
   3 x o o x o o o o o o o o o o
   4 x x o o o o o o o o o o o o
   5 x x x o o o o o o o o o o o
   6 x x x o o o o o o o o o o o
   7 x x o o o o o o o o o x o o
```

ARMA orders

- AR(2) and ARMA(1,3)
  - Conclusions?

```
> arima(rec, order=c(2,0,0))

Call:
arima(x = rec, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      1.3512  -0.4612    61.8585
s.e.  0.0416   0.0417     4.0039

sigma^2 estimated as 89.33:  log likelihood = -1661.51,  aic = 3331.02
> arima(rec, order=c(1,0,3))

Call:
arima(x = rec, order = c(1, 0, 3))

Coefficients:
         ar1     ma1     ma2     ma3  intercept
      0.7826  0.5484  0.3239  0.2119    61.8609
s.e.  0.0390  0.0554  0.0621  0.0530     4.1953

sigma^2 estimated as 88.43:  log likelihood = -1659.24,  aic = 3330.48
>
```

Model selection

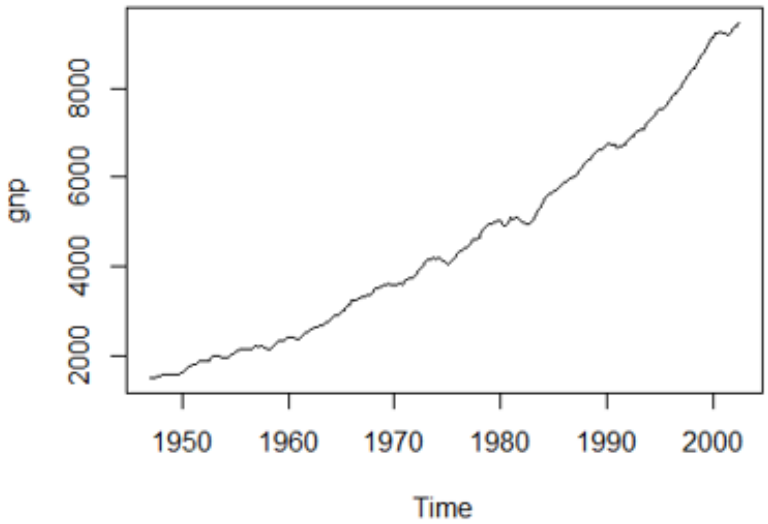
- Which model is suitable?
  - What is p, d, q is ARIMA(p,d,q)?
  - d is defined before!
  -
- Step 1: Check ACF, PACF and EACF to define a few tentative models

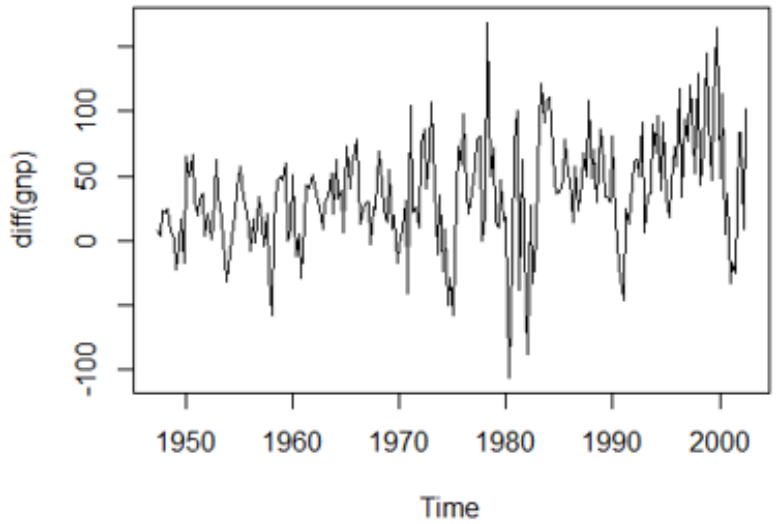
Model selection

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

Model selection

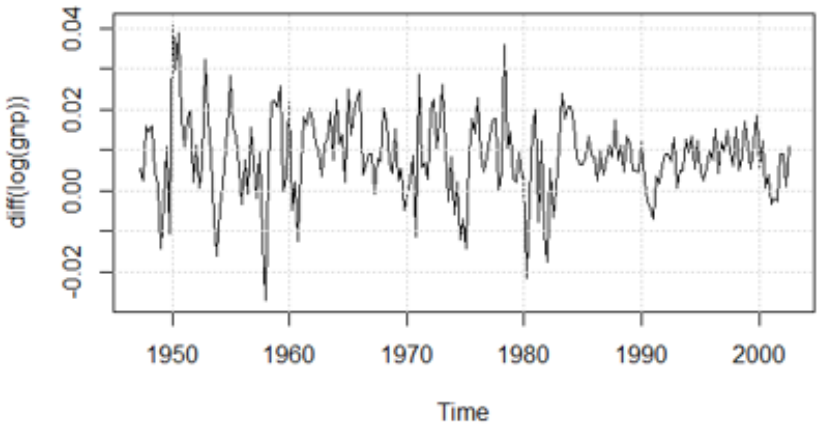
- Example: GNP data
  - Trying differencing $\rightarrow$ non-constant variance and maybe trend? $\rightarrow$ transformation





Model selection

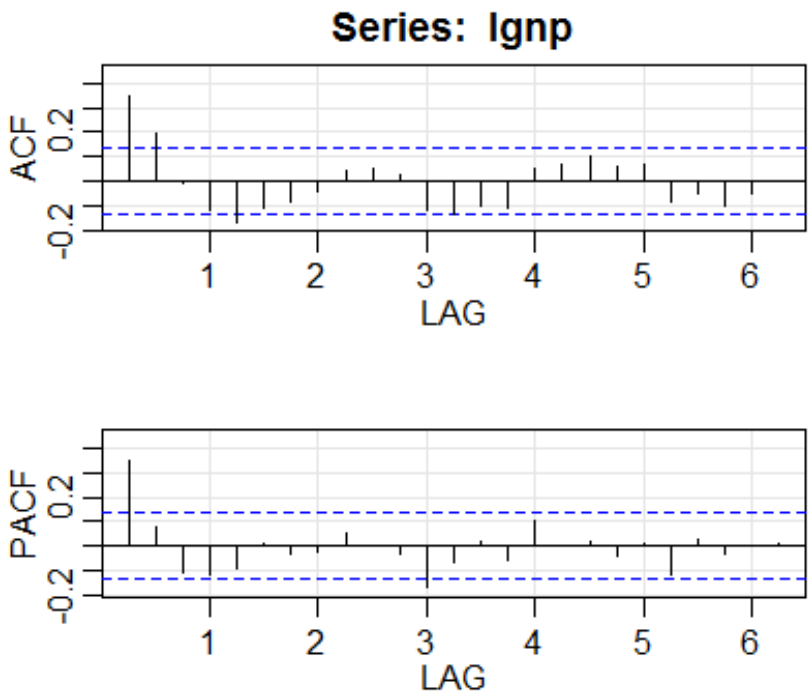
- Example: GNP data
  - Taking log and then differncing $\rightarrow$ still not perfect, but ... keep it as is.




```
> adf.test(lgnp)

        Augmented Dickey-Fuller Test

data:  lgnp
Dickey-Fuller = -6.1756, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Model selection

- Example: GNP data
  - Testing ACF and PACF



Conclusion?

Model selection

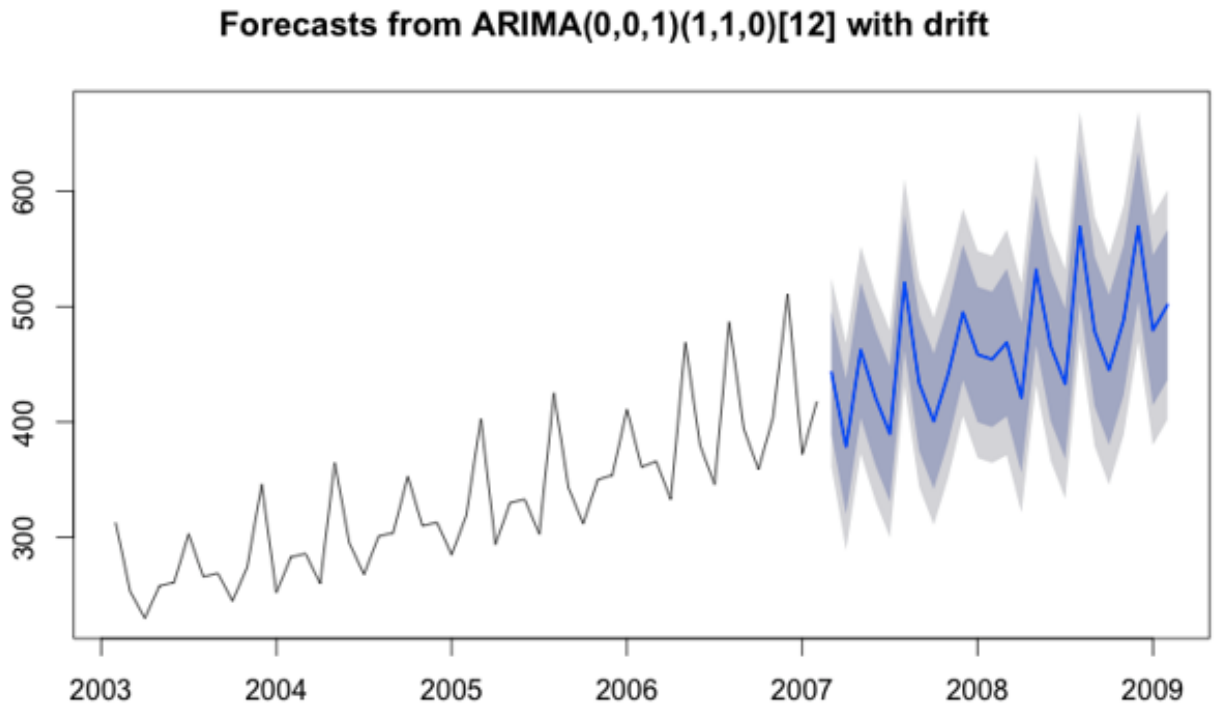
- Example: GMP data
  - Checking EACF

```
> TSA::eacf(lgnp)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o o x o o o o o o  o  o  o
1 x x o o o o o o o o o  o  o  o
2 x x o o o o o o o o o  o  o  o
3 x o x o o o o o o o o  o  o  o
4 x o x o o o o o o o o  o  o  o
5 o x x o x o o o o o o  o  o  o
6 x x x x x o o o o o o  o  o  o
7 x x o x o o x o o o o  o  o  o
```

Conclusion?

Forecasting

- We have our series $x_1 ... x_n$
- Use series to predict m steps ahead: x^n_{n+m} should be based on our observed data $x^n_{n+m} = g(x_1, ..., x_n)$



Forecasting

- Assume $g(x_1, ..., x_n) = \alpha_0 + \sum_{k=1}^{n} \alpha_k x_k$
  - Best linear predictors
- How to find α's?

$$minE\left[(x_{n+m} - g(x_1, ...x_n))^2\right]$$

- Prediction equations
  - Find α's by solving ($x_0 = 1$)

$$E\left[(x_{n+m} - x_{n+m}^n)x_k\right] = 0, k = 0, ..., n$$

- Note: n+1 equations, n+1 unknowns

One-step-ahead

- Denote $x_{n+1}^n = \phi_{n1}x_n + ...\phi_{nn}x_1$
- Prediction equations give

$$\Gamma_n \phi_n = \gamma_n$$

$$\Gamma_n = \begin{pmatrix} \gamma(1-1) & \gamma(2-1) & \dots & \gamma(n-1) \\ \gamma(2-1) & \gamma(2-2) & \dots & \gamma(n-2) \\ \dots & \dots & \dots & \dots \\ \gamma(n-1) & \gamma(n-2) & \dots & \gamma(n-n) \end{pmatrix}$$

$$\phi_n = \begin{pmatrix} \phi_{n1} \\ \dots \\ \phi_{nn} \end{pmatrix} \qquad \gamma_n = \begin{pmatrix} \gamma_1 \\ \dots \\ \gamma_n \end{pmatrix}$$

- Note: for ARMA models Γ_n is positive def $\rightarrow$ unique solution

One-step-ahead

- Causal AR(p): for $n \geq p$ best linear prediction is

$$x_{n+1}^{n} = \phi_1 x_n + ... + \phi_p x_{n-p+1}$$

- In general, solve system of equations $\rightarrow O(n^3)$ operations

- Much faster algorithms exist
  - Durbin-Levinson algorithm
  - Innovations algorithm
- Property: PACF of a stationary process can be obtained as ϕ_{nn} by solving $\Gamma_n \phi_n = \gamma_n$

One-step-ahead

- Mean square prediction error (MSPE)

$$P_{n+1}^{n} = E\left[(x_{n+1} - x_{n+1}^{n})^2\right] = \gamma(0) - \gamma_n' \Gamma_n^{-1} \gamma_n$$

- Confidence intervals for x_{n+1}

$$x_{n+1}^{n} \pm \alpha \sqrt{P_{n+1}^{n}}$$

- m-step ahead in general? Prediction equations
  - Difficult in general

Read home

- Ch 3.2-3.4
- Paper "Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation" by Tsay and Tiao

- R code: eacf in TSA package

m-step-ahead for ARMA

- Assume causal and invertible ARMA(p,q)
- Finite past prediction

$$x_{n+1}^n = E(x_{n+1}|x_n, ...x_1)$$

- Infinite past prediction

$$\tilde{x}_{n+m}^n = E(x_{n+m}|x_n, ...x_1, x_0, x_{-1}, ...)$$

- m-step-ahead forecast for infinite past
  - Compute recursively

$$\tilde{x}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \hat{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j \tilde{x}_{n+m-j}, m = 1, 2, ...$$

- m-step ahead prediction error: $P_{n+m}^n = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2$

Long-range forecasts

- What if $m \to \infty$?

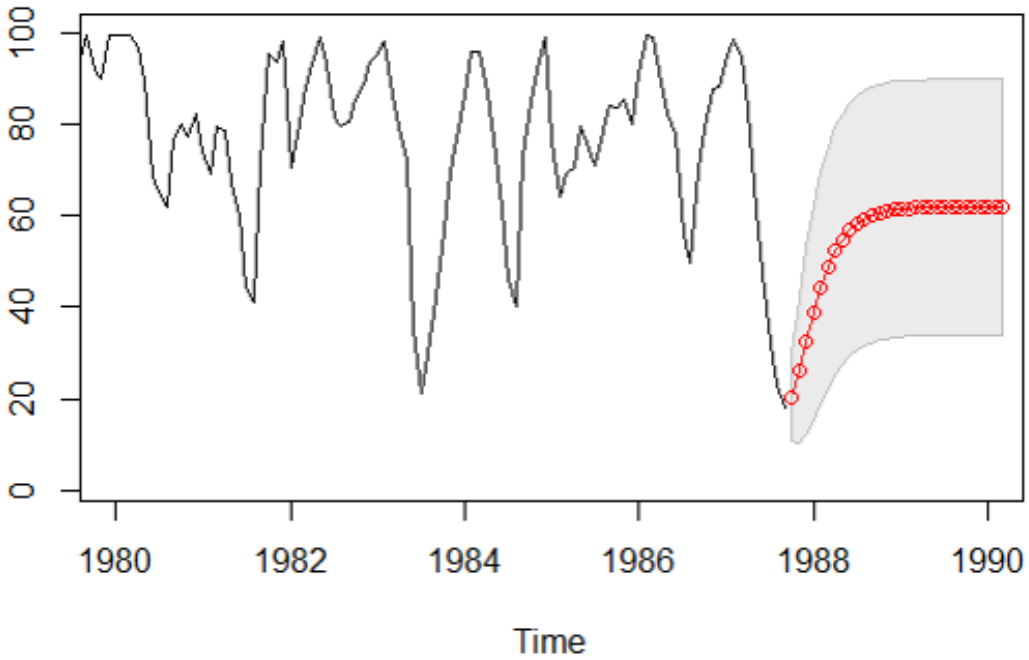
$$\tilde{x}_{n+m} \to 0 (\text{or} \mu)$$

$$P^n_{n+m} \to \sigma^2_x$$

m-step-ahead

- Recruitment, AR(2)

$$x_{n+m}^{n} \pm 2\sqrt{P_{n+m}^{n}}$$



Truncated prediction

- Ignore non-positive j in x_j

$$\tilde{x}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j}, m = 1, 2, ...$$

- For ARMA, truncated prediction formula:
  - Recursive computation, explicit

$$\tilde{x}_{n+m}^n = \phi_1 \tilde{x}_{n+m-1}^n + ...\phi_p \tilde{x}_{n+m-p}^n + \theta_1 \tilde{w}_{n+m-1}^n + ... + \theta_q \tilde{w}_{n+m-q}^n$$
$$\hat{w}_t^n = \tilde{x}_t^n - \phi_1 \tilde{x}_{t-1}^n - ... - \phi_p \tilde{x}_{t-p}^n - \theta_1 \tilde{w}_{t-1}^n - ... - \theta_q \tilde{w}_{t-q}^n$$

- Boundary conditions: $\tilde{x}_t^n = x_n, 1 \leq t \leq n, \tilde{x}_t^n = 0, t \leq 0$

$$\tilde{w}_t^n = 0, t \leq 0 \quad \text{or } t > n$$

Time Series Analysis

Lecture 6: ARIMA models summary
State space models

Tohid Ardeshiri

Linköping University
Division of Statistics and Machine Learning

September 27, 2019

Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

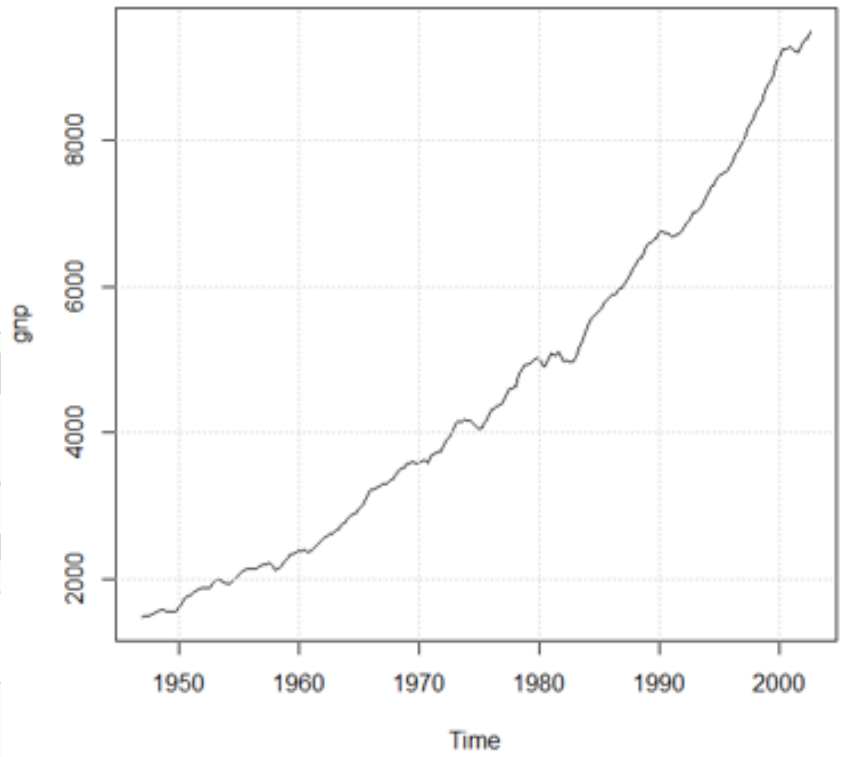
$\theta = (\theta_1, ..., \theta_q)$

Mod

- A
  ar

- Estimate $\phi's$, $\theta's$, ...

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

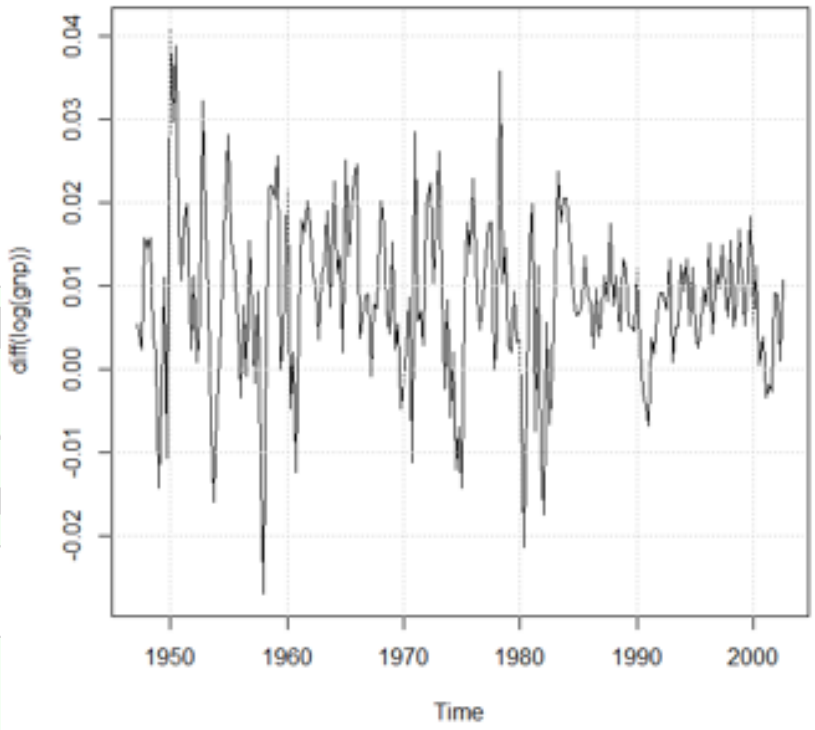
Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Prediction

$Y_t = \nabla(\log(X_t))$



Estimate $\phi's$, $\theta's$, ...

Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

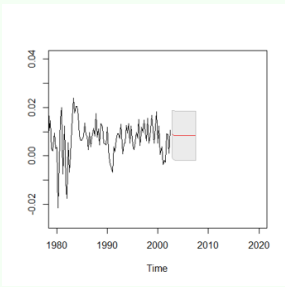
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

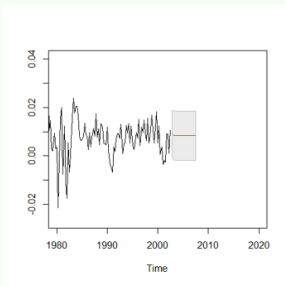
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

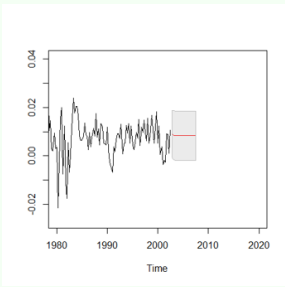
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

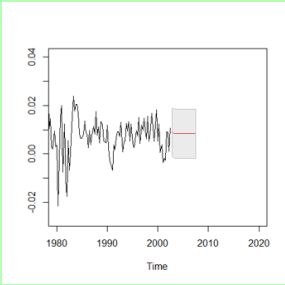
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

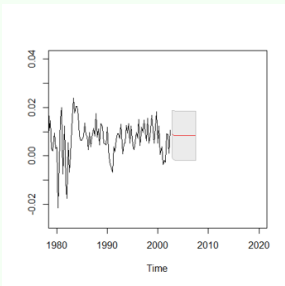
- Are there better models with other p and q or even different kind?

Revise the model

yes

no

Prediction



Model selection

Fit the tentative models, compare them

- Analytical measures: AIC, BIC
  - Penalize models with many parameters $\rightarrow$ simpler models
- Residual analysis

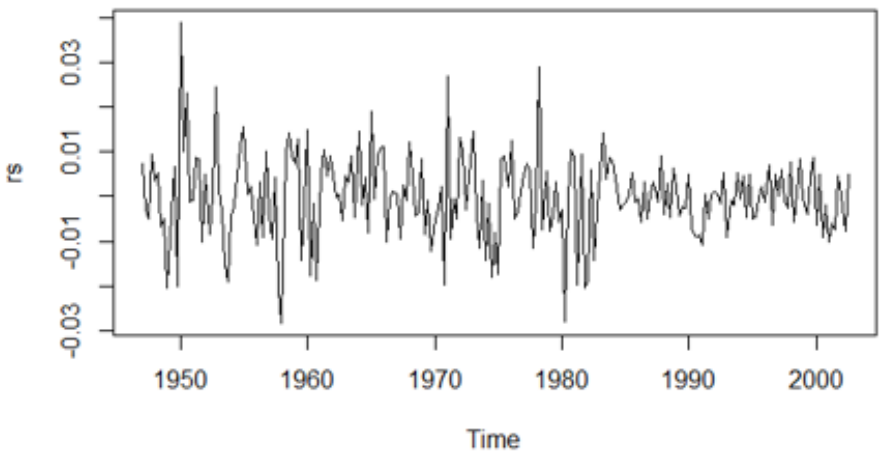
Residual analysis

- Residuals $r_t = x_t - \hat{x}_t^{t-1}$?they are innovations
  - Note: computed from one-step-ahead predictions!
  - Measures predictive quality of the model (compare OLS)

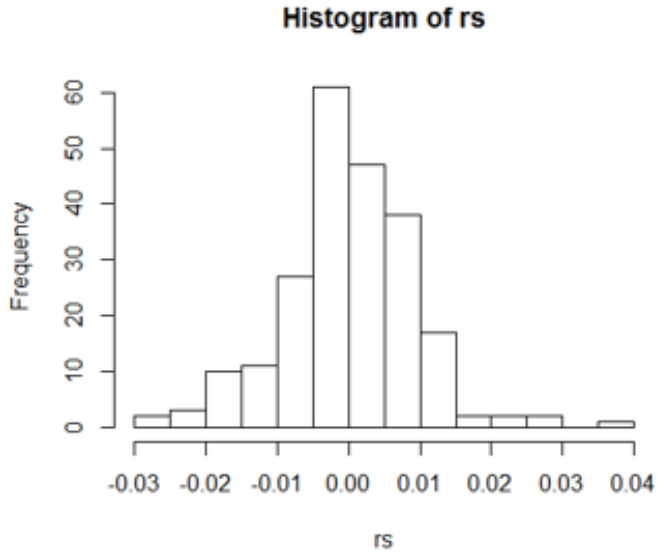
- Residual analysis
  - Visual inspection: stationary? Patterns?
  - Histograms, Q-Q plots
  - ACF, PACF
  - Runs test
  - Box-Ljung test

Residual analysis - Visual inspection

Histogram and visual inspection

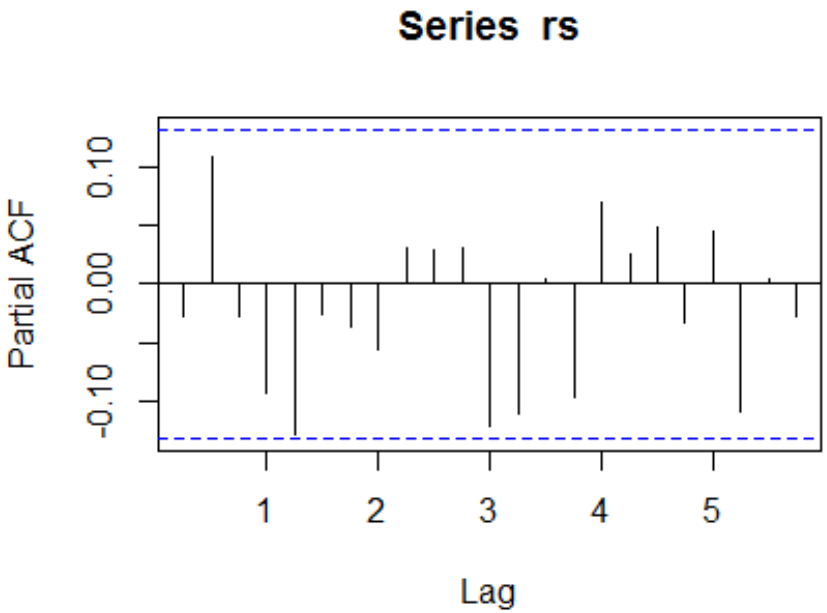


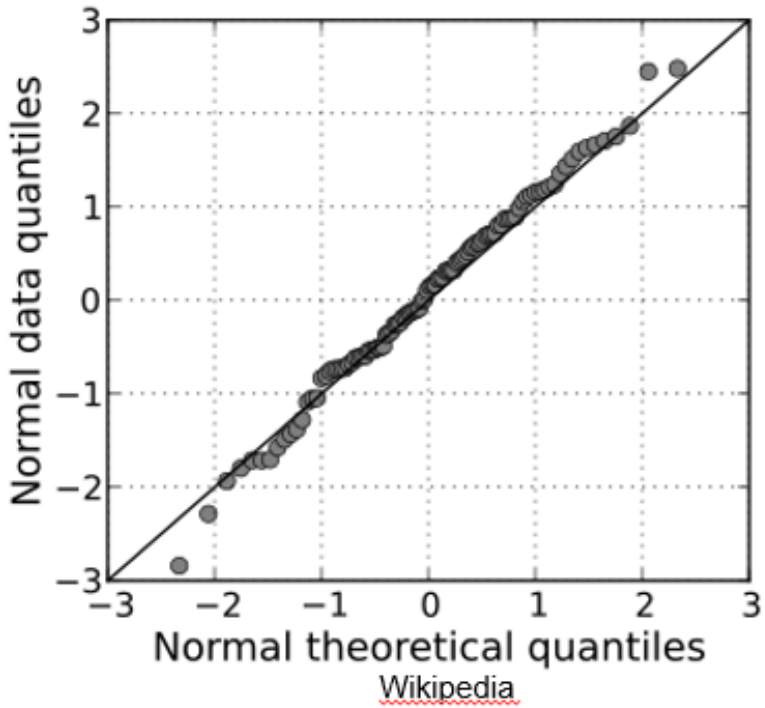
If looks white is good



If looks Normal is good

Residual analysis - ACF /PACF Q-Q plots





If between the blue lines good

If along the diagonal line GOOD

Statistical tests

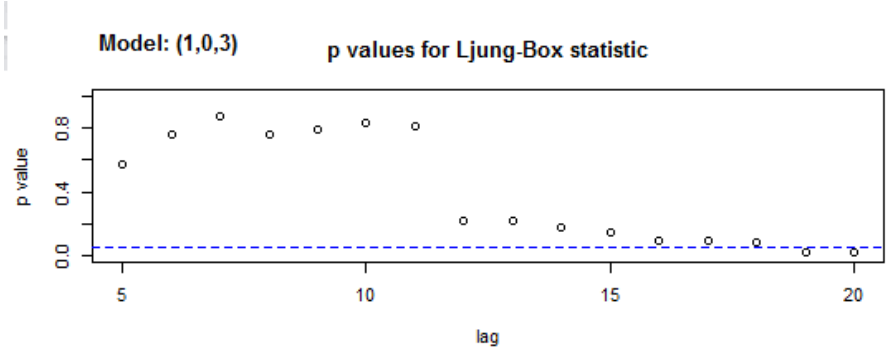
Tests are used to test independence

Runs test

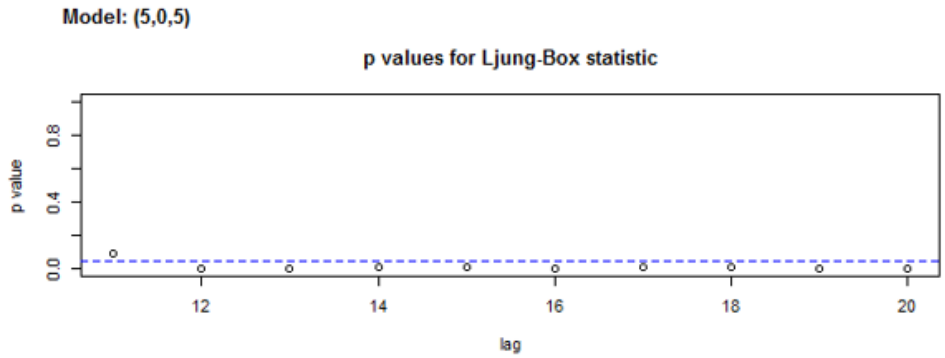
- H_0 : x_t values are i.i.d. **p-value NOT small**
- H_a : x_t values are not i.i.d. **p-value small**

Box-Ljung test

- H_0 : data are independent **p-value NOT small**
- H_a : data are not independent **p-value small**



GOOD



BAD

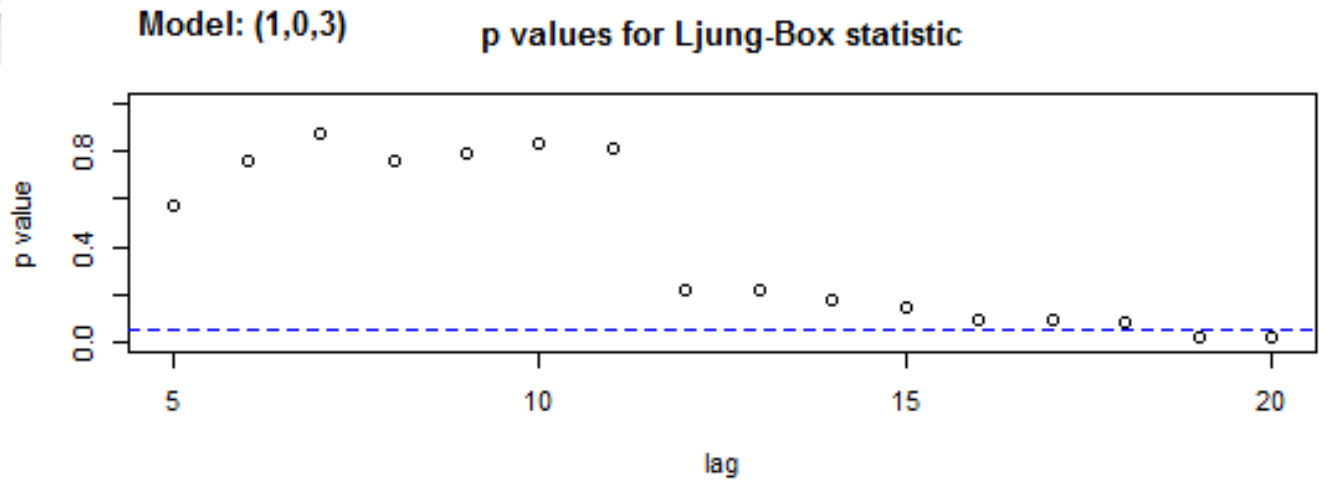
Overfitting

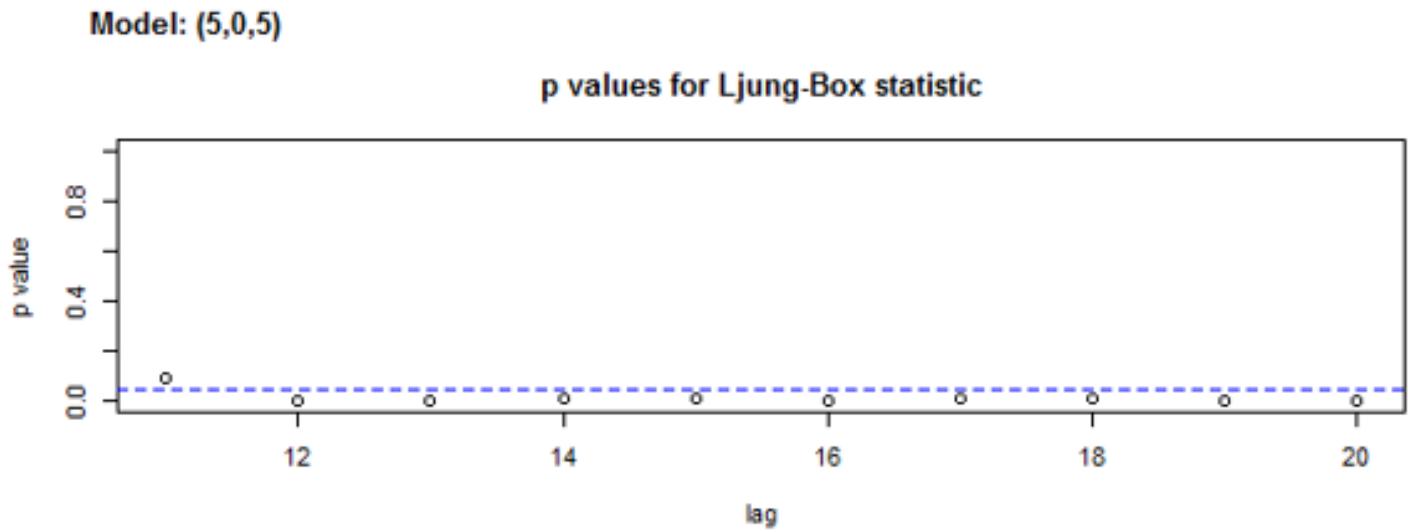
- Occams razor: among equally good models, choose the simplest one
- Overfitting: taking too complex models leads to bad predictions

- If ARIMA(p, d, q) has almost the same predictive quality as ARIMA(p', d', q') , take the one with less parameters

Overfitting

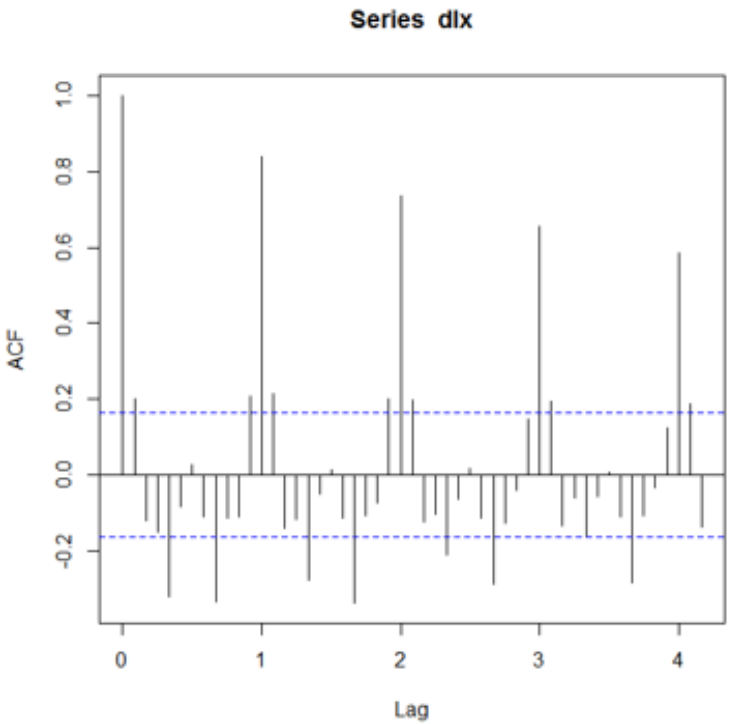
- Example: Recruitment series
  - Fit ARIMA(1,0,3) and ARIMA(5,0,5)

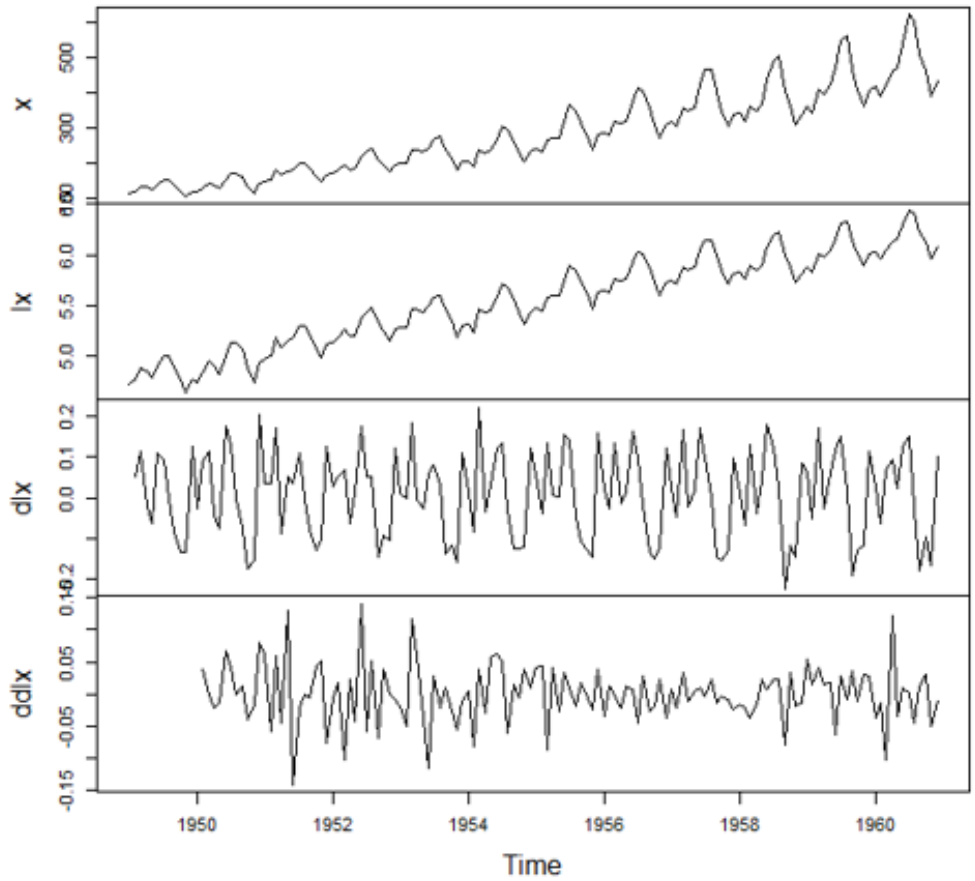




SARIMA - Air passangers

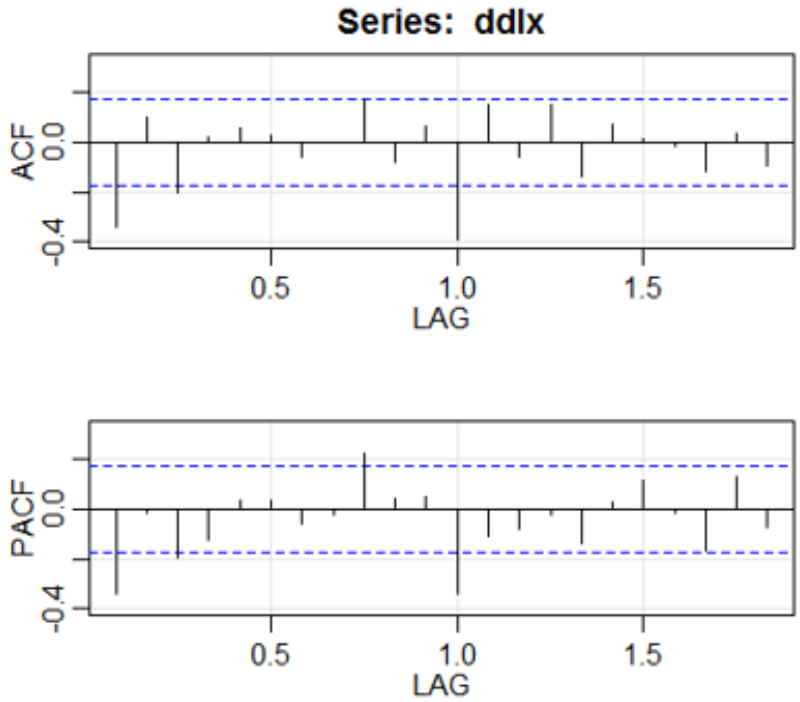
- Example: Air passangers





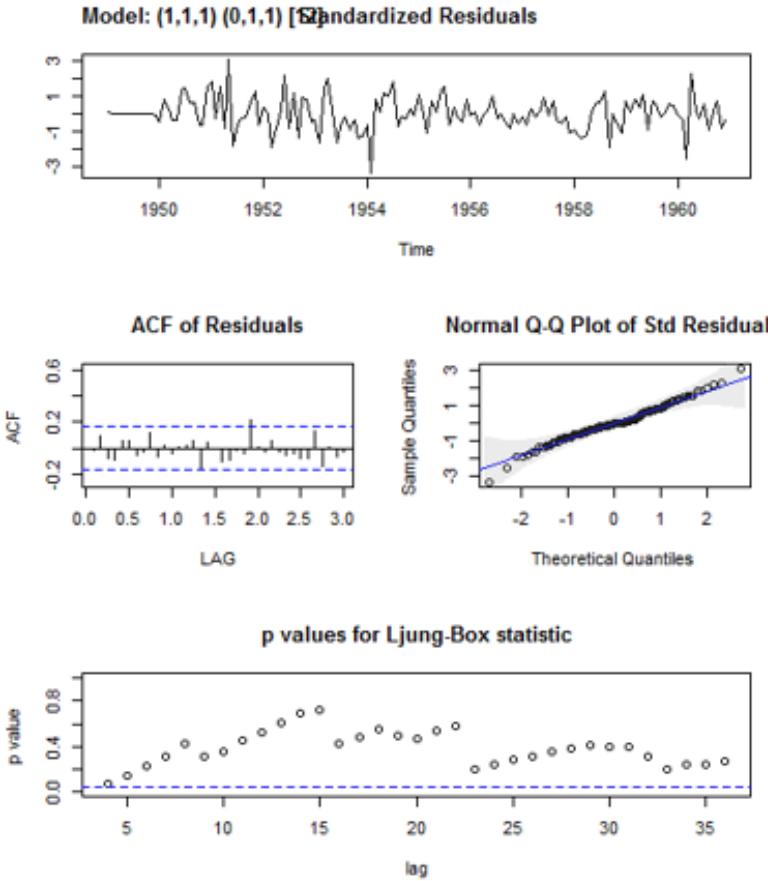
SARIMA - Air passangers

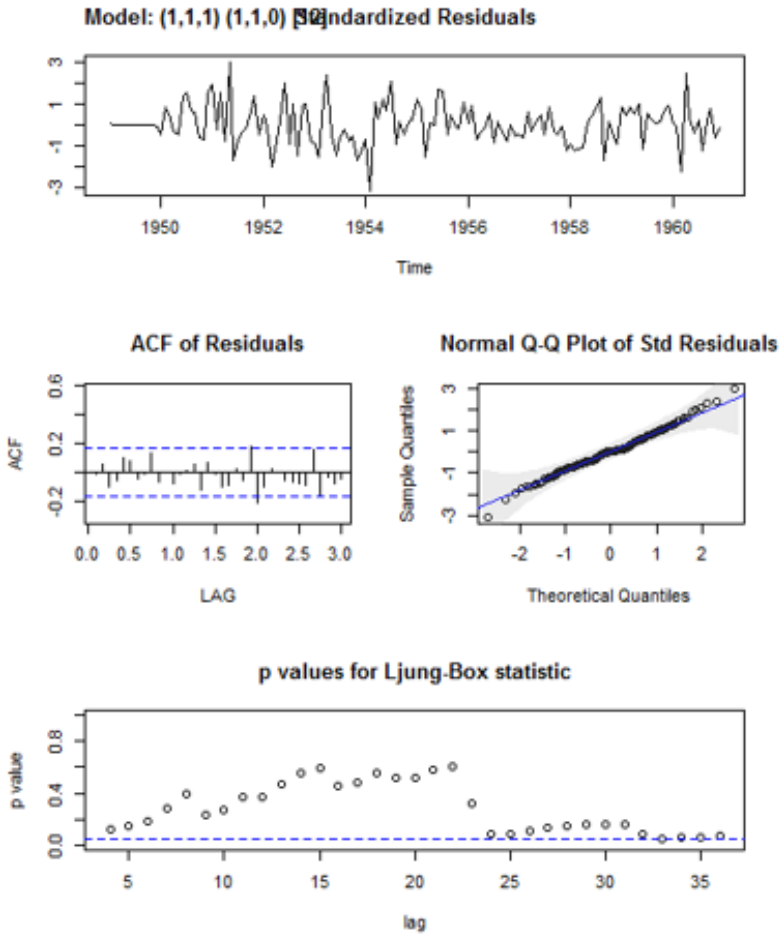
- Example: Air passangers



$\text{ARIMA}(0,1,1)_{12}$ or
$\text{ARIMA}(1,1,0)_{12}$

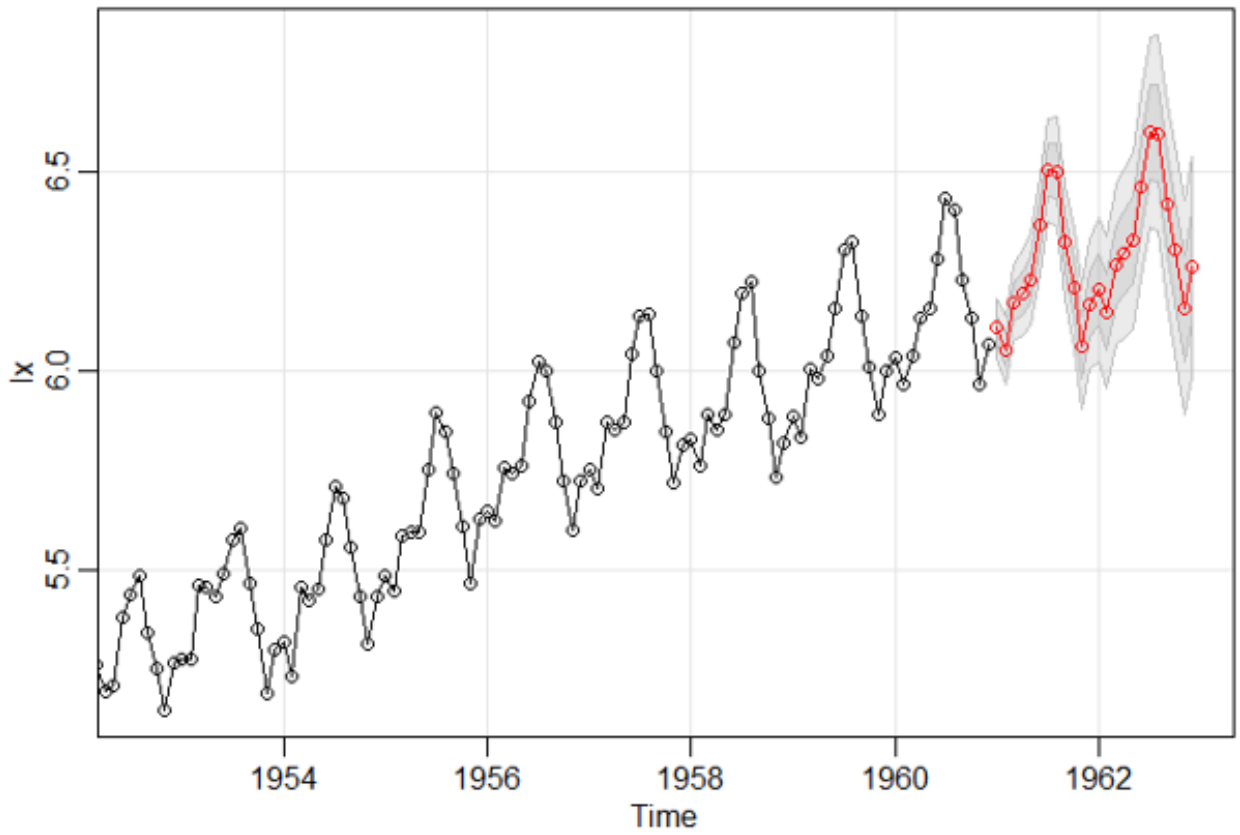
SARIMA - Air passangers





SARIMA

- Forecasting



Read home

- Shumway and Stoffer, Chapter 1, 2 and 3

ARIMA models

Time series models so far

$$\phi^p(B)x_t = \theta^q(B)w_t$$

Model	**Concise form**
AR(p)	$\phi^p(B)x_t = w_t$
MA(q)	$x_t = \theta^q(B)w_t$
ARMA(p, q)	$\phi^p(B)x_t = \theta^q(B)w_t$
ARIMA(p, d, q)	$\phi^p(B)(1-B)^d x_t = \theta^q(B)w_t$
ARMA(P, Q)$_s$	$\Phi^P(B^s)x_t = \Theta^Q(s)w_t$
ARIMA(P, D, Q)$_s$	$\Phi^P(B^s)(1-B^s)^D x_t = \Theta^Q(B^s)w_t$
ARMA(p, q) $\times$ (P, Q)$_s$	$\Phi^P(B^s)\phi^p(B)x_t = \Theta^Q(B^s)\theta^q(B)w_t$
ARIMA(p, d, q) $\times$ (P, D, Q)$_s$	$\Phi^P(B^s)\phi^p(B)(1-B^s)^D(1-B)^d x_t = \Theta^Q(B^s)\theta^q(B)w_t$

* The notation used in this slide deviates from the notation used in the course literature so far.

Vectorized AR(2) model

Whiteboard

Consider an AR(2) model

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

Let $\mathbf{z}_t = \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}$ and $e_t = \begin{bmatrix} w_t \\ 0 \end{bmatrix}$.

Show that we rewrite the AR(2) model in the state space form:

$$\mathbf{z}_t = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \mathbf{z}_{t-1} + e_t$$
$$x_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{z}_t,$$

ARIMA models in State Space form **Whiteboard**

$$\phi^p(B)x_t = \theta^q(B)w_t$$

Can we rewrite any model of this form as a state space model?

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + e_t,$$
$$\mathbf{x}_t = C\mathbf{z}_t + \nu_t,$$

ARIMA models in State Space form **Whiteboard**

$$\phi^p(B)x_t = \theta^q(B)w_t$$

Outline of the solution:

Let $r = \max(p, q+1)$,
$\phi^r(B) = 1 - \phi_1 B - \cdots - \phi_r B^r$,
$\theta^r(B) = 1 + \theta_1 B - \cdots - \theta_{r-1} B^{r-1}$,
$\phi^r(B)(\theta^r(B))^{-1} x_t = w_t$. Hence, for $z_t = (\theta^r(B))^{-1} x_t$ we can have
$\phi^r(B) z_t = w_t$

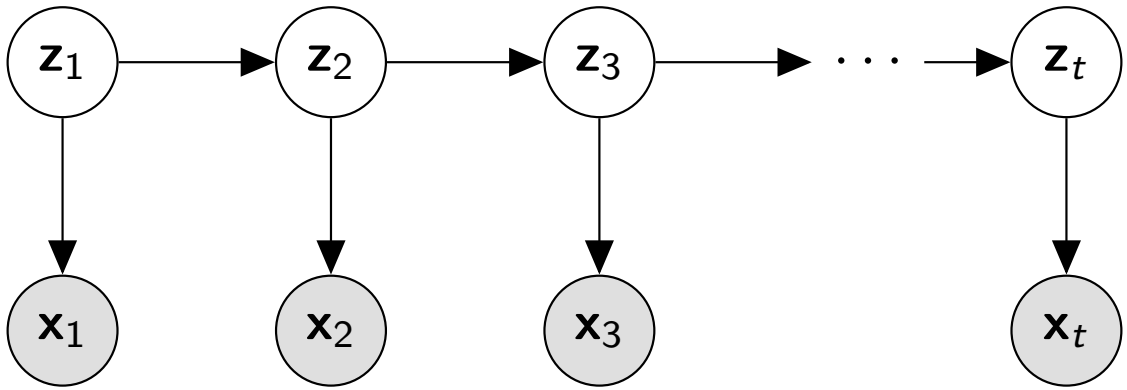
$$\mathbf{z}_t = \begin{bmatrix} z_t \\ z_{t-1} \\ z_{t-2} \\ \vdots \\ z_{t-r+1} \end{bmatrix} \text{ and } \mathbf{z}_t = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_r \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \mathbf{z}_{t-1} + \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$x_t = \begin{bmatrix} 1 & \theta_1 & \theta_2 & \cdots & \theta_r \end{bmatrix} \mathbf{z}_t$$

State Space models - graphical models

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + e_t, \qquad e_t \sim f_e(\cdot)$$
$$\mathbf{x}_t = C\mathbf{z}_t + \nu_t, \qquad \nu_t \sim f_\nu(\cdot)$$

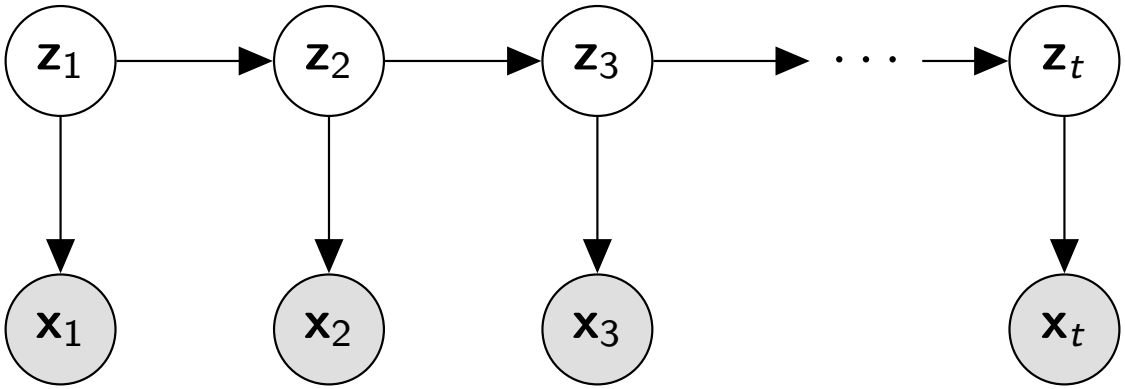
A probabilistic graphical model for stochastic dynamical system with latent state $\mathbf{z}_k$ and observations $\mathbf{x}_k$

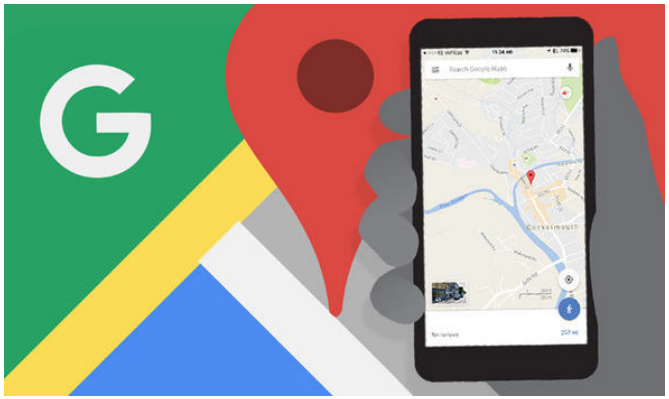


The main tool here is the probability Calculus; Bayes rule and marginalization.

Dynamical systems - more general case

$$\mathbf{z}_t = \mathcal{F}(\mathbf{z}_{t-1}) + e_t, \qquad e_t \sim f_e(\cdot)$$
$$\mathbf{x}_t = \mathcal{C}(\mathbf{z}_t) + \nu_t, \qquad \nu_t \sim f_\nu(\cdot)$$

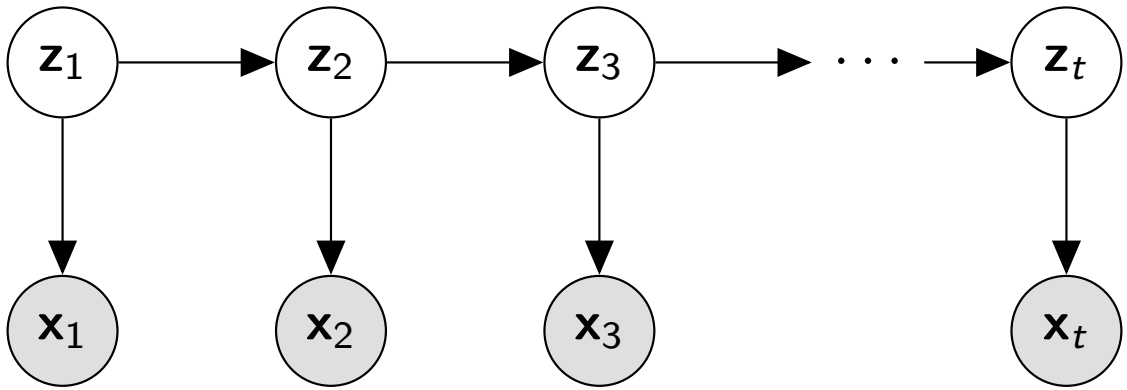




State Space models - Linear and Gaussian

Our main focus will be on linear and Gaussian models:

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + e_t, \qquad e_t \sim N(0, Q)$$
$$\mathbf{x}_t = C\mathbf{z}_t + \nu_t, \qquad \nu_t \sim N(0, R)$$



Bayesian Inference

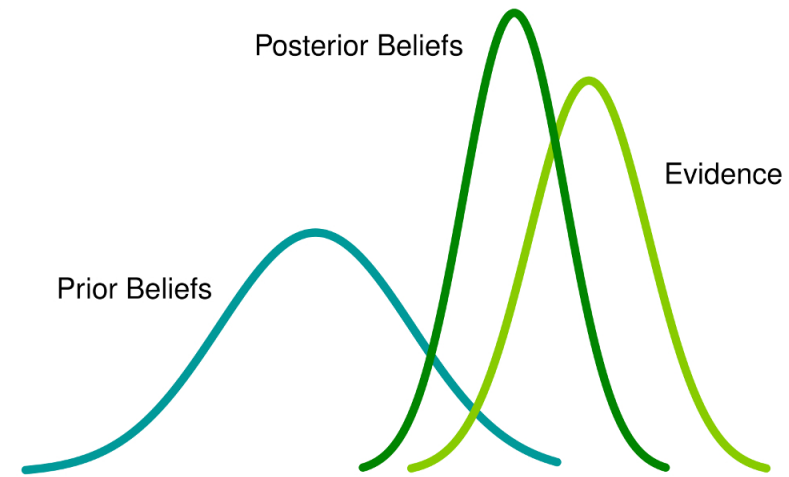
Bayesian inference is a means of combining prior beliefs with the data (evidence) to obtain posterior beliefs.

Example: likelihood update

$$f(\mathbf{z}|\mathbf{x}) \propto f(\mathbf{x}|\mathbf{z})f(\mathbf{z})$$

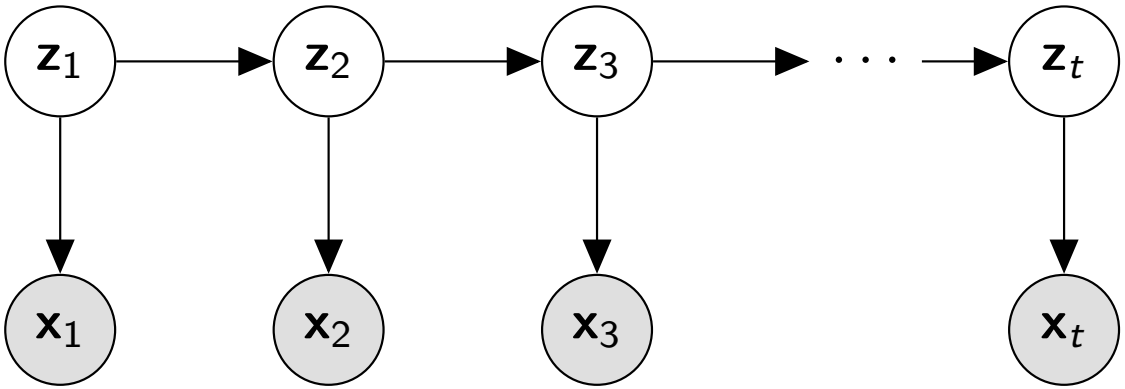
Probability Calculus

$$f(\mathbf{z},\mathbf{x}) = f(\mathbf{z}|\mathbf{x})f(\mathbf{x})$$
$$f(\mathbf{z},\mathbf{x}) = f(\mathbf{x}|\mathbf{z})f(\mathbf{z})$$



Online recursive algorithms

Consider a stochastic dynamical system represented by the following recursion



$$\mathbf{z}_1 \sim f(\mathbf{z}_1), \tag{1a}$$
$$\mathbf{x}_k \sim f(\mathbf{x}_k|\mathbf{z}_k), \tag{1b}$$
$$\mathbf{z}_{k+1} \sim f(\mathbf{z}_{k+1}|\mathbf{z}_k). \tag{1c}$$

The Bayesian filtering recursion corresponds to computing the posterior distributions $f(\mathbf{z}_k|\mathbf{x}_{1:k})$;

$$f(\mathbf{z}_k|\mathbf{x}_{1:k}) = \frac{f(\mathbf{z}_k|\mathbf{x}_{1:k-1})f(\mathbf{x}_k|\mathbf{z}_k)}{\int f(\mathbf{z}_k|\mathbf{x}_{1:k-1})f(\mathbf{x}_k|\mathbf{z}_k)\,\mathrm{d}\mathbf{z}_k}. \tag{2}$$

The density $f(\mathbf{z}_k|\mathbf{x}_{1:k-1})$ in the numerator of (2) which is called the predicted density of $\mathbf{z}_k$ and is obtained by integration as in

$$f(\mathbf{z}_k|\mathbf{x}_{1:k-1}) = \int f(\mathbf{z}_k|\mathbf{z}_{k-1})f(\mathbf{z}_{k-1}|\mathbf{x}_{1:k-1})\,\mathrm{d}\mathbf{z}_{k-1}. \tag{3}$$

Properties of the Normal density function

Property 1: $f(\mathbf{z})f(\mathbf{x}|\mathbf{z}) = f(\mathbf{z},\mathbf{x})$

$$N(\mathbf{z};\mu,\Sigma)N(\mathbf{x};C\mathbf{z},R) = N\left(\begin{bmatrix}\mathbf{z}\\ \mathbf{x}\end{bmatrix};\begin{bmatrix}\mu\\ C\mu\end{bmatrix},\begin{bmatrix}\Sigma & \Sigma C^{\mathrm{T}}\\ C\Sigma & C\Sigma C^{\mathrm{T}}+R\end{bmatrix}\right)$$

Property 2: marginalization and conditioning

If x, y were jointly normal:

$$f(x,y) = N\left(\begin{bmatrix}x\\ y\end{bmatrix};\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$

then

$$\begin{aligned}
f(x) &= N(x;\mu_1,\Sigma_{11})\\
f(y) &= N(y;\mu_2,\Sigma_{22})\\
f(x|y) &= N(x;\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(y-\mu_2),\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\\
f(y|x) &= N(y;\mu_2+\Sigma_{21}\Sigma_{11}^{-1}(x-\mu_1),\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})
\end{aligned}$$

The Kalman Filter's Foundation

Let $\mathbf{z}$ have a normal prior distribution with mean μ and covariance Σ , i.e., $\mathbf{z} \sim N(\mathbf{z}; \mu, \Sigma)$.

An observation $\mathbf{x}$ with the likelihood function $f(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}; C\mathbf{z}, R)$ is in hand where C is a matrix with proper dimensions and R is a covariance matrix. The posterior distribution of $\mathbf{z}$ can be obtained using the Bayes' rule

$$f(\mathbf{z}|\mathbf{x}) = \frac{f(\mathbf{z})f(\mathbf{x}|\mathbf{z})}{\int f(\mathbf{z})f(\mathbf{x}|\mathbf{z})\,\mathrm{d}\mathbf{z}} \tag{4}$$

$$= \frac{N(\mathbf{z}; \mu, \Sigma)N(\mathbf{x}; C\mathbf{z}, R)}{\int N(\mathbf{z}; \mu, \Sigma)N(\mathbf{x}; C\mathbf{z}, R)\,\mathrm{d}\mathbf{z}}. \tag{5}$$

The posterior distribution $f(\mathbf{z}|\mathbf{x})$ has an analytical solution and turns out to be the normal distribution $N(\mathbf{z}; \mu', \Sigma')$ where

$$\mu' = \mu + K(\mathbf{x} - C\mu), \tag{6a}$$

$$\Sigma' = \Sigma - KC\Sigma, \tag{6b}$$

where

$$K = \Sigma C^{\mathrm{T}}(C\Sigma C^{\mathrm{T}} + R)^{-1}. \tag{7}$$

Time Series Analysis

Lecture 7: State Space Model - Estimation

Tohid Ardeshiri

Linköping University
Division of Statistics and Machine Learning

October 2, 2019

Kalman filter

Kalman filter is an algorithm that uses time series data, containing statistical noise and unknown innovations, and produces estimates of latent (hidden) process that tend to be more accurate than those based on a single observations using a probabilistic framework.

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + e_t,$$
$$\mathbf{x}_t = C\mathbf{z}_t + \nu_t,$$

Kalman filtering output is

$$f(\mathbf{z}_t|\mathbf{x}_{1:t}).$$

That is, it computes the the posterior density of $\mathbf{z}_t$ using the observations up to time t.

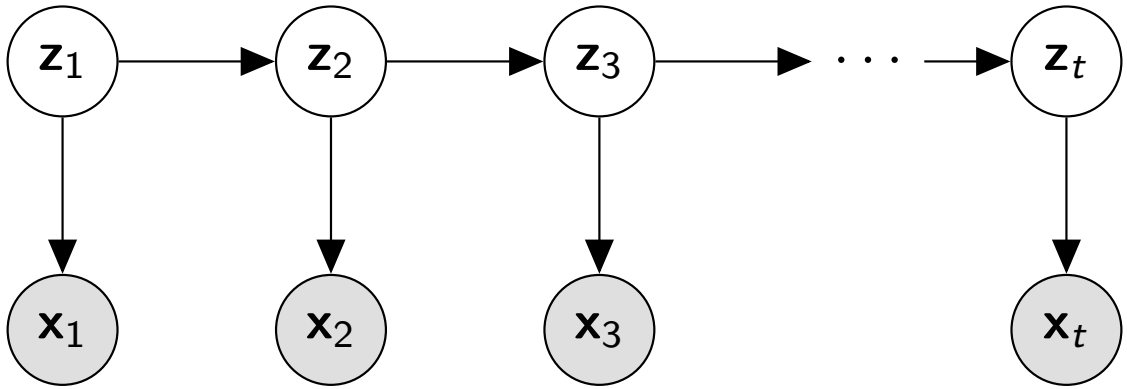
Kalman filtering recursion

1. initial estimate at $t = 1 \rightarrow N(\mathbf{z}_1; m_0, P_0)$
2. observation update using $\mathbf{x}_t$ and $\mathbf{x}_t = C\mathbf{z}_t + \nu_t \rightarrow N(\mathbf{z}_t; m_{t|t}, P_{t|t})$
3. prediction using $\mathbf{z}_{t+1} = A\mathbf{z}_t + e_{t+1} \rightarrow N(\mathbf{z}_t; m_{t+1|t}, P_{t+1|t})$
4. $t \leftarrow t + 1$
5. go to 2

State Space models - Time varying

State space models can be time-varying

$$\mathbf{z}_t = A_t \mathbf{z}_{t-1} + e_t, \qquad e_t \sim N(0, Q_t)$$
$$\mathbf{x}_t = C_t \mathbf{z}_t + \nu_t, \qquad \nu_t \sim N(0, R_t)$$



State space models with known deterministic input

State space model with input $\mathbf{u}$.

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + B\mathbf{u}_{t-1} + e_t,$$
$$\mathbf{x}_t = C\mathbf{z}_t + \nu_t,$$

Initialization:

$$f(\mathbf{z}_1) = N(\mathbf{z}_1; m_{1|0}, P_{1|0})$$

Inputs: A, B, C, Q, R, $\mathbf{u}_{1:T}$, $\mathbf{x}_{1:T}$, $m_{1|0}$, $P_{1|0}$
for $t = 1$ to T **do**
Kalman filter observation update step
    $K_t \leftarrow P_{t|t-1}C^{\mathrm{T}}(CP_{t|t-1}C^{\mathrm{T}} + R)^{-1}$
    $m_{t|t} \leftarrow m_{t|t-1} + K_t(\mathbf{x}_t - Cm_{t|t-1})$
    $P_{t|t} \leftarrow P_{t|t-1} - K_t CP_{t|t-1}$
Kalman filter prediction step
    $m_{t+1|t} \leftarrow Am_{t|t} + B\mathbf{u}_t$
    $P_{t+1|t} \leftarrow AP_{t|t}A^{\mathrm{T}} + Q$
end for
Outputs: $m_{t|t}$ **and** $P_{t|t}$ **for** $t = 1 : T$

Kalman Smoothing

The purpose of Kalman smoothing is to compute the marginal posterior distribution of $\mathbf{z}_t$ at time t after receiving observations up to time T where $T > t$:

$$f(\mathbf{z}_t|\mathbf{x}_{1:T}) = N(\mathbf{z}_t; m_{t|T}, P_{t|T})$$

The RTS smoother uses a Kalman filter in its forward path. In its backwards path it updates the densities using the relation

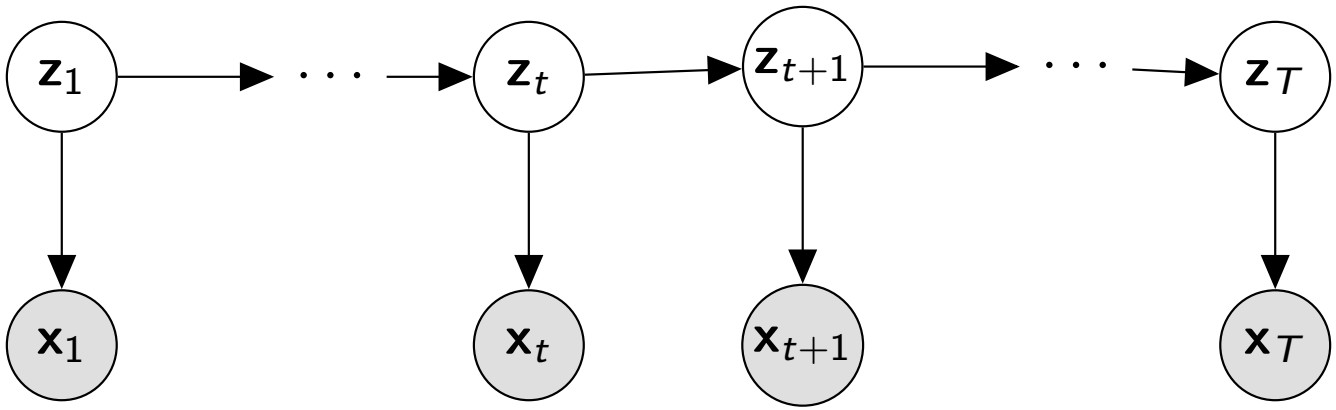
$$\mathbf{z}_t = A_{t-1}\mathbf{z}_{t-1} + e_t$$

RTS Smoother's derivation

Assume $f(\mathbf{z}_{t+1}|\mathbf{x}_{1:T})$ is available as in

$$f(\mathbf{z}_{t+1}|\mathbf{x}_{1:T}) = N(\mathbf{z}_{t+1}; m_{t+1|T}, P_{t+1|T})$$

For example $f(\mathbf{z}_T|\mathbf{x}_{1:T})$ which is the filtering density of $\mathbf{z}_T$ is available after filtering.



The objective is to compute $f(\mathbf{z}_t, \mathbf{z}_{t+1}|\mathbf{x}_{1:T})$.

RTS Smoother's derivation

The joint posterior $f(\mathbf{z}_t, \mathbf{z}_{t+1}|\mathbf{x}_{1:t})$ can be written as

$$\begin{aligned} f(\mathbf{z}_t, \mathbf{z}_{t+1}|\mathbf{x}_{1:t}) =& N(\mathbf{z}_t; m_{t|t}, P_{t|t}) N(\mathbf{z}_{t+1}; A\mathbf{z}_t, Q) \\ =& N\left(\begin{bmatrix} \mathbf{z}_t \\ \mathbf{z}_{t+1} \end{bmatrix}, \begin{bmatrix} m_{t|t} \\ Am_{t|t} \end{bmatrix}, \begin{bmatrix} P_{t|t} & P_{t|t}A^{\mathrm{T}} \\ AP_{t|t} & AP_{t|t}A^{\mathrm{T}} + Q \end{bmatrix} \right) \end{aligned}$$

Using the conditioning property of the multivariate normal distribution $f(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x}_{1:t})$ can be computed as a normal density as given in the following:

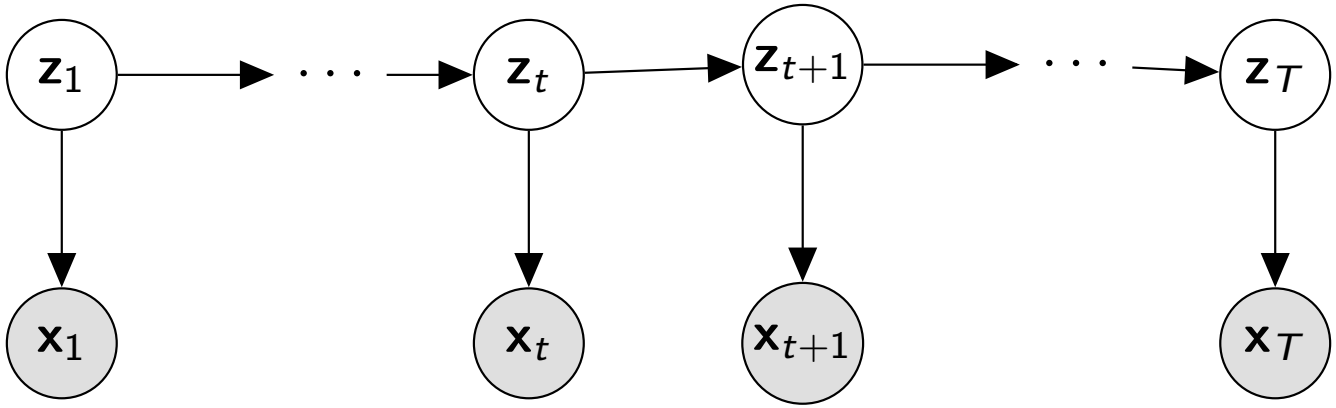
$$f(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x}_{1:t}) = N(\mathbf{z}_t; \tilde{m}_t, \tilde{P}_t)$$

where $\tilde{m}_t$ is a function of $\mathbf{z}_{t+1}$.

RTS Smoother's derivation

Note the Markov property

$$f(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x}_{1:T}) = f(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x}_{1:t})$$



Assume $f(\mathbf{z}_{t+1}|\mathbf{x}_{1:T})$ is available as in

$$f(\mathbf{z}_{t+1}|\mathbf{x}_{1:T}) = N(\mathbf{z}_{t+1}; m_{t+1|T}, P_{t+1|T})$$

Recall that

$$\begin{aligned} f(\mathbf{z}_{t+1}, \mathbf{z}_t|\mathbf{x}_{1:T}) &= f(\mathbf{z}_{t+1}|\mathbf{x}_{1:T}) f(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x}_{1:T}) \\ &= f(\mathbf{z}_{t+1}|\mathbf{x}_{1:T}) f(\mathbf{z}_t|\mathbf{z}_{t+1}, \mathbf{x}_{1:t}) \\ &= N(\mathbf{z}_{t+1}; m_{t+1|T}, P_{t+1|T}) N(\mathbf{z}_t; \tilde{m}_t, \tilde{P}_t) \end{aligned}$$

RTS Smoother's derivation **Whiteboard**

where

$$G_t = P_{t|t} A_t^{\mathrm{T}} (A P_{t|t} A^{\mathrm{T}} + Q)^{-1} = P_{t|t} A_t^{\mathrm{T}} P_{t+1|t}^{-1}$$
$$\tilde{m}_t = m_{t|t} + G_t (\mathbf{z}_{t+1} - A m_{t|t})$$
$$\tilde{P}_t = P_{t|t} - G_t (A P_{t|t} A^{\mathrm{T}} + Q) G_t^{\mathrm{T}} = P_{t|t} - G_t P_{t+1|t} G_t^{\mathrm{T}}$$

Hence,

$$f(\mathbf{z}_{t+1}, \mathbf{z}_t | \mathbf{x}_{1:T}) = N(\mathbf{z}_{t+1}; m_{t+1|T}, P_{t+1|T}) N(\mathbf{z}_t; \tilde{m}_t, \tilde{P}_t)$$
$$= N\left(\begin{bmatrix} \mathbf{z}_t \\ \mathbf{z}_{t+1} \end{bmatrix}, \begin{bmatrix} m_{t|t} + G_t (m_{t+1|T} - A m_{t|t}) \\ m_{t+1|T} \end{bmatrix}, \begin{bmatrix} G_t P_{t+1|T} G_t^{\mathrm{T}} + \tilde{P}_t & G_t P_{t+1|T} \\ P_{t+1|T} G_t^{\mathrm{T}} & P_{t+1|T} \end{bmatrix} \right)$$

RTS Smoother's derivation **Whiteboard**

The smoothing density's parameters is given by

$$
\begin{aligned}
G_t &= P_{t|t} A_t^{\mathrm{T}} (A P_{t|t} A^{\mathrm{T}} + Q)^{-1} = P_{t|t} A_t^{\mathrm{T}} P_{t+1|t}^{-1} \\
m_{t|T} &= m_{t|t} + G_t (m_{t+1|T} - A m_{t|t}) \\
P_{t|T} &= \tilde{P}_t + G_t P_{t+1|T} G_t^{\mathrm{T}} = P_{t|t} - G_t P_{t+1|t} G_t^{\mathrm{T}} + G_t P_{t+1|T} G_t^{\mathrm{T}} \\
&= P_{t|t} + G_t (P_{t+1|T} - P_{t+1|t}) G_t^{\mathrm{T}}
\end{aligned}
$$

RTS smoother's backwards recursion

Prove the backwards recursion of the RTS smoother for following state space model with initial prior on the state $f(\mathbf{z}_1) = N(\mathbf{z}_1; m_0, P_0)$

$$\mathbf{z}_t = A_{t-1}\mathbf{z}_{t-1} + e_t, \qquad e_t \sim N(0, Q_t)$$
$$\mathbf{x}_t = C_t\mathbf{z}_t + \nu_t, \qquad \nu_t \sim N(0, R_t)$$

1: **Inputs:** A_t, Q_t, $m_{t|t}$, $P_{t|t}$, $m_{t+1|t}$, $P_{t+1|t}$ for $1 \leq t \leq T$
initialization
2: **for** $t =$ T-1 down to 1 **do**
3: $\quad G_t \leftarrow P_{t|t}A_t^{\mathrm{T}}P_{t+1|t}^{-1}$
4: $\quad m_{t|T} \leftarrow m_{t|t} + G_t(m_{t+1|T} - A_t m_{t|t})$
5: $\quad P_{t|T} \leftarrow P_{t|t} + G_t(P_{t+1|T} - P_{t+1|t})G_t^{\mathrm{T}}$
6: **end for**
7: **Outputs:** $m_{t|T}$, $P_{t|T}$

State Space models - Estimation

We consider three approaches.

1. (Variational Bayes)
   T. Ardeshiri, E. Özkan, U. Orguner and F. Gustafsson, "**Approximate Bayesian Smoothing with Unknown Process and Measurement Noise Covariances**," in IEEE Signal Processing Letters, vol. 22, no. 12, pp. 2450-2454, Dec. 2015.
2. Direct maximum likelihood estimate
3. Expectation maximization (EM)

Variational Bayes smoothing with unknwn time varying R_t and Q_t

Consider a Linear and Gaussian state space model with parameters

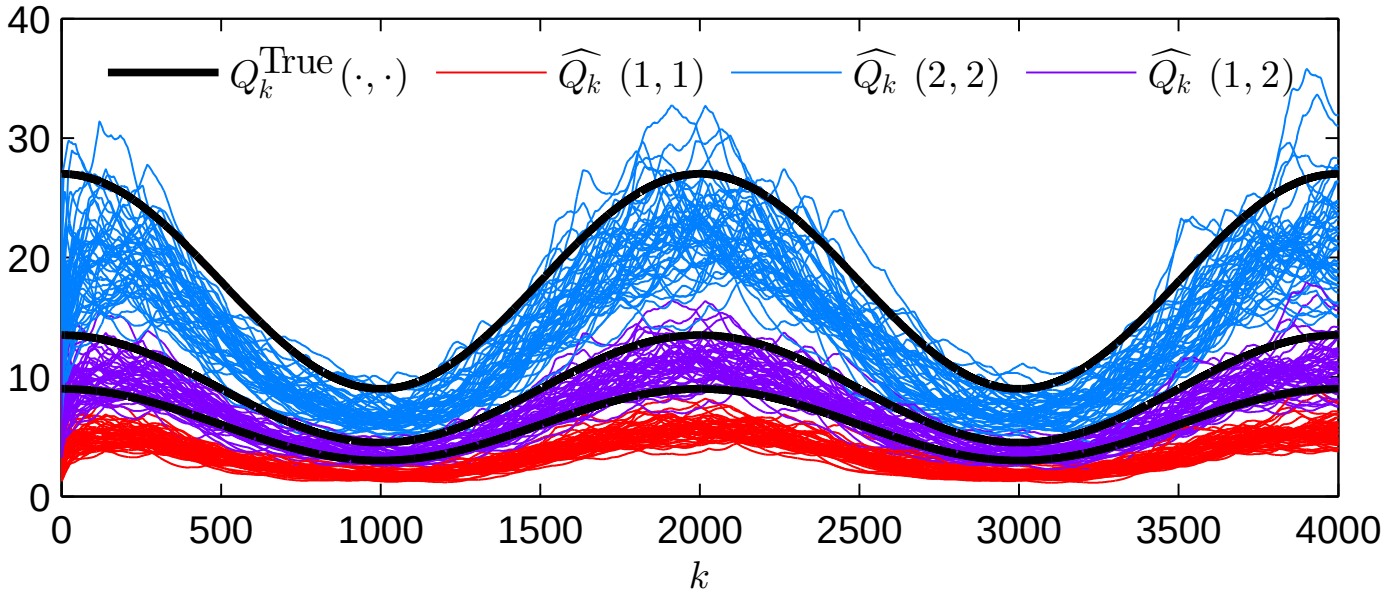
$$A_k = \text{Diag}(a, a), \qquad Q_0 = \text{Diag}(q, q),$$

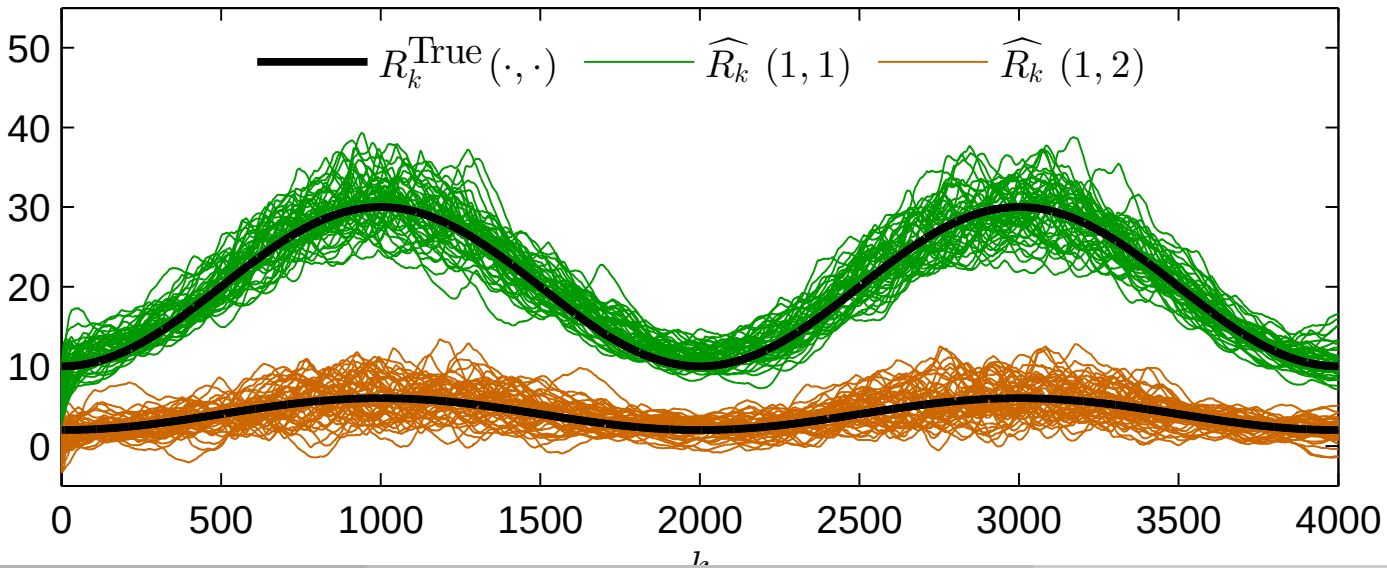
$$a = \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix}, \qquad q = \sigma_\nu^2 \begin{bmatrix} \tau^3/3 & \tau^2/2 \\ \tau^2/2 & \tau \end{bmatrix},$$

$$R_k^{\text{True}} = \left(2 - \cos\left(\frac{4\pi k}{K}\right)\right) R_0, \qquad R_0 = \sigma_e^2 \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix},$$

$$Q_k^{\text{True}} = \left(\frac{2}{3} + \frac{1}{3}\cos\left(\frac{4\pi k}{K}\right)\right) Q_0, \qquad C_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Variational Bayes smoothing with unknwn time varying R_t and Q_t





Maximum likelihood methods **Whiteboard**

Let $\theta = \{A, C, R, Q, m_0, P_0\}$ denote the unknown state space parameters

$$f(\mathbf{x}_{1:T}|\theta) = f(\mathbf{x}_1|\theta)f(\mathbf{x}_2|\mathbf{x}_1, \theta)f(\mathbf{x}_3|\mathbf{x}_{1:2}, \theta)\cdots f(\mathbf{x}_T|\mathbf{x}_{1:T-1}, \theta)$$

where

$$f(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \theta) = \int f(\mathbf{x}_{t+1}|\mathbf{z}_{t+1}, \mathbf{x}_{1:t}, \theta)f(\mathbf{z}_{t+1}|\mathbf{x}_{1:t}, \theta)\, \mathrm{d}\mathbf{z}_{t+1}$$

This can be computed using the Kalman filter

$$\begin{aligned} f(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \theta) &= \int f(\mathbf{x}_{t+1}|\mathbf{z}_{t+1}, \mathbf{x}_{1:t}, \theta)f(\mathbf{z}_{t+1}|\mathbf{x}_{1:t}, \theta)\, \mathrm{d}\mathbf{z}_{t+1} \\ &= \int N(\mathbf{x}_{t+1}; C\mathbf{z}_{t+1}, R)N(\mathbf{z}_{t+1}; m_{t+1|t}, P_{t+1|t})\, \mathrm{d}\mathbf{z}_{t+1} \\ &= N(\mathbf{x}_{t+1}; Cm_{t+1|t}, CP_{t+1|t}C^{\mathrm{T}} + R) \end{aligned}$$

Maximum likelihood methods **Whiteboard**

The negative logarithm of the likelihood becomes

$$
\begin{aligned}
l(\theta) &= -\sum_{t=1}^{T} \log f(\mathbf{x}_t | \mathbf{x}_{1:t-1}, \theta) \\
&= -\sum_{t=1}^{T} \log N(\mathbf{z}_{t+1}; Cm_{t+1|t}, CP_{t+1|t}C^{\mathrm{T}} + R) \\
&= \frac{1}{2}\sum_{t=1}^{T} \log |CP_{t+1|t}C^{\mathrm{T}} + R| \\
&+ \frac{1}{2}\sum_{t=1}^{T} (\mathbf{x}_t - Cm_{t+1|t})(CP_{t+1|t}C^{\mathrm{T}} + R)^{-1}(\mathbf{x}_t - Cm_{t+1|t})^{\mathrm{T}}
\end{aligned}
$$

which can be solved using for example Newton-Raphson method.

Maximum likelihood methods

The first two derivatives of the negative log-likelihood is computed with respect to the θ.
Then in the iterations of the Newton-Raphson method

1. An initial value for for θ is selected, say $\theta^{(0)}$.
2. A Kalman filter is run to compute the quantities for the first two derivatives of $l(\theta)$.
3. A new set of parameters are obtained from a Newton-Raphson procedure.
4. Iterations are repeated until convergence.

Expectation Maximization **Whiteboard**

- Expectation-maximization (EM) method can be used to compute the maximum likelihood (ML) estimate of the state space parameters.
- In the E (Expectation) step of the EM algorithm the conditional expectation of the joint log-likelihood is computed using the last estimates of the unknown parameters as in

$$\mathcal{Q} = E\left[\log f(\mathbf{z}_{1:T}, \mathbf{x}_{1:T}) \middle| \mathbf{x}_{1:T}, \theta^{(i)}\right] \tag{1}$$

where

$$\begin{aligned}\log f(\mathbf{z}_{1:T}, \mathbf{x}_{1:T}) = &\log N(\mathbf{z}_1; m_0, P_0) - \frac{T+1}{2}\log|R| \\ &- \frac{1}{2}\sum_{t=1}^{T}\mathrm{Tr}\left(R^{-1}(\mathbf{x}_t - C\mathbf{z}_t)(\mathbf{x}_t - C\mathbf{z}_t)^{\mathrm{T}}\right) - \frac{T}{2}\log|Q| \\ &- \frac{1}{2}\sum_{t=1}^{T-1}\mathrm{Tr}\left(Q^{-1}(\mathbf{z}_{t+1} - A\mathbf{z}_t)(\mathbf{z}_{t+1} - A\mathbf{z}_t)^{\mathrm{T}}\right) + c. \end{aligned} \tag{2}$$

Therefore,

$$\begin{aligned}\mathcal{Q} = &-\frac{1}{2}E[(\mathbf{z}_0 - m_0)P_0^{-1}(\mathbf{z}_0 - m_0)^{\mathrm{T}} + \log|P_0|] \\ &-\frac{T+1}{2}\log|R| - \frac{1}{2}\mathrm{Tr}\left(R^{-1}\sum_{t=0}^{T}E[(\mathbf{x}_t - C\mathbf{z}_t)(\mathbf{x}_t - C\mathbf{z}_t)^{\mathrm{T}}|\mathbf{x}_{1:T}]\right) \\ &-\frac{T}{2}\log|Q| - \frac{1}{2}\mathrm{Tr}\left(Q^{-1}\sum_{t=0}^{T-1}E[(\mathbf{z}_{t+1} - A\mathbf{z}_t)(\mathbf{z}_{t+1} - A\mathbf{z}_t)|\mathbf{x}_{1:T}]\right) + c,\end{aligned} \tag{3}$$

Computing the Expectiations

Whiteboard

In order to compute the expectations the RTS smoother's posterior $f(\mathbf{z}_t|\mathbf{z}_{1:T})$ is used.
Then in the iterations of the EM method

1. An initial value for for θ is selected, say $\theta^{(0)}$.
2. A Kalman smoother is run using $\theta^{(i)}$
3. In the expectation step $\mathcal{Q}$ function as a function of θ is derived.
4. A new set of parameters $\theta^{(i+1)}$ are obtained from maximization of the $\mathcal{Q}$ function.
5. Iterations are repeated until convergence.

Read home

- Shumway and Stoffer, Chapter 6.3

Time Series Analysis

Lecture 8: State Space Model
Stochastic Volatility

Tohid Ardeshiri

Linköping University
Division of Statistics and Machine Learning

October 4, 2019

Remaining Course topics

- ARIMA models
- State space models (2 lectures, 1 teaching session with hand-in, 1 computer lab with short report)
  - Linear and Gaussian state space models (Chapter 6.1)
  - Kalman filtering, Kalman smoothing and Forecasting (Chapter 6.2)
  - Maximum likelihood estimate of the state space models (Chapter 6.3)
  - Stochastic volatility (Chapter 6.11)
- Recurrent Neural Networks (RNNs) (1 lecture and 1 Computer lab No examination)
- Summary lecture

Why Stochastic volatility

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + e_t, \qquad e_t \sim N(0, Q)$$
$$\mathbf{x}_t = C\mathbf{z}_t + \nu_t, \qquad \nu_t \sim N(0, R)$$

- Filtering: Kalman filtering, $f(\mathbf{z}_t|\mathbf{x}_{1:t})$
- Smooting: Kalman smoothing, $f(\mathbf{z}_t|\mathbf{x}_{1:T})$
- Modelling: Maximum likelihood and EM, $\widehat{\theta} = \arg\max_\theta f(\mathbf{x}_{1:T}|\theta)$
- Case study on Stochastic volatility via a generalization of the above tools

Stochastic Volatility

In finance, **return** is a profit on an investment. It comprises any change in value of the investment, and/or cash flows which the investor receives from the investment, such as interest payments or dividends.

Stochastic volatility models are those in which the variance of a stochastic process is itself randomly distributed.

In the following:

- r_t denote the return of some financial asset. A common model for the return is

$$r_t = \beta \sigma_t \epsilon_t$$

- σ_t is the volatility process and
- ϵ_t is an iid sequence and $\epsilon_t \sim iid(0, 1)$ and ϵ_t is independent of past σ_s $(s \leq t)$

Stochastic Volatility

In the following:

- r_t denote the return of some financial asset. A common model for the return is

$$r_t = \beta \sigma_t \epsilon_t$$

- σ_t is the volatility process and
- ϵ_t is an iid sequence and $\epsilon_t \sim iid(0, 1)$ and ϵ_t is independent of past σ_s $(s \leq t)$
- Let $\mathbf{z}_t = \log \sigma_t^2$ and consider the hidden autoregressive model

$$\mathbf{z}_t = \phi \mathbf{z}_{t-1} + w_t$$
$$r_t = \beta \exp(\mathbf{z}_t/2)\epsilon_t$$

In this model $w_t \sim iidN(0, \sigma_w^2)$ and ϵ_t is iid noise with finite moments.

$$\mathbf{z}_t = \phi \mathbf{z}_{t-1} + w_t$$
$$r_t = \beta \exp(\mathbf{z}_t/2)\epsilon_t$$

Furthermore, let $\mathbf{x}_t = \log r_t^2$ and $\nu_t = \log \epsilon_t^2$. We obtain

$$\mathbf{x}_t = \alpha + \mathbf{z}_t + \nu_t$$

We can move the α to the state equation and rewrite it as

$$\mathbf{z}_t = \phi_0 + \phi_1 \mathbf{z}_{t-1} + w_t$$
$$\mathbf{x}_t = \mathbf{z}_t + \nu_t$$

where the ϕ_0 is called the leverage effect.

Stochastic Volatility

The distribution of ν_t is not Gaussian because

$$\nu_t = \log \epsilon_t^2 \text{ and}$$
$$\epsilon_t \sim iidN(0,1)$$

Hence, ν is distributed as a log of a chi-squared distribution with defree of freedom 1 with density

$$f(\nu) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(e^{\nu} - \nu)\} \qquad -\infty < \nu < \infty$$

Stochastic Volatility - Gaussian mixture approximation

Instead let us approximate $f(\nu)$ by a Gaussian mixture

$$f(\eta) = \pi_0 N(\eta; 0, \sigma_0^2) + \pi_1 N(\eta; \mu_1, \sigma_1^2)$$

That is,

$$\eta_t = I_t n_{t0} + (1 - I_t) n_{t1}$$

where I_t is an iid Bernoulli process where $Pr\{I = 0\} = \pi_0$ and $Pr\{I = 1\} = \pi_1$, $\pi_0 + \pi_1 = 1$. Also,

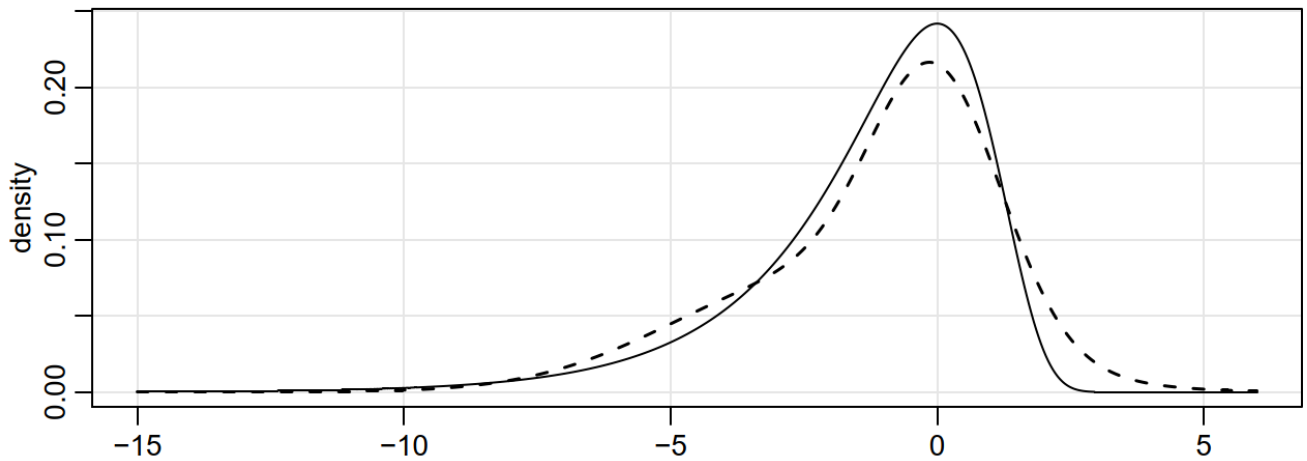
$$n_{t0} \sim N(0, \sigma_0^2)$$
$$n_{t1} \sim N(\mu_1, \sigma_1^2)$$

Stochastic Volatility - Gaussian sum approximation

$$f(\eta) = \pi_0 N(\eta; 0, \sigma_0^2) + \pi_1 N(\eta; \mu_1, \sigma_1^2) \qquad -\infty < \eta < \infty$$

$$f(\nu) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(e^{\nu} - \nu)\} \qquad -\infty < \nu < \infty$$

$f(\nu)$ and $f(\eta)$ are plotted for comparison. The dashed line is the Gaussian sum approximation, $f(\eta)$.



Stochastic Volatility - Gaussian sum formulation

The problem is finding the filtering distribution of $\mathbf{z}_t|\mathbf{x}_{1:t}$ when

$$\mathbf{z}_t = \phi_0 + \phi_1 \mathbf{z}_{t-1} + w_t$$
$$\mathbf{x}_t = \mathbf{z}_t + \eta_t$$

and

$$w_t \sim iidN(0, \sigma_w^2)$$
$$\eta_t \sim \pi_0 N(0, \sigma_0^2) + \pi_1 N(\mu_1, \sigma_1^2)$$

where $\pi_0 + \pi_1 = 1$

Gaussian sum filter - derivation

Whiteboard

The problem is finding the filtering distribution of $\mathbf{z}_t|\mathbf{x}_{1:t}$ when

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + w_t$$
$$\mathbf{x}_t = C\mathbf{z}_t + \eta_t$$

and

$$w_t \sim iidN(0, Q)$$
$$\eta_t \sim \pi_0 N(\mu_0, R_1) + \pi_1 N(\mu_1, R_2)$$

where $\pi_0 + \pi_1 = 1$

Read home

- Shumway and Stoffer, Chapter 6.11

LINKÖPING UNIVERSITY

Time Series Analysis – Lecture 9

Recurrent and Temporal Convolutional Networks

Fredrik Lindsten, Linköping University

2019-10-14

Aim and outline

Aim:

- Introduce two popular deep-learning-based methods for time series analysis
- Highlight some formal connections with classical models (state space and auto-regressive) that you have seen in the course.

Outline:

1. Basics of neural network models — the multi-layer perceptron
2. Linear Gaussian state space models on innovation form
3. A nonlinear generalization — Recurrent Neural Networks
4. Nonlinear auto-regressive models
5. Temporal Convolutional Networks

ex) Generating text

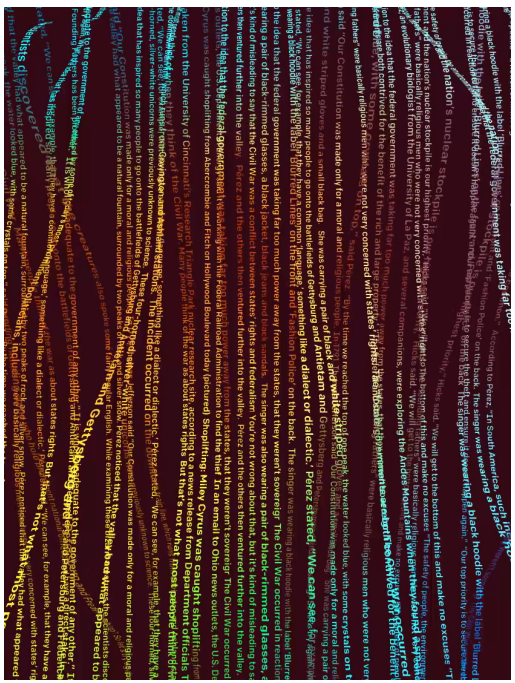
Input (human-written) In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model completion (machine-written) The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. ''By the time we reached the top of one peak, the water looked blue, with some crystals on top,'' said Pérez.



https://openai.com/blog/better-language-models/

State Space Models $\Longrightarrow$ Recurrent Neural Networks

Linear state space models

Linear state space model:

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + e_t,$$
$$x_t = C\mathbf{z}_t + \nu_t.$$

Limitation:

The next state $\mathbf{z}_{t+1}$ as well as the observation x_t depend **linearly** on the current state $\mathbf{z}_t$.

The model flexibility is limited.

Going nonlinear

Aim: Increase the flexibility of the model by replacing the linear functions by **generic** and **flexible** nonlinear functions.

Linear function: $y = A\mathbf{z}$, where the matrix A is the **parameter**.

Nonlinear function: $y = f_\theta(\mathbf{z})$. Here, θ is a vector of **parameters** determining the shape of the function $f_\theta(\cdot)$.

ex) Let $\theta = (\theta_1, \theta_2, \theta_3)$, and

$$f_\theta(z) = \frac{\theta_1}{\theta_2 + z^{\theta_3}}.$$

Neural networks

Recall: We want to use **generic** and **flexible** nonlinear functions.

This is precisely what **neural networks** provide!

Fully connected, 1-layer network:

We **construct** a function $f_\theta : \mathbb{R}^p \mapsto \mathbb{R}$ by

$$\mathbf{h} = \sigma(W^{(1)}\mathbf{z} + b^{(1)})$$
$$y = W^{(2)}\mathbf{h} + b^{(2)}.$$

That is,

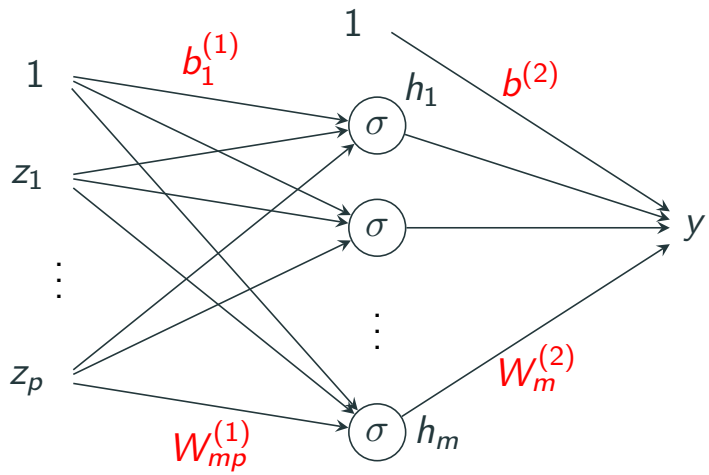
$$f_\theta(\mathbf{z}) = W^{(2)}\sigma(W^{(1)}\mathbf{z} + b^{(1)}) + b^{(2)}$$

Neural network – graphical illustration

The equations

$$\mathbf{h} = \sigma(W^{(1)}\mathbf{z} + b^{(1)})$$
$$y = W^{(2)}\mathbf{h} + b^{(2)}.$$

can be illustrated graphically.

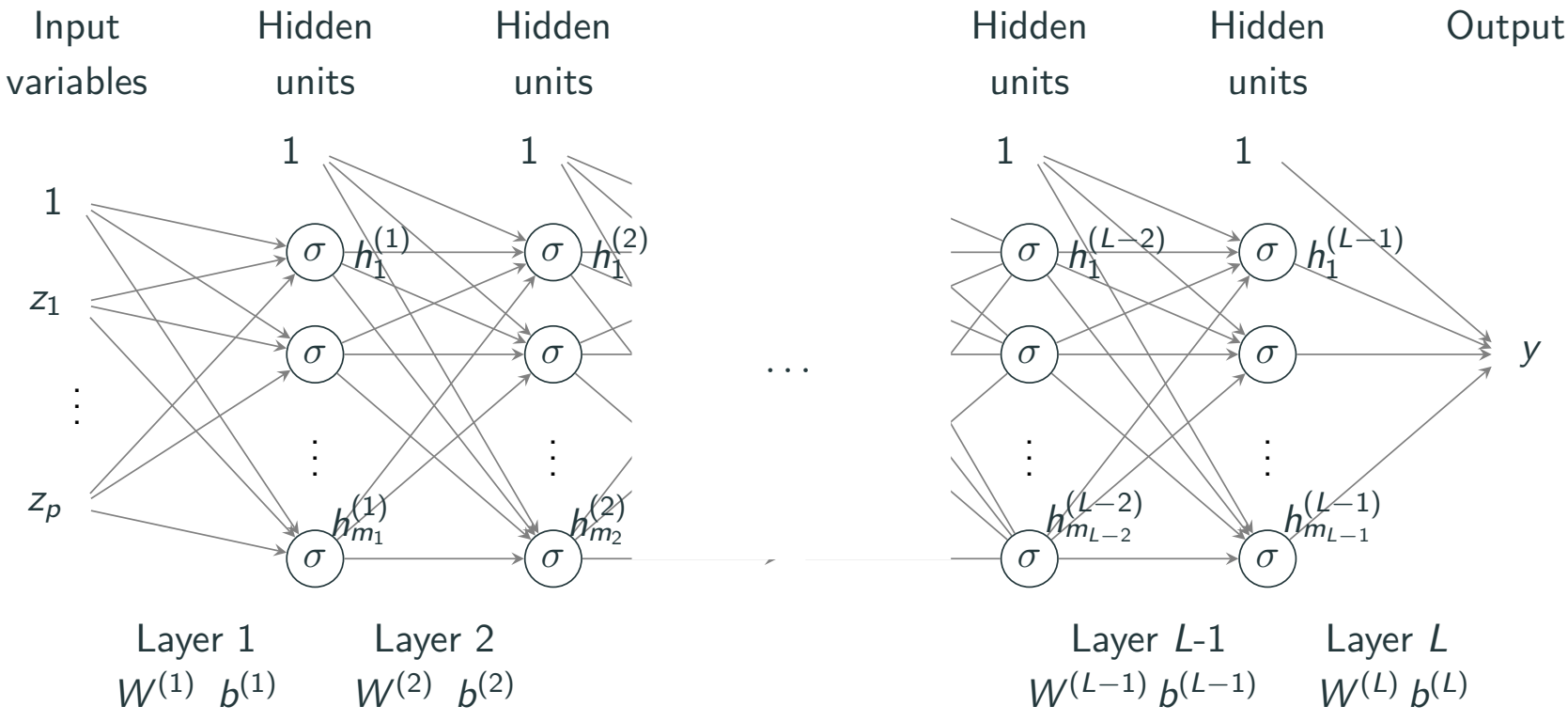


- The variables $\mathbf{h} = (h_1, \ldots, h_m)$ are referred to as a **hidden layer**.
- The function $\sigma(\cdot)$ is an element-wise nonlinearity, referred to as an **activation function**. Typical choices are

$$\sigma(x) = \tanh(x) \qquad \text{or} \qquad \sigma(x) = \text{ReLU}(x) = x\mathbb{1}(x \geq 0)$$

- The model **parameters** are the weight matrices and bias vectors $\theta = \{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$.

Multi-layer perceptron



Innovation form

Linear state space model:

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + e_t,$$
$$x_t = C\mathbf{z}_t + \nu_t.$$

Innovation form. There exists an **equivalent** representation given by

$$\mathbf{h}_t = W\mathbf{h}_{t-1} + Ux_{t-1},$$
$$x_t = C\mathbf{h}_t + \nu'_t.$$

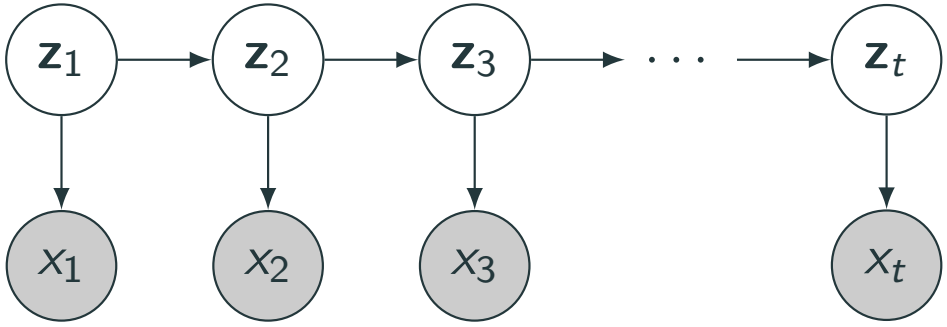
(Assuming stationarity for simplicity.)

Proof. Let $\mathbf{h}_t = m_{t|t-1}$, the Kalman predictive mean.

Innovation form

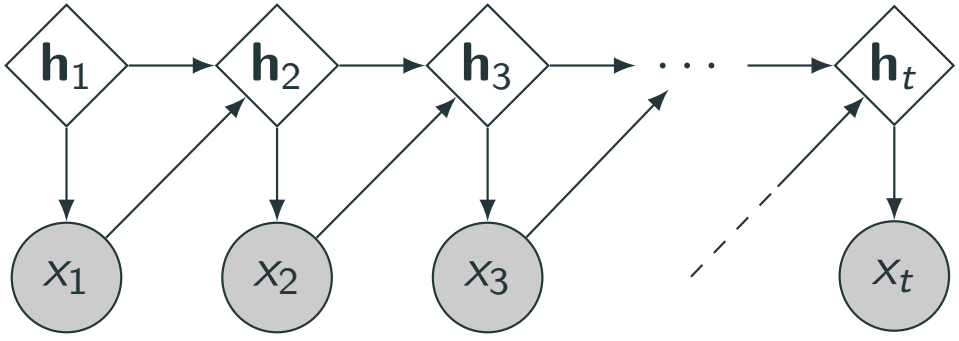
Original form:

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + e_t,$$
$$x_t = C\mathbf{z}_t + \nu_t.$$



Innovation form:

$$\mathbf{h}_t = W\mathbf{h}_{t-1} + Ux_{t-1},$$
$$x_t = C\mathbf{h}_t + \nu'_t.$$



The hidden state variables can be **deterministically and recursively computed** from the data.

Going nonlinear

Doesn't this look suspiciously similar to an MLP...?

$$\mathbf{h}_t = W\mathbf{h}_{t-1} + Ux_{t-1},$$
$$x_t = C\mathbf{h}_t + \nu'_t,$$

for some **nonlinear activation function** $\sigma(\cdot)$.

This is a basic **Recurrent Neural Network (RNN).**

Learning the parameters

The model parameters are the weight matrices and bias vectors:

$$\mathbf{h}_t = \sigma(W\mathbf{h}_{t-1} + Ux_{t-1} + b),$$
$$x_t = C\mathbf{h}_t + c + \nu'_t,$$

with $\theta = \{W, U, b, C, c\}$.

Note:

- The parameters are the same for all time steps ("weight sharing").
- The fact that there is no state noise means that we can compute

$$p_\theta(x_t \mid x_{1:t-1}) = N(x_t \mid C\mathbf{h}_t + c, \sigma^2_{\nu'}).$$

Learning the parameters

We can thus learn the parameters θ directly by optimizing the negative log-likelihood,

$$L(\theta; x_{1:T}) = -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{1:t-1}),$$

using gradient-based optimization.

The gradient $\nabla_\theta L(\theta; x_{1:T})$ is computed using the chain rule of differentiation, propagating information from $t = 1$ to $t = T$ and then back again.

$\Longrightarrow$ **Back-propagation through time.**

RNN extensions

- GRU/LSTM
- Non-Gaussian likelihood (e.g., for discrete data)
- Conditioning on context (input)
- Stochastic hidden layers
- Bidirectional connections
- ...

Autoregressive Models $\Longrightarrow$ Temporal Convolutional Nets

Autoregressive models

State space models and RNNs use a latent state vector to model temporal dependencies.

An alternative is to model the dependency of the current data point x_t on the past data points $x_{1:t-1}$ by a **direct functional relationship**.

Auto-regressive model, AR(p):

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t, \qquad w_t \sim N(0, \sigma_w^2).$$

The AR model is linear in the parameters:

- ▲ Learning of parameters easy $\Leftrightarrow$ linear regression
- ▼ Flexibility/ability to model complex temporal dependencies is limited
- ▼ Memory/receptive field is just p time steps

Going nonlinear

Nonlinear auto-regressive model, NAR(p):

$$x_t = \sigma(\phi_1 x_{t-1} + \cdots + \phi_p x_{t-p}) + w_t, \qquad w_t \sim N(0, \sigma_w^2),$$

for some nonlinear activation function σ.

- ▲ Flexibility increased...
- ▼ ...but only slightly!
- ▼ Memory/receptive field is *still* just p steps

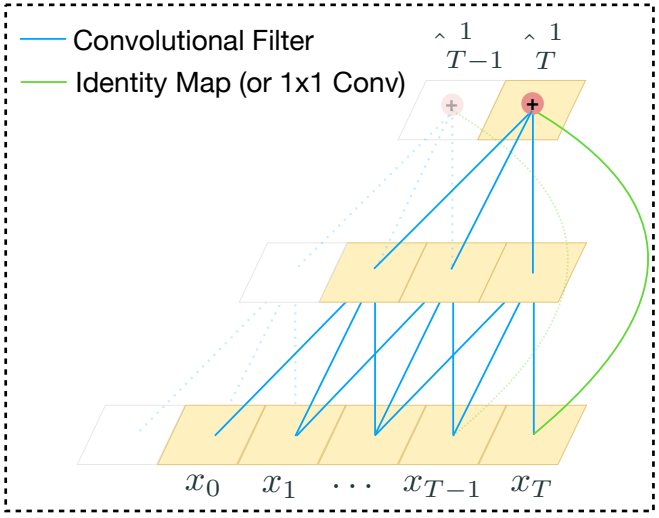
We can address these issues with a **multi-layer network architecture!**

Temporal Convolutional Network

2-layer TCN:

$$h_{t-1} = \sigma(\phi_1^{(1)} x_{t-1} + \cdots + \phi_p^{(1)} x_{t-p}),$$
$$x_t = \sigma(\phi_1^{(2)} h_{t-1} + \cdots + \phi_p^{(2)} h_{t-p}) + w_t,$$

with $w_t \sim N(0, \sigma_w^2)$.

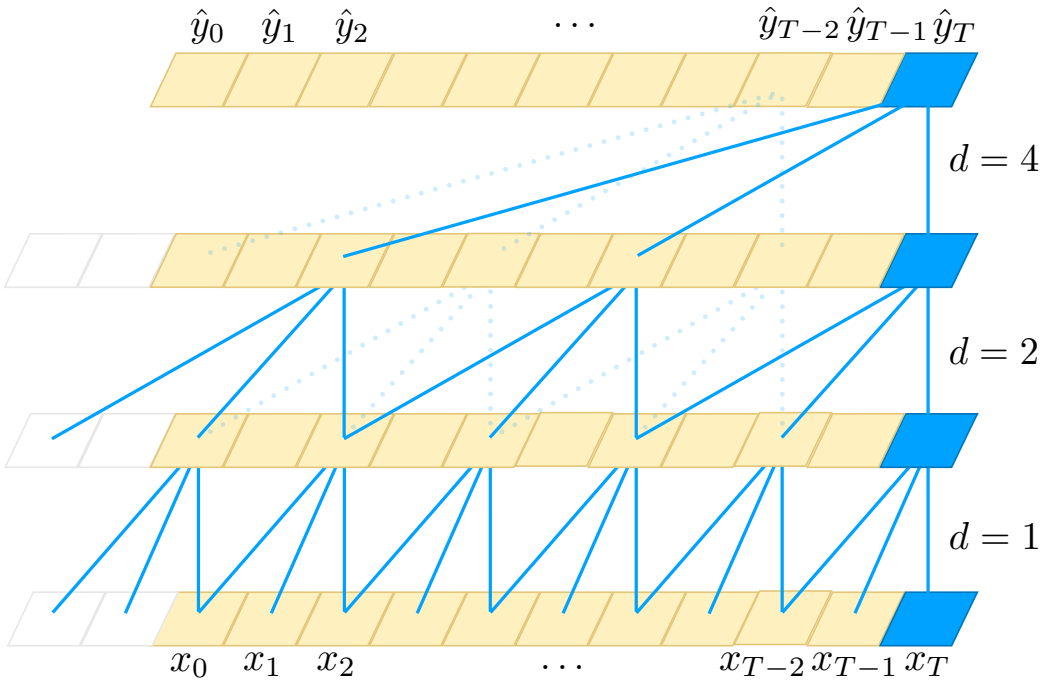


Can extend to multiple hidden layers, $h_t^{(1)}$, $h_t^{(2)}$, ...

- ▲ Multiple layers $\Rightarrow$ very flexible models
- ▲ Receptive field increases with depth...
- ▼ ...but only linearly

TCN with dilated convolutions

By using **dilated convolutions** we can increase receptive field **exponentially** with depth.



RNN vs TCN

RNNs are still the *de facto* standard deep learning approach to time series modeling, *but...*

...TCNs have outperformed them on many benchmark problems

Shaojie Bai, J. Zico Kolter, Vladlen Koltun. **An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.** *arXiv.org: 1803.01271*, 2018.

...and have other advantages too:

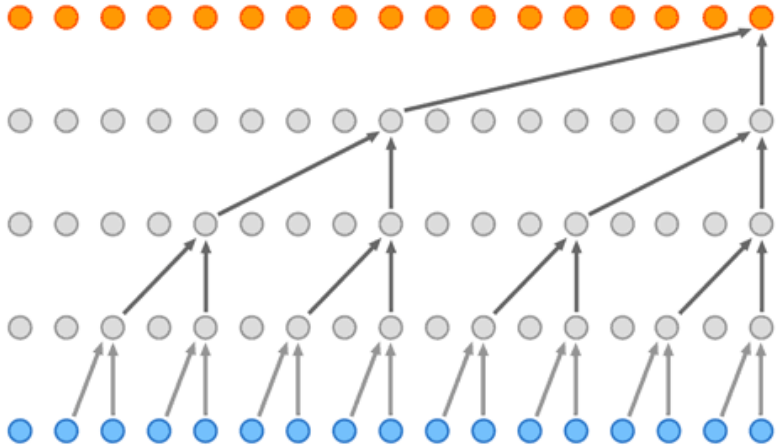
- Easier parallelization
- Better control over receptive field
- Lower memory requirements during training
- ...

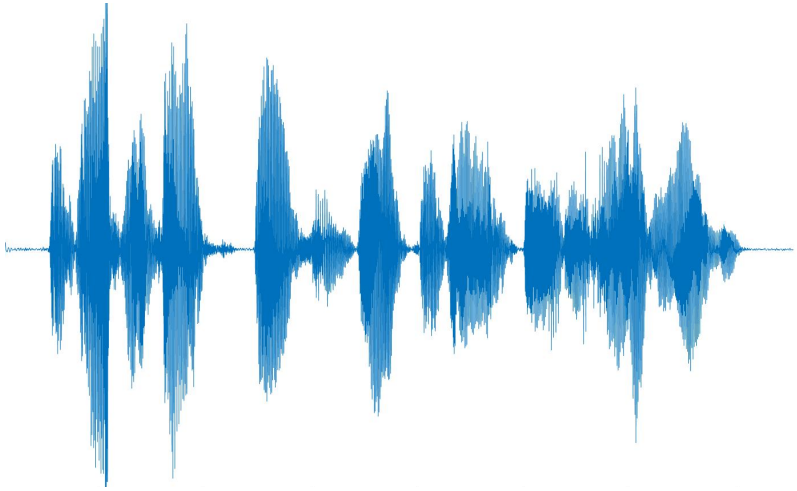
...but also some disadvantages:

- More difficult to reuse model for multiple tasks
- Larger memory requirements after deployment
- ...

ex) WaveNet

WaveNet by DeepMind powers Google's text-to-speech technology.





Deep Learning for TSA

So is this what we should always do for modeling sequential data?!

No!

- Only makes sense to use something as complex as RNN or TCN when classical methods fail — **Try Simple Things First!**
- Methods based on deep learning work best if we have multiple sequences, or one long sequence that can be split into segments
- For a single univariate time series, classical methods (ARIMA, state space, . . .) often work better.

Time Series Analysis

Lecture **X**: Summary
Questions and Answers

Tohid Ardeshiri

Linköping University
Division of Statistics and Machine Learning

October 16, 2019

Course topics

- Time series, time series regression and exploratory analysis
  - Autocovariance, ACF
  - Sample ACF
  - Stationarity, detrending, differencing,
  - transformation and smoothing
- ARIMA models
  - AR, MA, ARMA, ARIMA, seasonal ARIMA
  - PACF
  - Model selection
  - Estimation
  - Forecasting
- State space models
  - Linear and Gaussian state space models
  - Kalman filtering, Kalman smoothing and Forecasting
  - Maximum likelihood estimate of the state space models
  - Stochastic volatility
- Recurrent Neural Networks (RNNs)

Stationarity

- Time series x_t is weakly stationary (stationary) if
  - $Ex_t = const$
  - $\gamma(s,t) = \gamma(|s-t|)$
  - $\text{var}(x_t) < \infty$
- $\gamma(t, t+h) = \gamma(|t+h-t|) = \gamma(h)$
  - Autocovariance depends on lag only!
- Autocovariance for stationary process $\gamma(h) = \text{cov}(x_t, x_{t+h})$
- ACF for stationary process $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$

Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

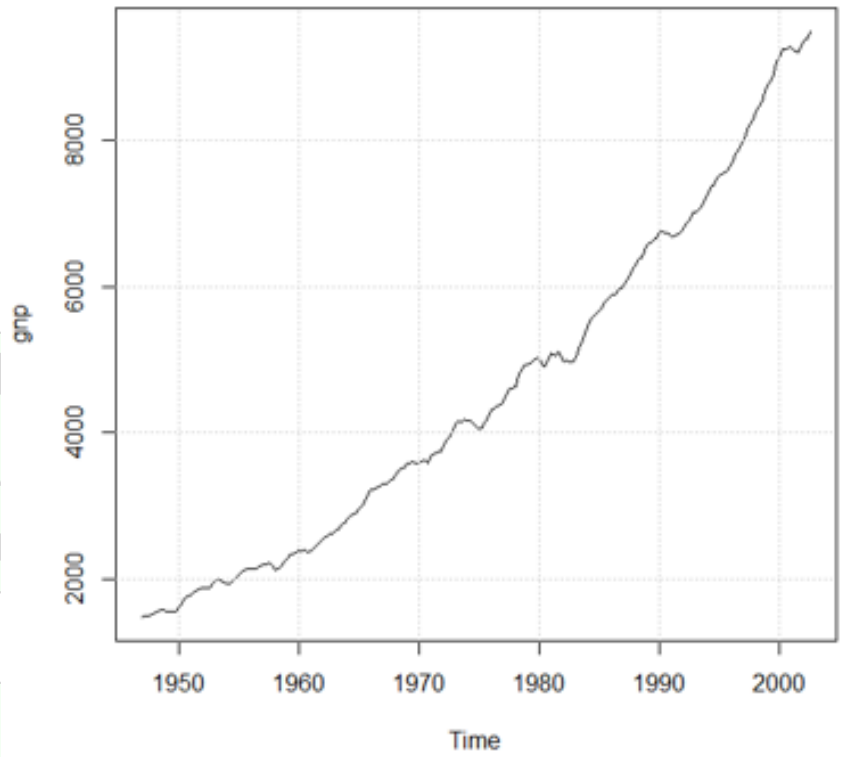
$\theta = (\theta_1, ..., \theta_q)$

Mod

- A
  ar

- Estimate $\phi's$, $\theta's$, ...

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

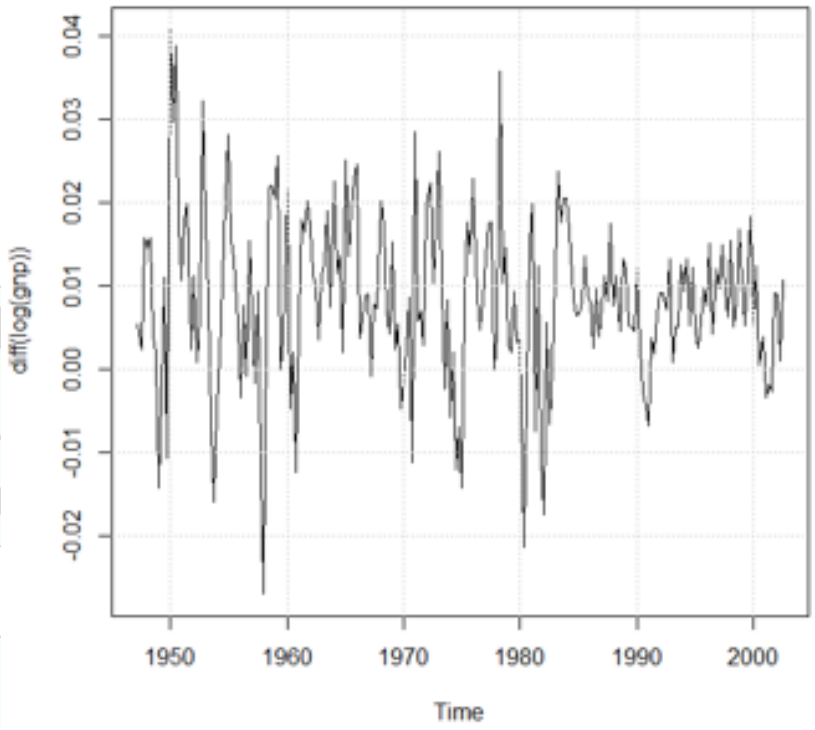
Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Prediction

$$Y_t = \nabla(\log(X_t))$$



Estimate $\phi's$, $\theta's$, ...

Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

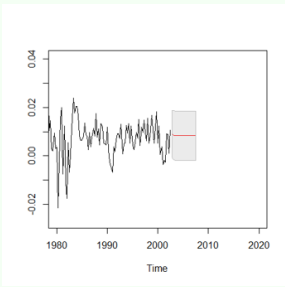
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

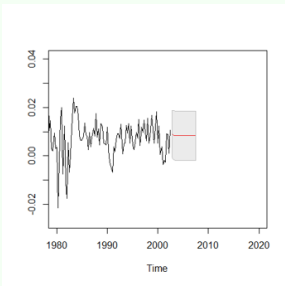
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

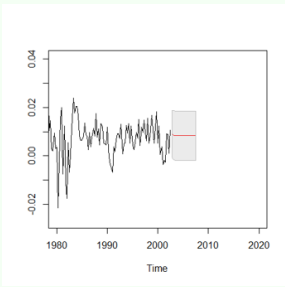
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

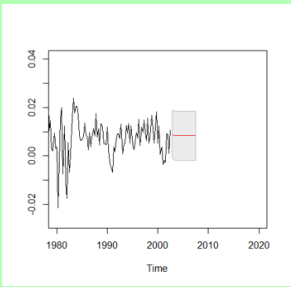
- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

- Are there better models with other p and q or even different kind?

no

Prediction



Time domain: The Big Picture

Time Series data

Exploratory data analysis

Make data stationary

Suggest a model M(ϕ, θ)

$\phi = (\phi_1, ..., \phi_p)$

$\theta = (\theta_1, ..., \theta_q)$

Model Estimation

- Estimate $\phi's$, $\theta's$, ...

Model diagnostics and Model selection

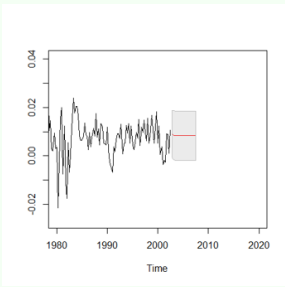
- Are there better models with other p and q or even different kind?

Revise the model

yes

no

Prediction



ARIMA modelling

- ARIMA models
  - AR, MA, ARMA, ARIMA, seasonal ARIMA
  - PACF
  - Model selection
  - Estimation
  - Forecasting

ARIMA models

Time series models so far

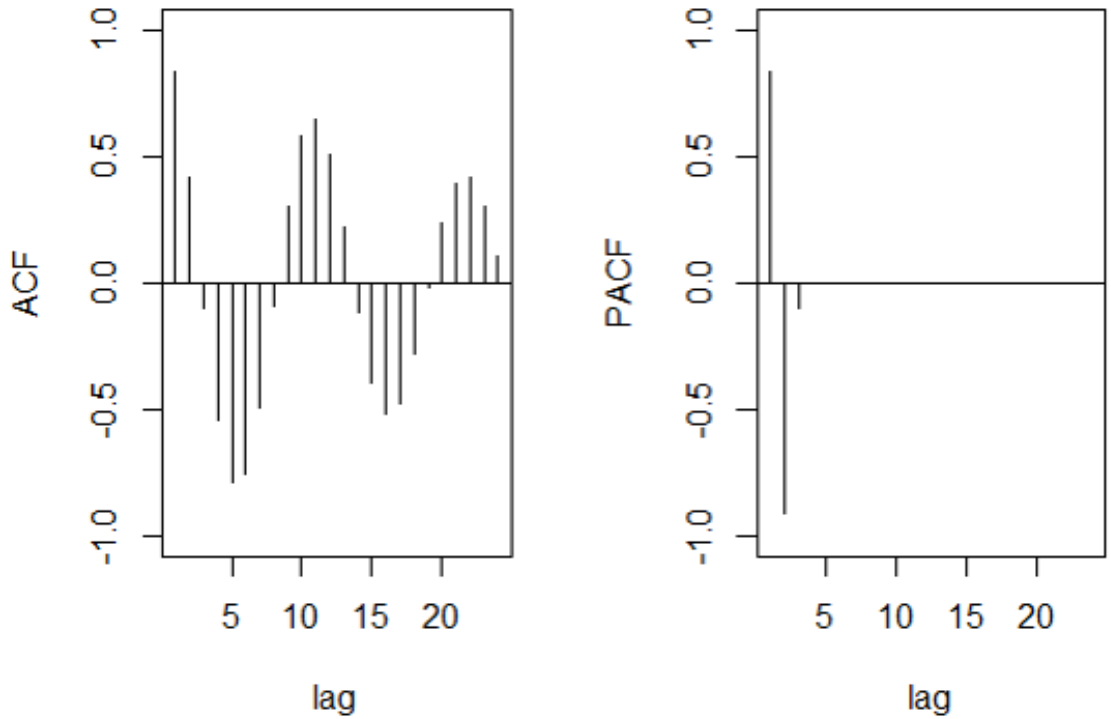
$$\phi^p(B)x_t = \theta^q(B)w_t$$

Model	**Concise form**
AR(p)	$\phi^p(B)x_t = w_t$
MA(q)	$x_t = \theta^q(B)w_t$
ARMA(p, q)	$\phi^p(B)x_t = \theta^q(B)w_t$
ARIMA(p, d, q)	$\phi^p(B)(1-B)^d x_t = \theta^q(B)w_t$
ARMA(P, Q)$_s$	$\Phi^P(B^s)x_t = \Theta^Q(s)w_t$
ARIMA(P, D, Q)$_s$	$\Phi^P(B^s)(1-B^s)^D x_t = \Theta^Q(B^s)w_t$
ARMA(p, q) $\times$ (P, Q)$_s$	$\Phi^P(B^s)\phi^p(B)x_t = \Theta^Q(B^s)\theta^q(B)w_t$
ARIMA(p, d, q) $\times$ (P, D, Q)$_s$	$\Phi^P(B^s)\phi^p(B)(1-B^s)^D(1-B)^d x_t = \Theta^Q(B^s)\theta^q(B)w_t$

* The notation used in this slide deviates from the notation used in the course literature so far.

PACF for AR(p)

- Example: AR(3) $\phi_1 = 1.5$, $\phi_2 = -0.75, \phi_3 = -0.1$



Seasonal?

ACF and PACF

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

How to differentiate between ARMA(p,q)?

Empirical ACF (EACF)

Idea:

- ARMA(p,q): $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + \sum_{j=1}^{q} \theta_j w_{t-j} + w_t$
- If we can estimate $\phi_j \rightarrow x'_t = x_t - \sum_{j=1}^{p} \phi_j x_{t-j}$ is linear function in $w_t, ..., w_{t-q}$
- If we run regression x'_t against $w_t ... w_{t-j}$:
  - Residuals are white noise, $j \geq q \rightarrow$ ACFs not significant
    - ⋆ Some of the coefficients will be 0
  - Residuals are not white noise, $j < q \rightarrow$ ACFs significant
  - Note: $w_t s$ substituted by lagged residuals from a series of regressions
- If $x'_t = x_t - \sum_{j=1}^{k} \phi_j x_{t-j}, k < p \rightarrow$ white noise will never be achieved $\rightarrow$ ACFs are not zero

Empirical ACF (EACF)

- $k > p$ General result: ACFs are 0 for $j > q + (k - p)$
  - Example: ARMA(0,1)
- General conclusion for AR,MA =(k,j):
  - This is theoretical one! $\rightarrow$ not exactly the same for the samples

AR/MA	0	1	2	...	...	...	...
0	X	X	X	X	X	X	X
1	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X
...	X	X	X	X	X	X	X
...	X	X	X	X	X	X	X
...	X	X	0	0	0	0	0
...	X	X	X	0	0	0	0
...	X	X	X	X	0	0	0
...	X	X	X	X	X	0	0

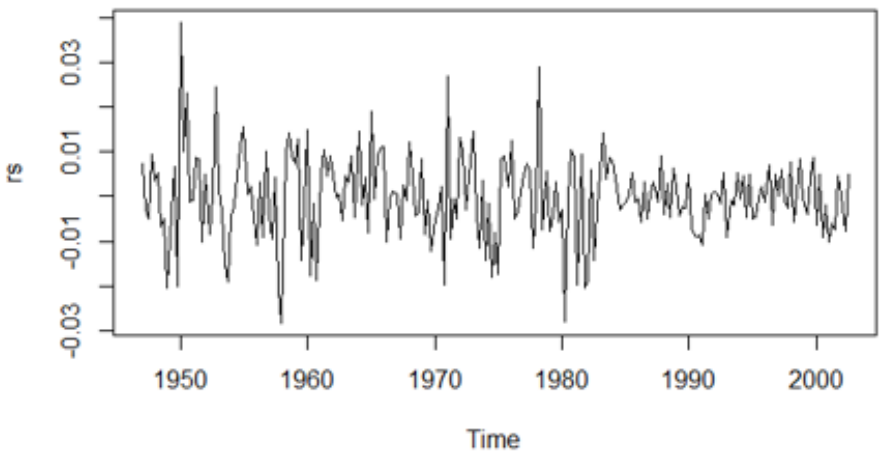
Residual analysis

- Residuals $r_t = x_t - \hat{x}_t^{t-1}$?they are innovations
  - Note: computed from one-step-ahead predictions!
  - Measures predictive quality of the model (compare OLS)

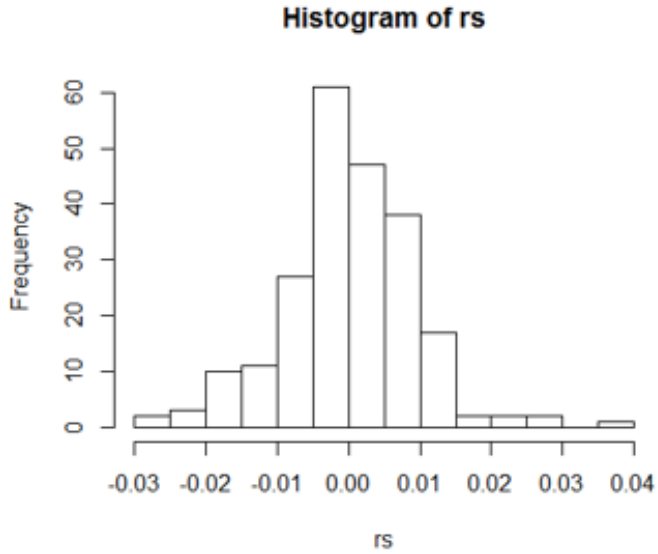
- Residual analysis
  - Visual inspection: stationary? Patterns?
  - Histograms, Q-Q plots
  - ACF, PACF
  - Runs test
  - Box-Ljung test

Residual analysis - Visual inspection

Histogram and visual inspection

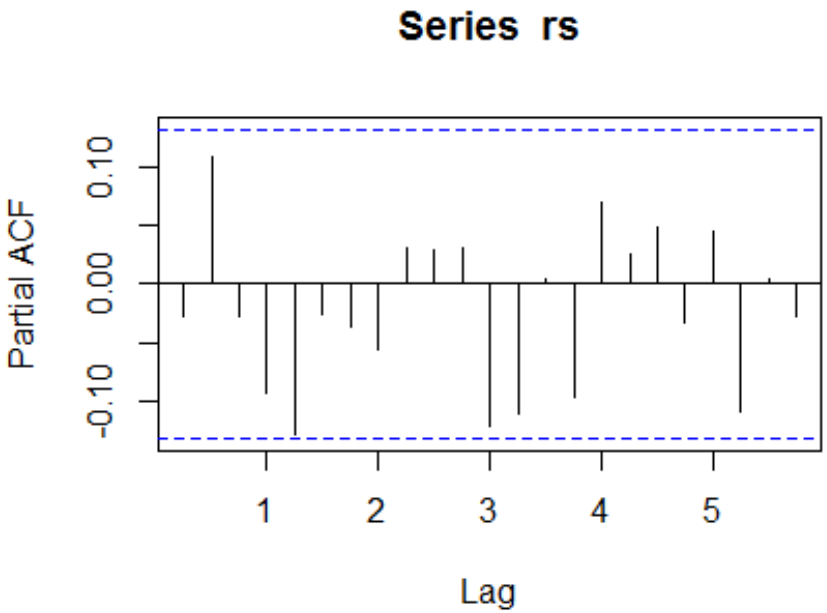


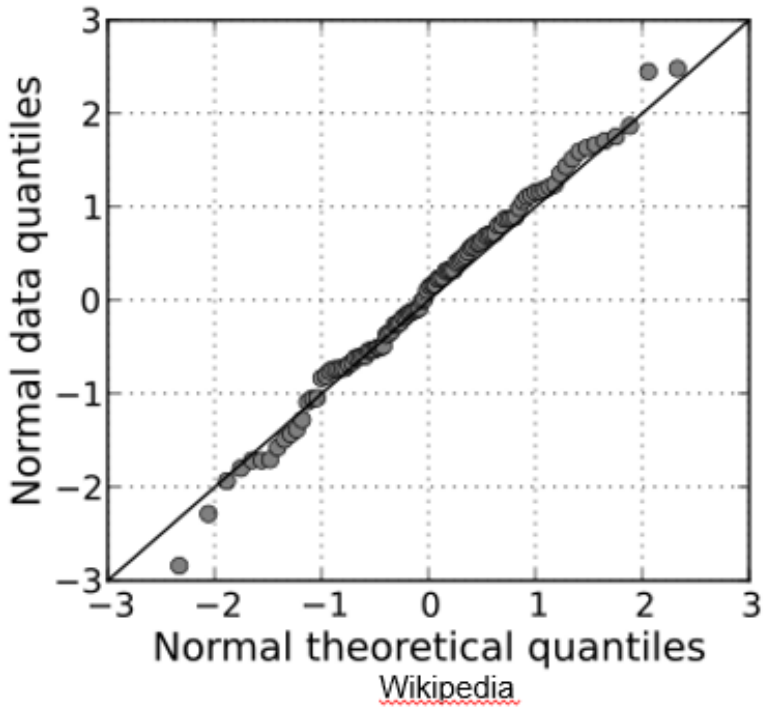
If looks white is good



If looks Normal is good

Residual analysis - ACF /PACF Q-Q plots





If between the blue lines good

If along the diagonal line GOOD

Statistical tests

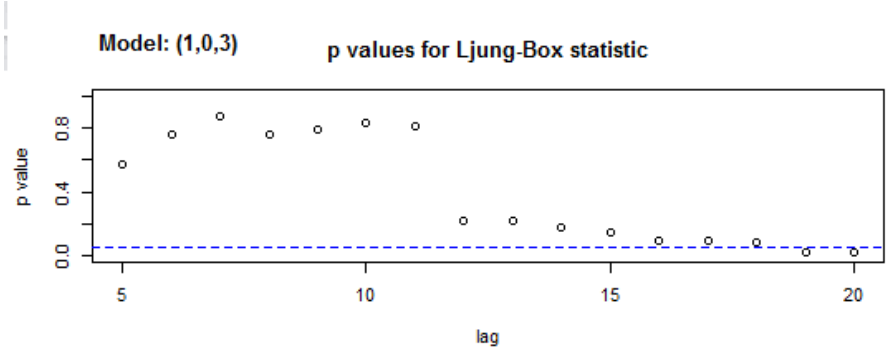
Tests are used to test independence

Runs test

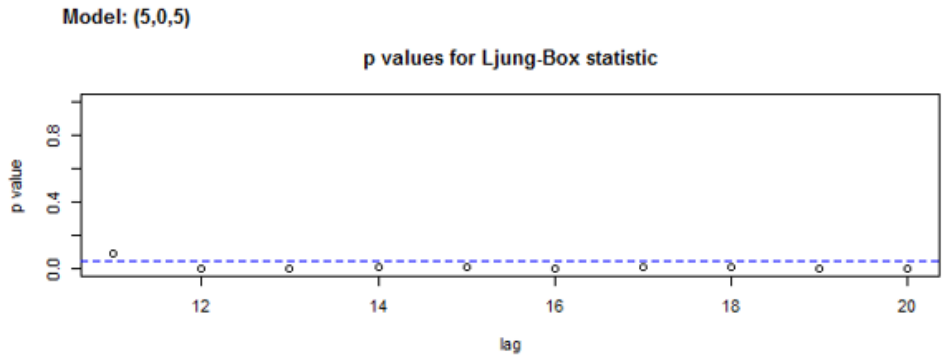
- H_0 : x_t values are i.i.d. **p-value NOT small**
- H_a : x_t values are not i.i.d. **p-value small**

Box-Ljung test

- H_0 : data are independent **p-value NOT small**
- H_a : data are not independent **p-value small**



GOOD



BAD

SARIMA

- Multiplicative seasonal autoregressive integrated moving average model $ARIMA(p, d, q)\ x\ (P, D, Q)_s$

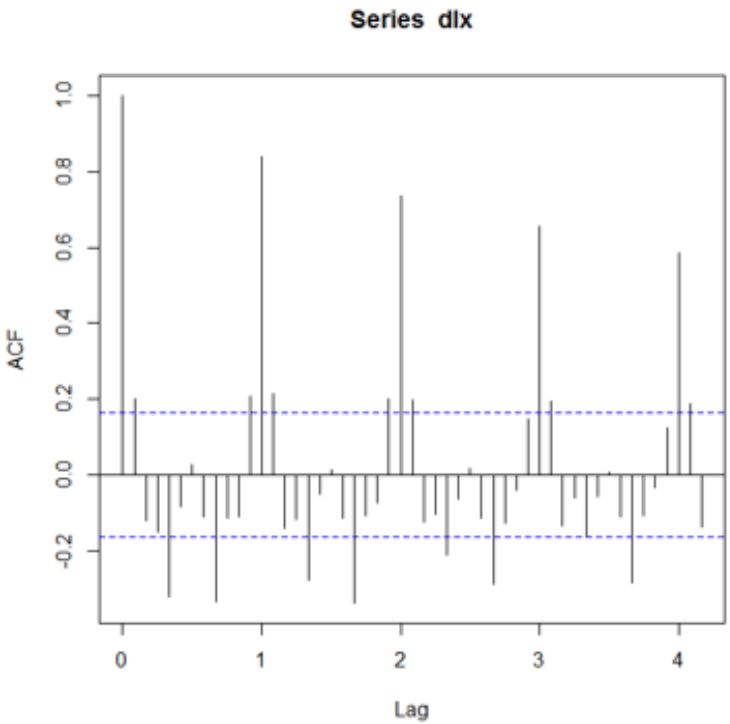
$$\Phi_p(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t$$

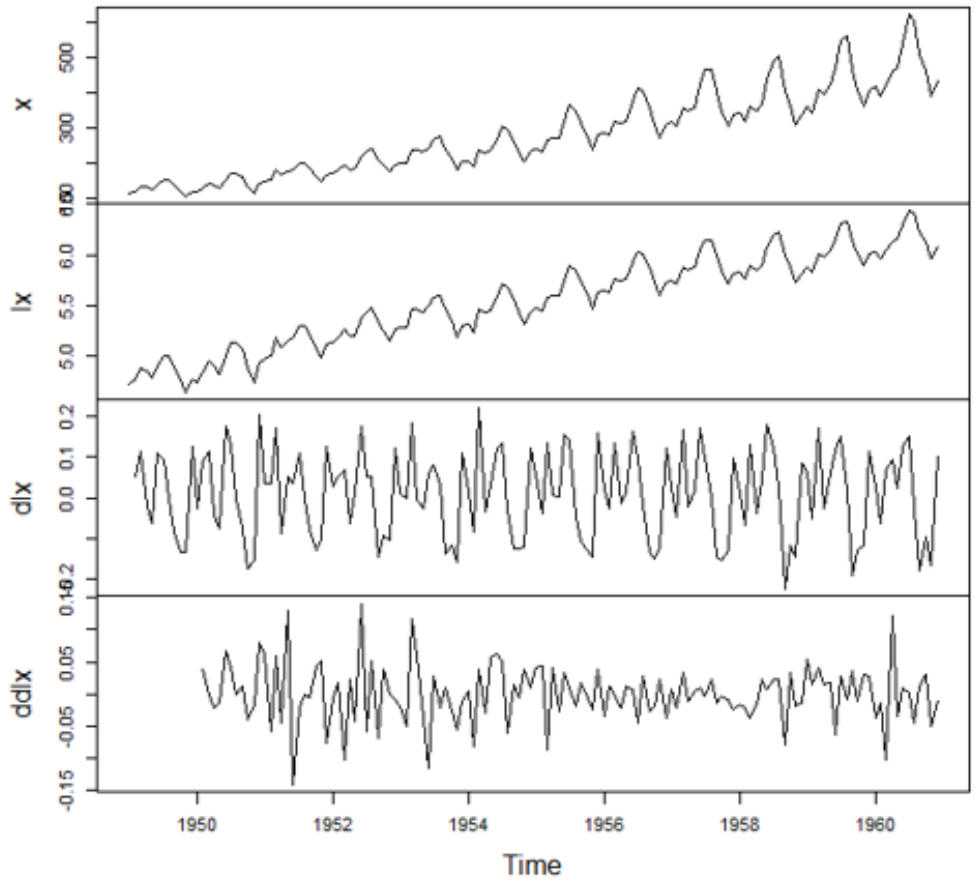
$$\nabla_s^D = (1 - B^s)^D$$

- How to identify SARIMA?
  1. Perform differencing first (trend)
  2. Investigate ACF $\rightarrow$ slowly decays at peaks?
     1. Yes $\rightarrow$ Additional differencing by ∇_s^D
  3. Model non-seasonal part
  4. Model seasonal part (check peaks), check ACF and PACF of residuals

SARIMA

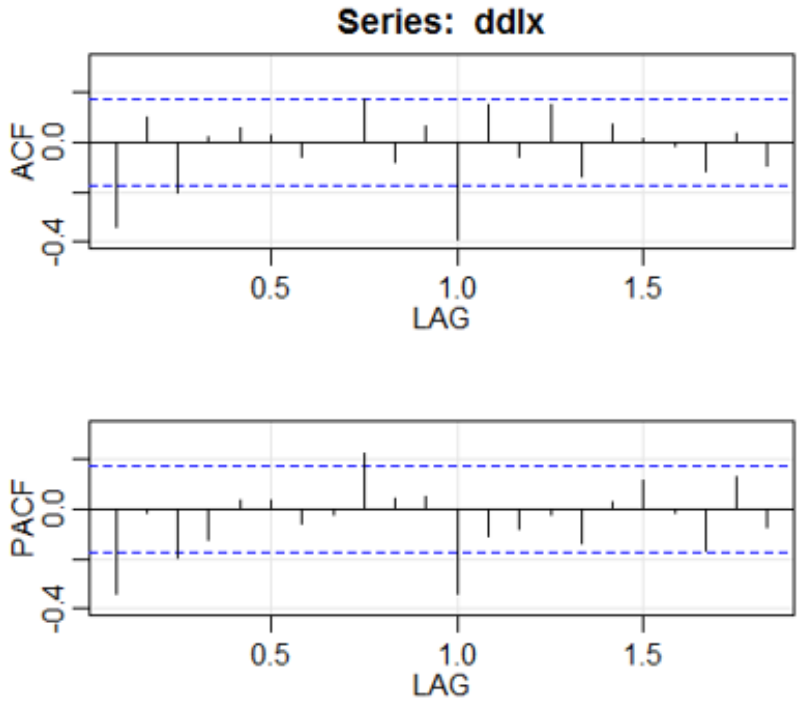
- Example: Air passangers





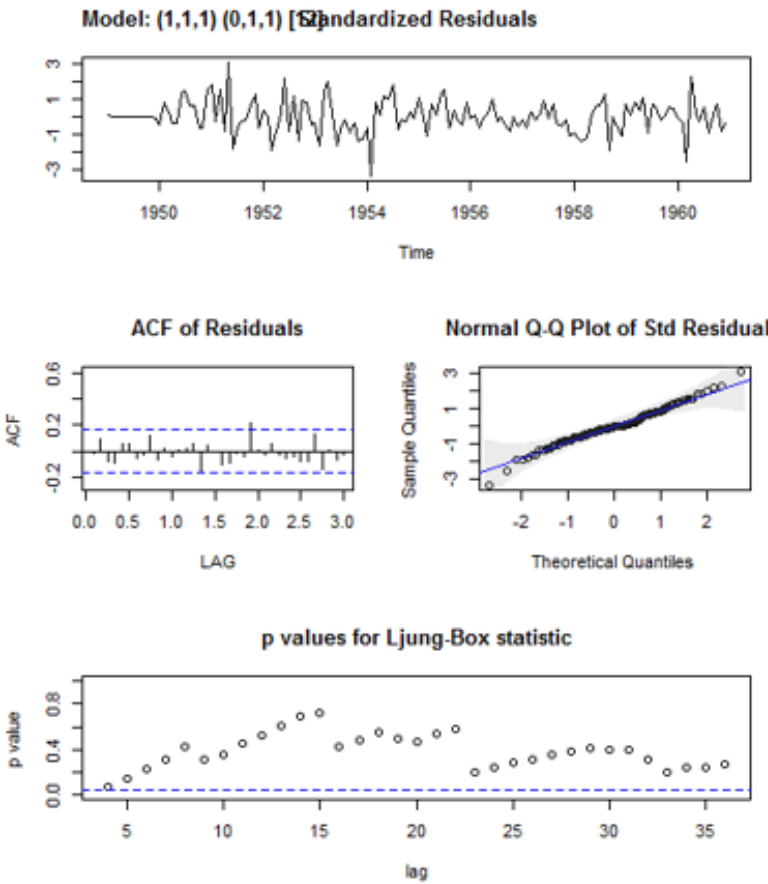
SARIMA

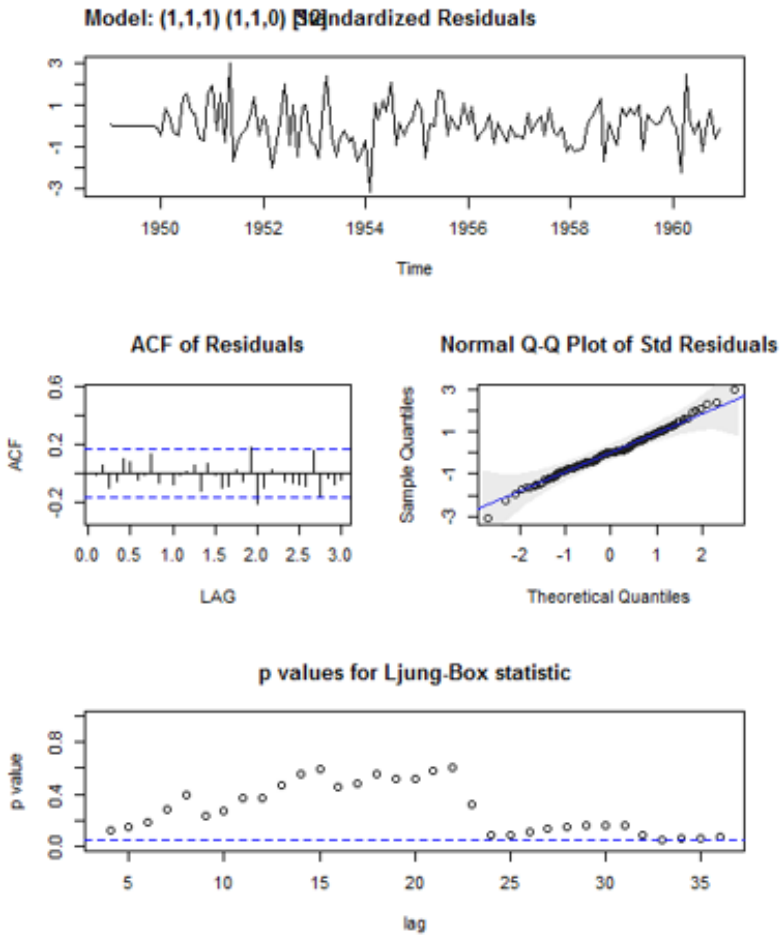
- Example: Air passangers



$(0,1,1)_{12}$ or $(1,1,0)_{12}$

SARIMA


Model: (1,1,1) (0,1,1) [12] Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung-Box statistic
p value
lag


Model: (1,1,1) (1,1,0) [12] Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung-Box statistic
p value
lag

SARIMA

- Remove AR term!

Is one model much better the other one?

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

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

```
> m1$fit

Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
    Q), period = S), include.mean = !no.constant, optim.control = list(trace = trc,
    REPORT = 1, reltol = tol))

Coefficients:
          ar1      ma1     sar1
      0.0547  -0.4886  -0.4731
s.e.  0.2161   0.1933   0.0800

sigma^2 estimated as 0.001425:  log likelihood = 241.73,  aic = -475.47
> m2$fit

Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
    Q), period = S), include.mean = !no.constant, optim.control = list(trace = trc,
    REPORT = 1, reltol = tol))

Coefficients:
          ar1      ma1     sma1
      0.1960  -0.5784  -0.5643
s.e.  0.2475   0.2132   0.0747

sigma^2 estimated as 0.001341:  log likelihood = 244.95,  aic = -481.9
```

State space modelling

- State space models
  - Linear and Gaussian state space models
  - Kalman filtering, Kalman smoothing and Forecasting
  - Maximum likelihood estimate of the state space models
  - Stochastic volatility

Vectorized AR(2) model **Whiteboard**

Consider an AR(2) model

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

Let $\mathbf{z}_t = \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}$ and $e_t = \begin{bmatrix} w_t \\ 0 \end{bmatrix}$.

Show that we rewrite the AR(2) model in the state space form:

$$\mathbf{z}_t = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \mathbf{z}_{t-1} + e_t$$
$$x_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{z}_t,$$

ARIMA models in State Space form **Whiteboard**

$$\phi^p(B)x_t = \theta^q(B)w_t$$

Can we rewrite any model of this form as a state space model?

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + e_t,$$
$$\mathbf{x}_t = C\mathbf{z}_t + \nu_t,$$

ARIMA models in State Space form **Whiteboard**

$$\phi^p(B)x_t = \theta^q(B)w_t$$

Outline of the solution:

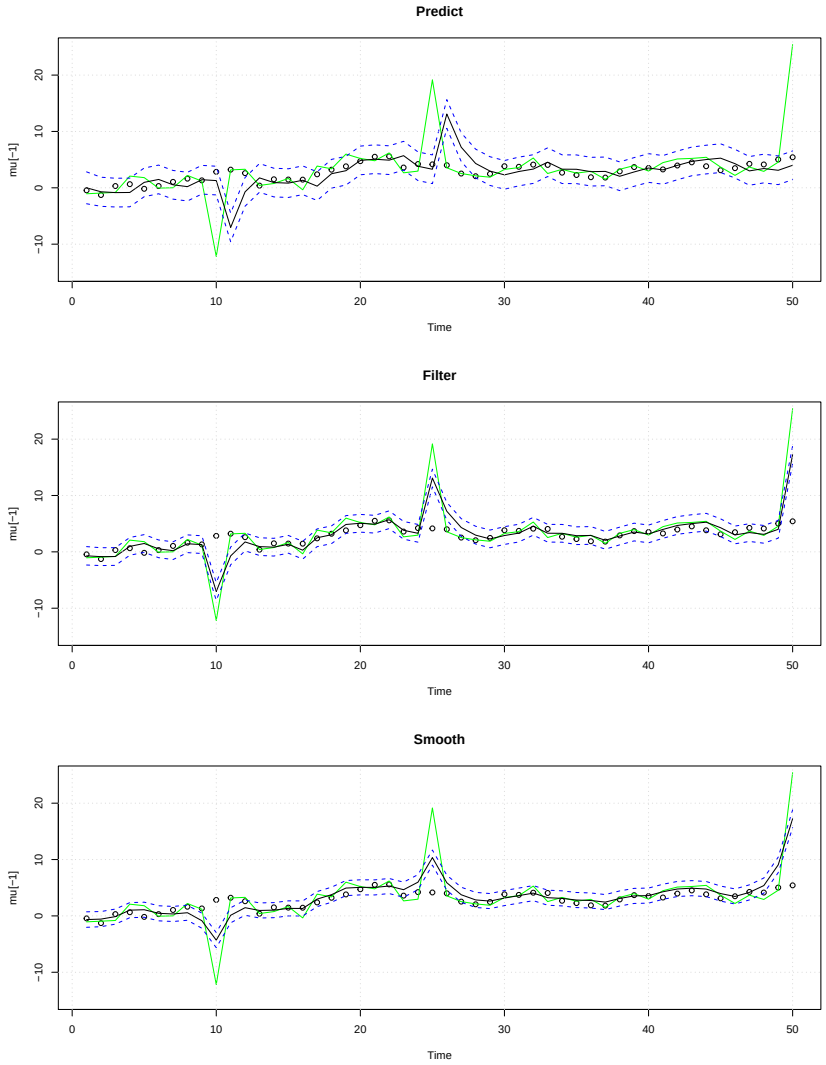
Let $r = \max(p, q+1)$,
$\phi^r(B) = 1 - \phi_1 B - \cdots - \phi_r B^r$,
$\theta^r(B) = 1 + \theta_1 B - \cdots - \theta_{r-1} B^{r-1}$,
$\phi^r(B)(\theta^r(B))^{-1} x_t = w_t$. Hence, for $z_t = (\theta^r(B))^{-1} x_t$ we can have
$\phi^r(B) z_t = w_t$

$$\mathbf{z}_t = \begin{bmatrix} z_t \\ z_{t-1} \\ z_{t-2} \\ \vdots \\ z_{t-r+1} \end{bmatrix} \text{ and } \mathbf{z}_t = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_r \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \mathbf{z}_{t-1} + \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$x_t = \begin{bmatrix} 1 & \theta_1 & \theta_2 & \cdots & \theta_r \end{bmatrix} \mathbf{z}_t$$

Robustness to outliers:filter versus smoother

Live example in Rstudio



Stochastic Volatility : Gaussian sum filter

The problem is finding the filtering distribution of $\mathbf{z}_t|\mathbf{x}_{1:t}$ when

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + w_t$$
$$\mathbf{x}_t = C\mathbf{z}_t + \eta_t$$

and

$$w_t \sim iidN(0, Q)$$
$$\eta_t \sim \pi_0 N(\mu_0, R_1) + \pi_1 N(\mu_1, R_2)$$

where $\pi_0 + \pi_1 = 1$

Examination

- Most of the examination will be your Computer labs and assignments from the teaching sessions with a twist.
- You need to have a deep knowledge of the subjects covered in the lectures to get a B+ score.
- Study them over and over and make sure you have the correct solutions with you on the examination day.

1 Lectures 1-3

- Probability density function for x: $f(x)$
- Marginal density
  $f_i(x_i) = \int f(x)dx_1...dx_{i-1}dx_{i+1}...dx_p$
- Expected (mean) value $Ex = \int xf(x)dx$
- Covariance $\text{cov}(x,y) = E\{(x-Ex)(y-Ey)\}$
- Correlation $\rho_{x,y} = \text{corr}(x,y) = \frac{\text{cov}(x,y)}{\sigma_x\sigma_y}$
- Variance $\text{var}(x) = E\{(x-Ex)^2\} = \text{cov}(x,x)$
- Relationships (a is a constant)
  - $E(x+a) = Ex + a$, $E(ax) = aEx$
  - $E(x+y) = Ex + Ey$
  - $\text{cov}(x+a,y) = \text{cov}(x,y)$
  - $\text{cov}(x+z,y) = \text{cov}(x,y) + \text{cov}(z,y)$
  - $\text{var}(ax) = a^2\,\text{var}(x)$

uncorrelated $\iff$ $E(XY) = EX.EY$
independent $\iff$ $f_{X,Y}(x,y) = f_X(x).f_Y(y)$

- Autocovariance function

$$\gamma(s,t) = \text{cov}(x_s,x_t) = E[(x_s-\mu_s)(x_t-\mu_t)]$$

Note: $\text{var}(x_t) = \gamma(t,t)$

- Autocorrelation function (ACF)

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

Useful fact: If $U = \sum_{j=1}^{m} a_j x_j$ and $V = \sum_{k=1}^{r} b_k y_k$

$\text{cov}(U,V) = \sum_{j=1}^{m}\sum_{k=1}^{r} a_j b_k \,\text{cov}(x_j,y_k)$

1.1 stationarity

- Time series x_t is weakly stationary (stationary) if
  - $Ex_t = const$
  - $\gamma(s,t) = \gamma(|s-t|)$
  - $\text{var}(x_t) < \infty$
- $\gamma(t,t+h) = \gamma(|t+h-t|) = \gamma(h)$
  - Autocovariance depends on lag only!
- Autocovariance for stationary process
  $\gamma(h) = \text{cov}(x_t, x_{t+h})$
- ACF for stationary process $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$

Properties of stationary process:

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

$$|\gamma(h)| \le \gamma(0) \qquad \rho(h) \le 1, \rho(0) = 1$$

If x_t is stationary,

- Sample mean

$$Ex \approx \bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

- Sample autocovariance function

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

Theorem: Under weak conditions, if x_t is white noise and $n \to \infty$ then $\hat{\rho}(h)$ is approximately $N(0, \frac{1}{n})$

Consequence: If some $|\hat{\rho}(h)| > \frac{2}{\sqrt{n}}$ then the time series is not a white noise (with approximately 95 % confidence).

1.2 Backshift operator

- Backshift operator $Bx_t = x_{t-1}$,
  Powers $B^k x_t = x_{t-k}$
- Forward-shift operator $B^{-1}x_t = x_{t+1}$
- Note $BB^{-1}x_t = x_t$ (i.e. $BB^{-1} = 1$)
- Differencing $\nabla x_t = (1-B)x_t$
- Differences of order d: $\nabla^d = (1-B)^d$
- Property: Operators can be manipulated as polynomials
- Example Check that $\nabla^2 x_t = x_t - 2x_{t-1} + x_{t-2}$
- Property: Differencing of order p can remove polynomial trend of order p

1.3 MA, AR, ARMA

- **Moving average model of order q, MA(q)**

$$\begin{aligned} x_t &= w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} \\ &= \sum_{j=0}^{q} \theta_j w_{t-j} \end{aligned}$$

  - $w_t \sim wn(0, \sigma_w^2)$
  - $\theta_1, ...\theta_q$ constants, $\theta_q \neq 0$ and $\theta_0 = 1$

- **Moving average operator**

$$\theta(B) = \sum_{j=0}^{q} \theta_j B^j$$

- MA(q):

$$\boxed{x_t = \theta(B) w_t}$$

- **Autoregressive model of order p, $AR(p)$**

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t$$

  - x_t is stationary if x_0 is sampled from the stationary distribution
  - $w_t \sim wn(0, \sigma_w^2)$
  - $\phi_1, ...\phi_p$ constants, $\phi_p \neq 0$
  - $Ex_t = 0$

- **Autoregressive operator**

$$\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p$$

- AR(p) model

$$\boxed{\phi(B) x_t = w_t}$$

- **ARMA(p,q)**

$$\begin{aligned} x_t =& \phi_1 x_{t-1} + ... + \phi_p x_{t-p} \\ &+ w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} \end{aligned}$$

  - $\phi_p \neq 0, \theta_q \neq 0$
  - Is stationary
  - $Ex_t = 0$

- p-autoregressive order, q-moving average order
- Alternative form

$$\boxed{\phi(B) x_t = \theta(B) w_t}$$

- Note: $x_t = \phi^{-1}(B)\theta(B) w_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}$
  - But series might be non-convergent

1.4 Causality / invertiblity

A stationary process is **causal** if it is only dependent on the past values of the process

Def: A linear process is nonexplosive and causal if it can be written as a one-sided sum:

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

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

Def: An MA process is **invertible** if it has a causal AR representation,

$$w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}$$

Def: Linear process is **causal** and **nonexplosive** if

- $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$ (depends on the past only)
- $\sum_{j=0}^{\infty} |\psi_j| < \infty$
- We set $\psi_0 = 1$ by convention.

Property: ARMA(p,q) is **causal** iff roots $\phi(z') = 0$ are outside unit circle, i.e. $|z'| > 1$

$$\boxed{\phi(B) x_t = \theta(B) w_t}$$

Def: ARMA(p,q) is **invertible** if

- $w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}$ (depends on the past only)
- $\sum_{j=0}^{\infty} |\pi_j| < \infty$

Property: ARMA(p,q) is **invertible** iff roots $\theta(z') = 0$ are outside unit circle, i.e. $|z'| > 1$

$$\boxed{\phi(B) x_t = \theta(B) w_t}$$