

Time Series Analysis

Lasse Engbo Christiansen

DTU Applied Mathematics and Computer Science
Technical University of Denmark

September 8, 2017

Outline of the lecture

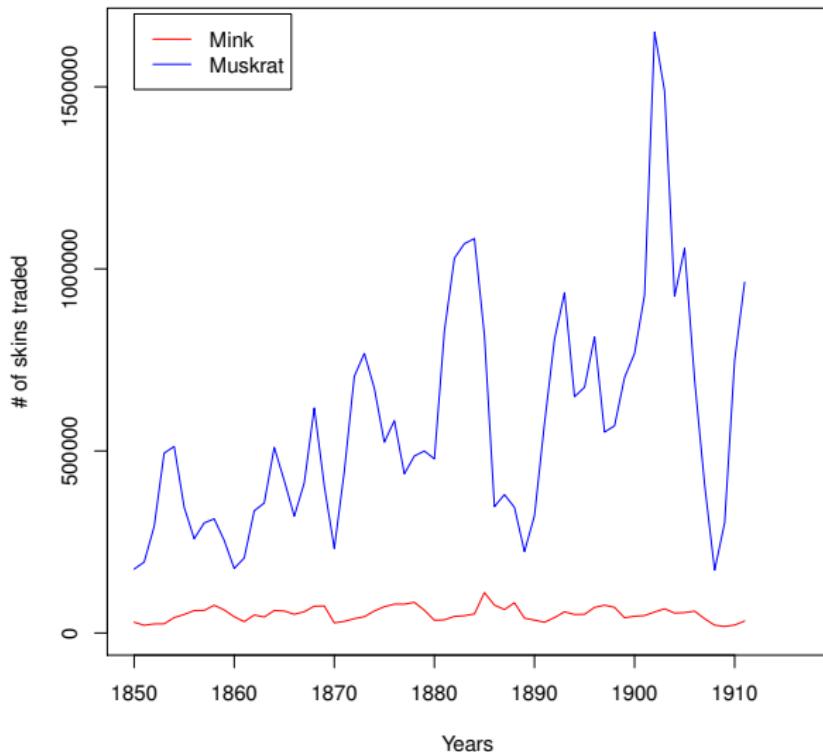
- ▶ Practical information
- ▶ Introductory examples (see also Chapter 1)
- ▶ A brief outline of the course
- ▶ Chapter 2:
 - ▶ Multivariate random variables
 - ▶ The multivariate normal distribution
 - ▶ Linear projections
- ▶ Example

Contact info

- ▶ Lasse Engbo Christiansen
Technical University of Denmark
Applied Mathematics and Computer Science
Section for Dynamical Systems
From October 1st: Building 303B, room 010
Email laec@dtu.dk
- ▶ Teaching assistants: Jesper, Sebastian & me.
For consultation besides that – please drop me an email

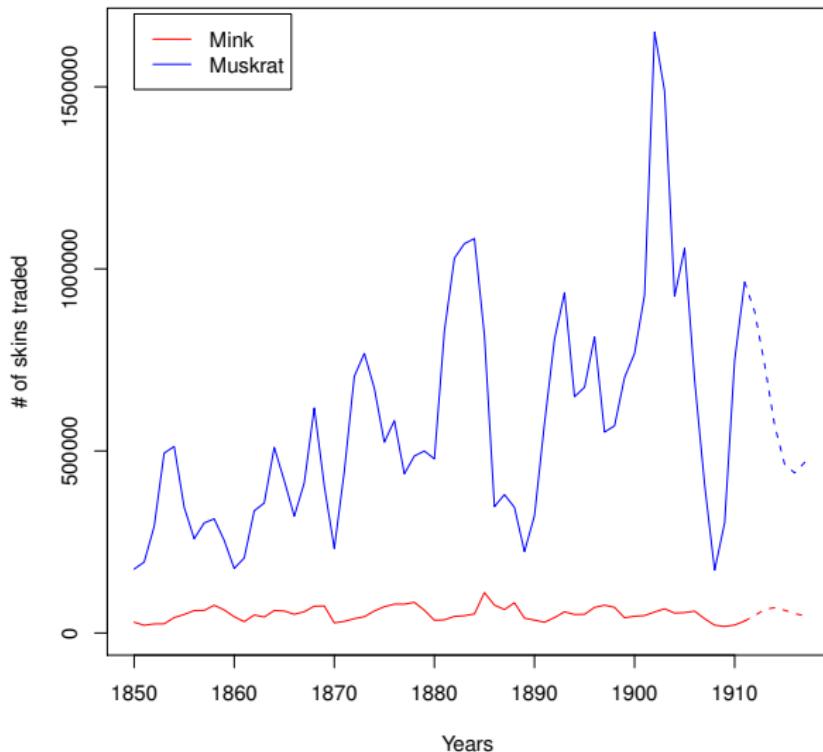
What you should be able to do

Mink and Muskrat skins traded
in Canada 1850–1911

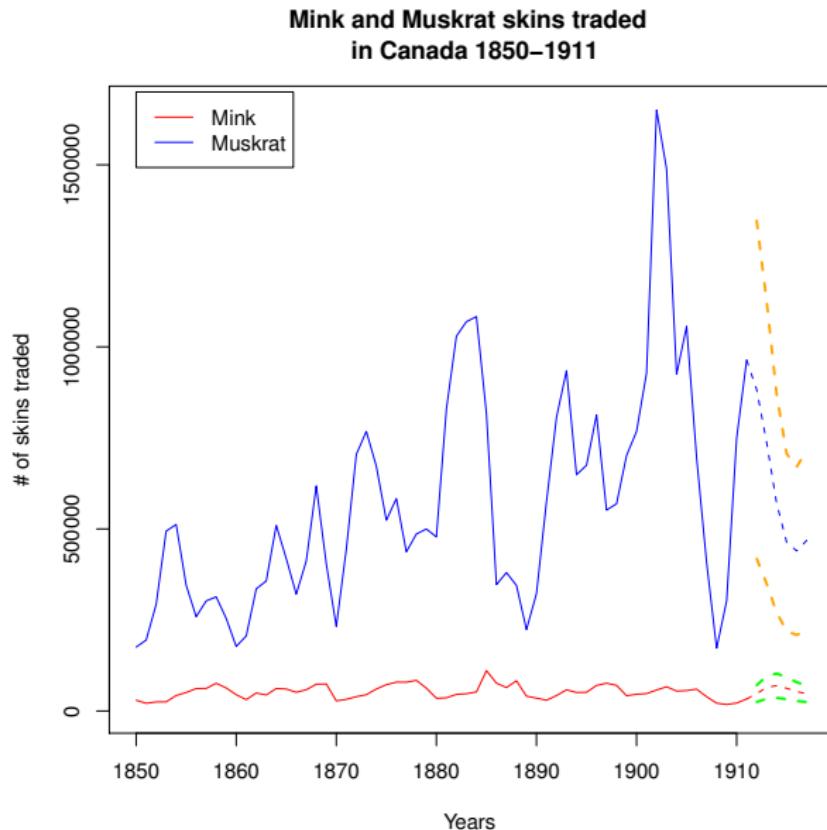


What you should be able to do

Mink and Muskrat skins traded
in Canada 1850–1911



What you should be able to do



Introductory example – shares (COLO B 1 month)



From finance.yahoo.com

Introductory example – shares (COLO B 1 year)

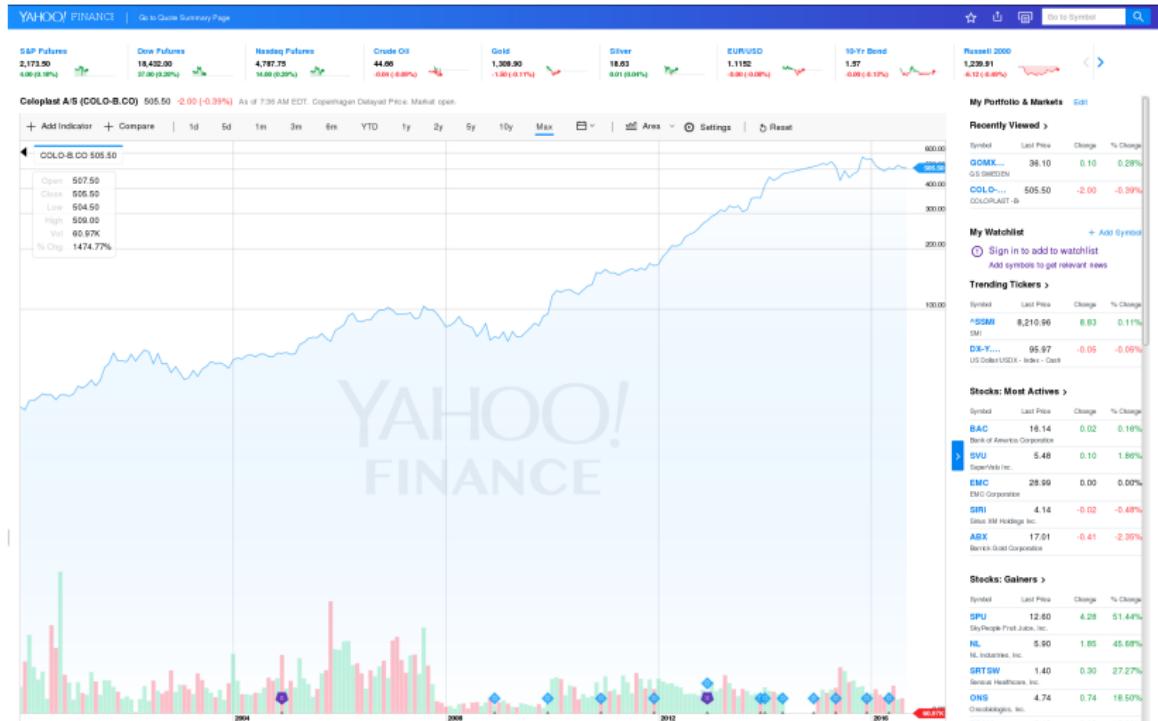


Introductory example – shares (COLO B all)



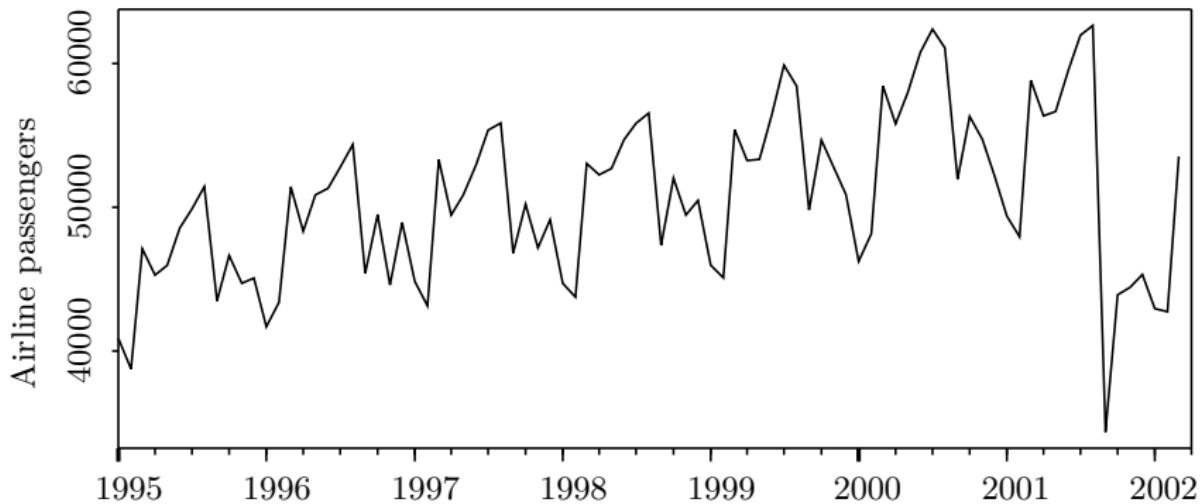
From finance.yahoo.com

Introductory example – shares (COLO B log(all))

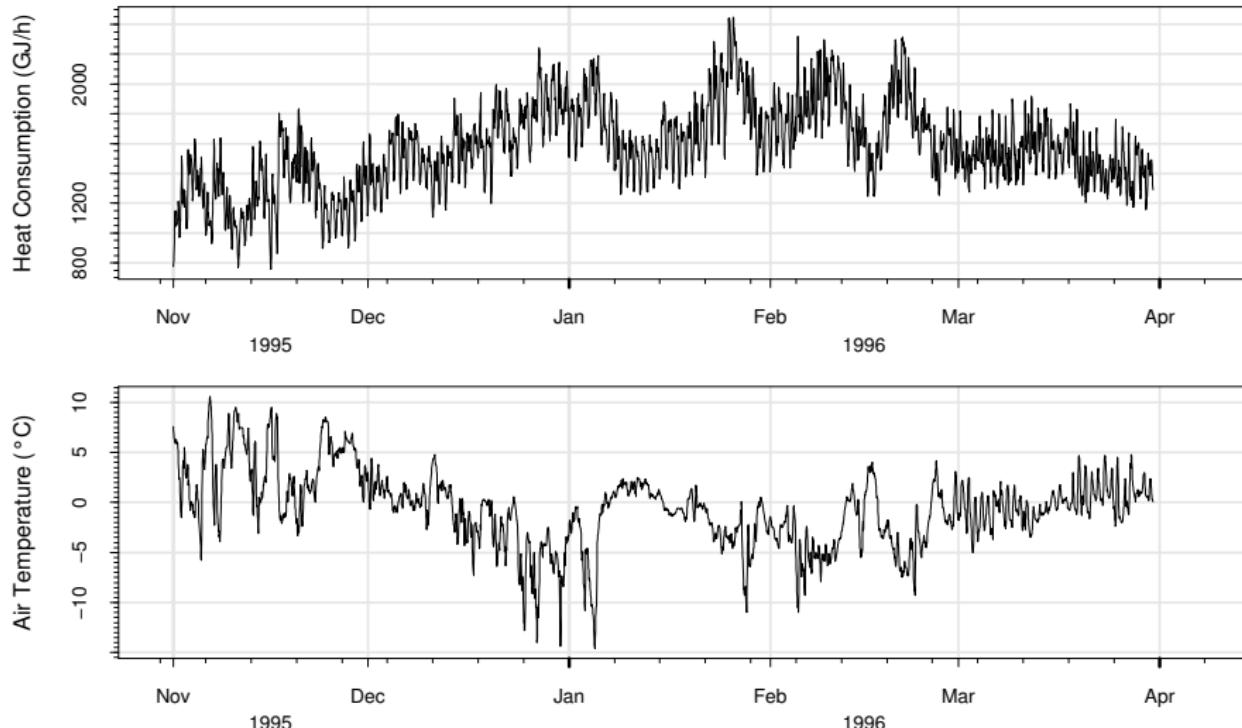


From finance.yahoo.com

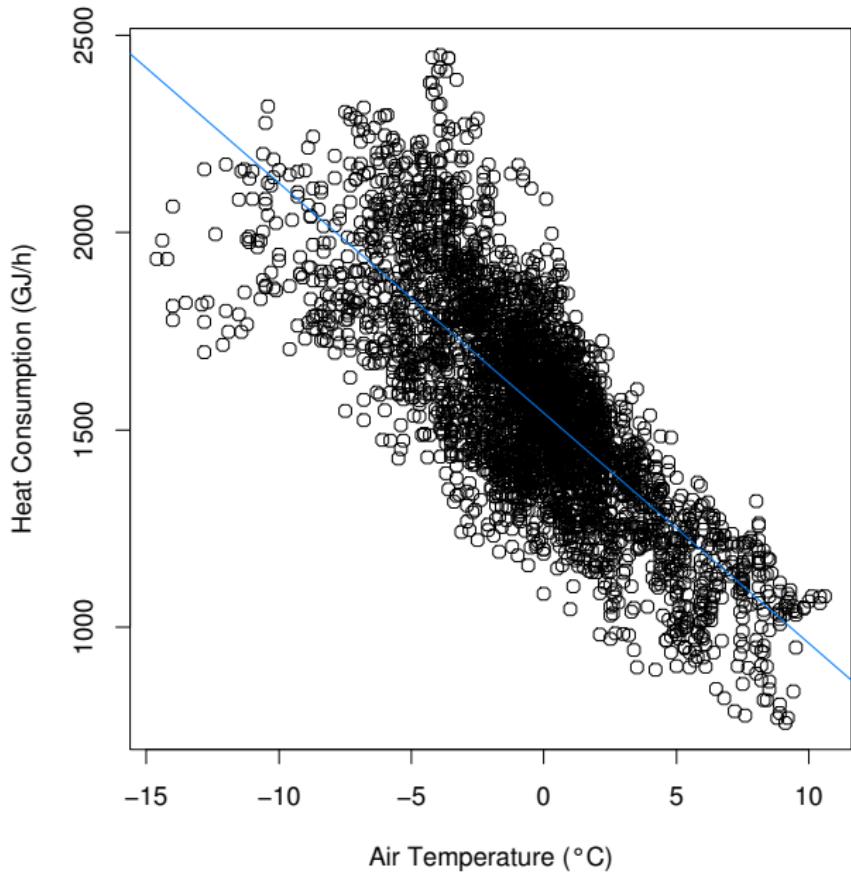
Number of Monthly Airline Passengers in the US



Consumption of District Heating (VEKS) – data

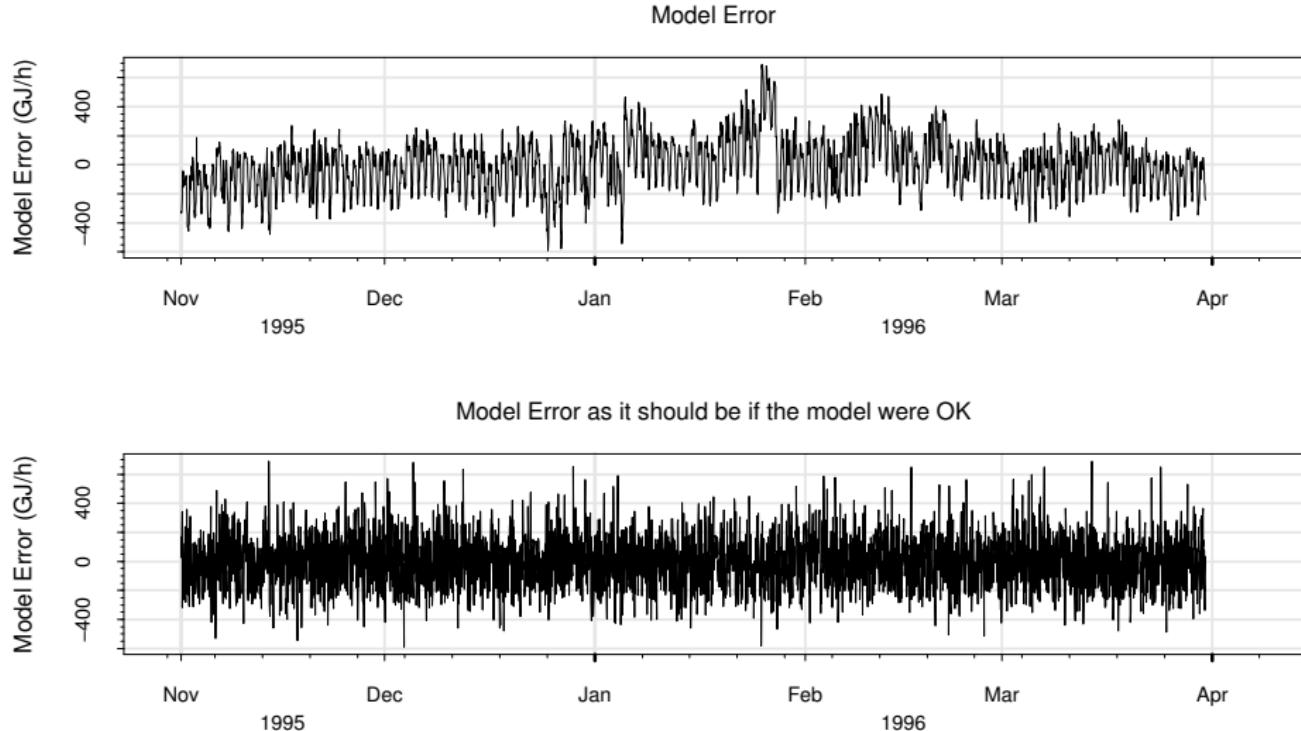


Consumption of DH – simple model



Discussion: What is a dynamical system?

Consumption of DH – model error



A brief outline of the course

- ▶ General aspects of multivariate random variables
- ▶ Prediction using the general linear model
- ▶ Time series models
- ▶ Some theory on linear systems
- ▶ Time series models with external input

Some goals:

- ▶ Characterization of time series / signals; correlation functions, covariance functions, stationarity, linearity, . . .
- ▶ Signal processing; filtering and smoothing
- ▶ Modelling; with or without external input
- ▶ Prediction with uncertainty

Today: Multivariate random variables

- ▶ Distribution functions
- ▶ Density functions
- ▶ The multivariate normal distribution
- ▶ Marginal densities
- ▶ Conditional distributions and independence
- ▶ Expectations and moments
- ▶ Moments of multivariate random variables
- ▶ Conditional expectation
- ▶ Distributions derived from the normal distribution
- ▶ Linear projections and relations to conditional means

Multivariate random variables – distr. functions

- ▶ Definition (n -dimensional random variable; random vector)

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

- ▶ Joint distribution function:

$$F(x_1, \dots, x_n) = P\{X_1 \leq x_1, \dots, X_n \leq x_n\}$$

- ▶ Notice notation (lowercase, capital letters, bold font)

Multivariate random variables - joint densities

- ▶ Joint distribution function (repeated from last slide):

$$F(x_1, \dots, x_n) = \mathbb{P}\{X_1 \leq x_1, \dots, X_n \leq x_n\}$$

- ▶ Joint density function - continuous case:

$$f(x_1, \dots, x_n) = \frac{\partial^n F(x_1, \dots, x_n)}{\partial x_1 \dots \partial x_n}$$

- ▶ and back to the joint distribution function:

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f(t_1, \dots, t_n) dt_1 \dots dt_n$$

- ▶ Joint density function - discrete case:

$$f(x_1, \dots, x_n) = \mathbb{P}\{X_1 = x_1, \dots, X_n = x_n\}$$

The Multivariate Normal Distribution

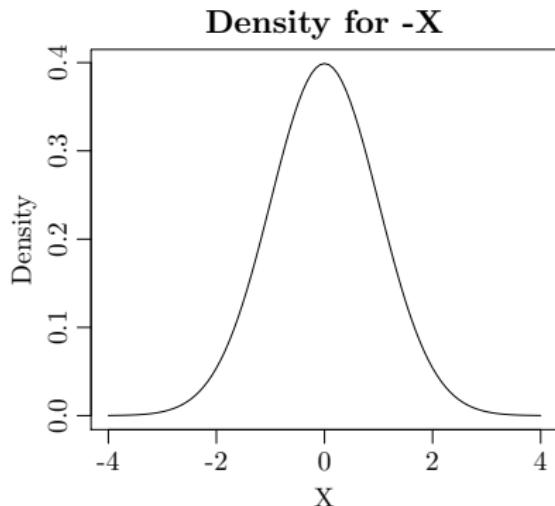
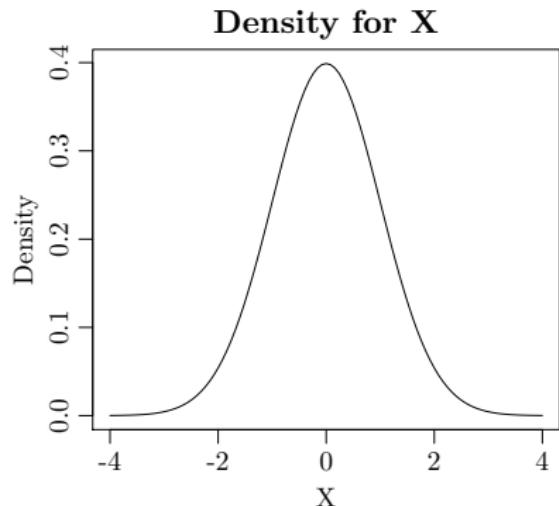
- ▶ The joint p.d.f.

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp \left[-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right]$$

- ▶ Σ is symmetric and positive semi-definite
- ▶ Notation: $X \sim N(\mu, \Sigma)$
- ▶ Standard multivariate normal: $Z \sim N(\mathbf{0}, I)$
- ▶ If $X = \mu + TZ$, where $\Sigma = TT^T$, then $X \sim N(\mu, \Sigma)$
- ▶ If $X \sim N(\mu, \Sigma)$ and $Y = a + BX$ then $Y \sim N(a + B\mu, B\Sigma B^T)$
- ▶ More relations between distributions in Sec. 2.7

Stochastic variables and distributions

- If $X \sim N(0, 1)$, then $-X \sim N(0, 1)$

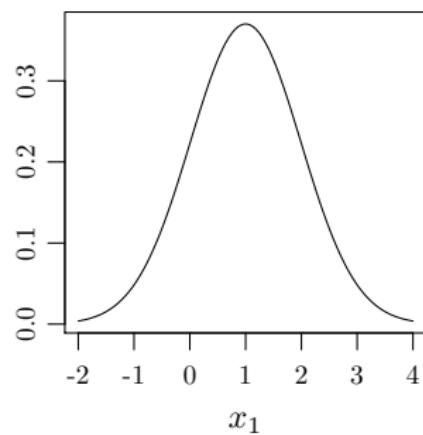
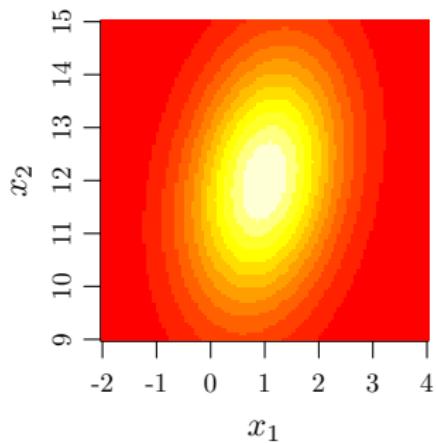
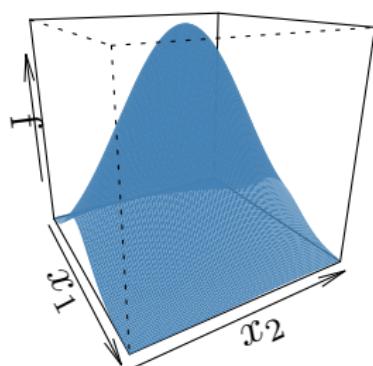


- $X, -X$ are different variables that have the same distribution

Marginal density function

- ▶ Sub-vector: $(X_1, \dots, X_k)^T$, $(k < n)$
- ▶ Marginal density function:

$$f_S(x_1, \dots, x_k) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_{k+1} \dots dx_n$$



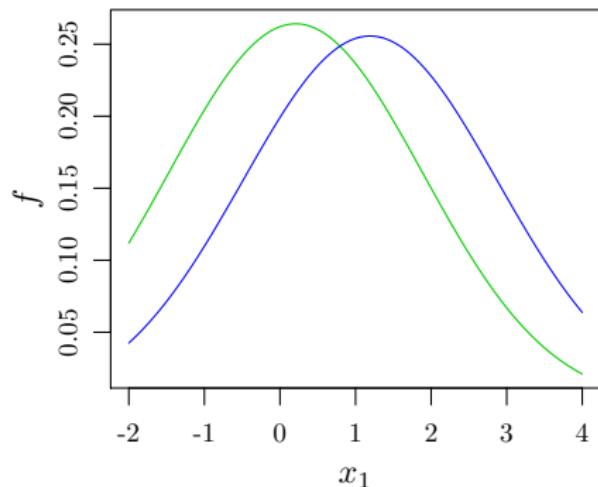
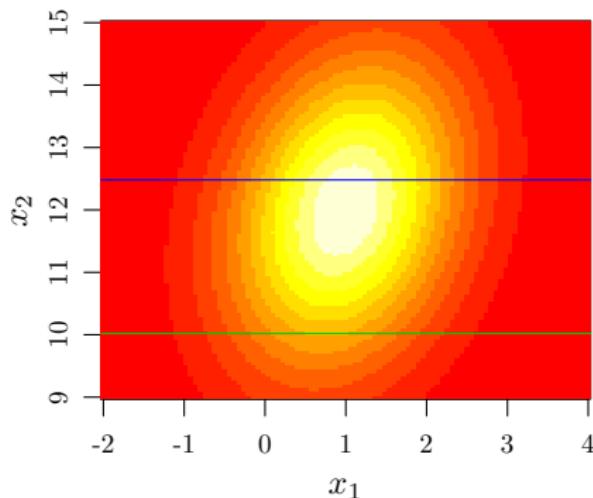
Blackboard - conditional probability

Conditional distributions

- The conditional density of X_1 given $X_2 = x_2$ is defined as ($f_{X_1}(x_1) > 0$):

(joint density of (X_1, X_2) divided by the marginal density of X_2 evaluated at x_2)

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$



Independence

- ▶ If knowledge of X does not give information about Y , we get that $f_{Y|X=x}(y) = f_Y(y)$
- ▶ This leads to the following definition of independence:

X, Y stochastically independent $\stackrel{\text{def}}{\Leftrightarrow}$

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

Expectation

- ▶ Let X be a univariate random variable with density $f_X(x)$. The expectation of X is then defined as:

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \quad (\text{continuous case})$$

$$E[X] = \sum_{\text{all } x} x P(X = x) \quad (\text{discrete case})$$

- ▶ Expectation is a linear operator
- ▶ Calculation rule:

$$E[a + bX_1 + cX_2] = a + b E[X_1] + c E[X_2]$$

Moments and Variance

- ▶ n 'th moment:

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x) dx$$

- ▶ n 'th central moment:

$$E[(X - E[X])^n] = \int_{-\infty}^{\infty} (x - E[X])^n f_X(x) dx$$

- ▶ The 2'nd central moment is called the variance:

$$V[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

Covariance

- ▶ Covariance:

$$\text{Cov}[X_1, X_2] = E[(X_1 - E[X_1])(X_2 - E[X_2])] = E[X_1 X_2] - E[X_1]E[X_2]$$

- ▶ Variance and covariance:

$$V[X] = \text{Cov}[X, X]$$

- ▶ Calculation rules:

$$\text{Cov}[aX_1 + bX_2, cX_3 + dX_4] =$$

$$ac \text{Cov}[X_1, X_3] + ad \text{Cov}[X_1, X_4] + bc \text{Cov}[X_2, X_3] + bd \text{Cov}[X_2, X_4]$$

- ▶ The calculation rule can be used for the variance as well. For instance:

$$V[a + bX_2] = b^2 V[X_2]$$

Moment representation

- ▶ All moments up to a given order.
- ▶ Second order moment representation:
 - ▶ Mean
 - ▶ Variance
 - ▶ Covariance (If relevant)

Expectation and Variance for Random Vectors

- ▶ Expectation: $E[\mathbf{X}] = [E[X_1], E[X_2], \dots, E[X_n]]^T$
- ▶ Variance-covariance (matrix): $\Sigma_{\mathbf{X}} = V[\mathbf{X}] = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] =$

$$\begin{bmatrix} V[X_1] & \text{Cov}[X_1, X_2] & \cdots & \text{Cov}[X_1, X_n] \\ \text{Cov}[X_2, X_1] & V[X_2] & \cdots & \text{Cov}[X_2, X_n] \\ \vdots & & & \vdots \\ \text{Cov}[X_n, X_1] & \text{Cov}[X_n, X_2] & \cdots & V[X_n] \end{bmatrix}$$

- ▶ Correlation:

$$\rho_{ij} = \frac{\text{Cov}[X_i, X_j]}{\sqrt{V[X_i]V[X_j]}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

Correlation and Independence

- ▶ If X and Y are independent stochastic variables then $\text{Cov}(X, Y) = 0$ and thus $\text{Corr}(X, Y) = 0$.
- ▶ However, if $X \in N(0, 1)$, then

$$\begin{aligned}\text{Cov}(X, X^2) &= E[X \cdot X^2] - E[X] \cdot E[X^2] = E[X^3] \\ &= \int x^3 f_X(x) dx = 0\end{aligned}$$

- ▶ Thus X and X^2 are uncorrelated, but $E[X^2|X = x] = x^2$.
- ▶ Independence implies no correlation, not the other way around.

Expectation and Variance for Random Vectors

- ▶ The correlation matrix $R = \rho$ is an arrangement of ρ_{ij} in a matrix
- ▶ Covariance matrix between X (dim. p) and Y (dim. q):

$$\begin{aligned}\Sigma_{XY} &= C[X, Y] = E[(X - \mu)(Y - \nu)^T] \\ &= \begin{bmatrix} \text{Cov}[X_1, Y_1] & \cdots & \text{Cov}[X_1, Y_q] \\ \vdots & & \vdots \\ \text{Cov}[X_p, Y_1] & \cdots & \text{Cov}[X_p, Y_q] \end{bmatrix}\end{aligned}$$

- ▶ Calculation rules – see the book.
- ▶ The special case of the variance $C[X, X] = V[X]$ results in

$$V[AX] = A V[X] A^T$$

Conditional expectation

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy$$

$E[Y|X] = E[Y]$ if X and Y are independent

$$E[Y] = E [E[Y|X]]$$

$$E[g(X)Y|X] = g(X)E[Y|X]$$

$$E[g(X)Y] = E [g(X)E[Y|X]]$$

$$E[a|X] = a$$

$$E[g(X)|X] = g(X)$$

$$E[cX + dZ|Y] = cE[X|Y] + dE[Z|Y]$$

Variance separation

- ▶ Definition of conditional variance and covariance:

$$V[Y|X] = E \left[(Y - E[Y|X])(Y - E[Y|X])^T | X \right]$$

$$C[Y, Z|X] = E \left[(Y - E[Y|X])(Z - E[Z|X])^T | X \right]$$

- ▶ The variance separation theorem:

$$V[Y] = E[V[Y|X]] + V[E[Y|X]]$$

$$C[Y, Z] = E[C[Y, Z|X]] + C[E[Y|X], E[Z|X]]$$

Linear Projections

- ▶ Consider two random vectors \mathbf{Y} and \mathbf{X} , then

$$E \left[\begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} \right] = \begin{pmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_X \end{pmatrix} \text{ and } V \left[\begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} \right] = \begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}$$

- ▶ Define the *linear projection* $\rho_X(\mathbf{Y}) \stackrel{def}{=} \boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\mathbf{X} - \boldsymbol{\mu}_X)$
- ▶ Then:
 - ▶ $\rho_X(\mathbf{Y})$ is of the form $\mathbf{a} + B\mathbf{X}$;
 - ▶ $V[\mathbf{Y} - \rho_X(\mathbf{Y})] = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{YX}^T$;
 - ▶ $\text{Cov}(\mathbf{Y} - \rho_X(\mathbf{Y}), \mathbf{X}) = 0$.

Linear projections and conditional means

- ▶ $\rho_X(Y)$ minimizes the variance among functions $a + BX$ that gives projection errors uncorrelated with X .
- ▶ If (X, Y) is multivariate normal, this is a property of $E[Y|X]$.
- ▶ Nevertheless, if $X \sim N(0, 1)$ then
 $\rho_X(X^2) = 1$ while $E[X^2|X] = X^2$
- ▶ We shall write $E[Y|X]$ for $\rho_X(Y)$ anyway.

BECAUSE:

1. The book and other time series literature does so
 2. It is true for Normal stochastic variables
 3. $\rho_X(Y)$ satisfies the same calculus rules as $E[Y|X]$ for normal distributions
- ▶ The differences should be kept in mind.

Air pollution in cities

- ▶ Carstensen (1990) has used time series analysis to set up models for NO and NO_2 at Jagtvej in Copenhagen
- ▶ Measurements of NO and NO_2 are available every third hour (00, 03, 06, 09, 12, ...)
- ▶ We have $\mu_{NO_2} = 48\mu g/m^3$ and $\mu_{NO} = 79\mu g/m^3$
- ▶ In the model $X_{1,t} = NO_{2,t} - \mu_{NO_2}$ and $X_{2,t} = NO_t - \mu_{NO}$ is used

Air pollution in cities – model and forecast

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} = \begin{pmatrix} 0.9 & -0.1 \\ 0.4 & 0.8 \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} + \begin{pmatrix} \xi_{1,t} \\ \xi_{2,t} \end{pmatrix}$$

$$\mathbf{X}_t = \boldsymbol{\Phi} \mathbf{X}_{t-1} + \boldsymbol{\xi}_t$$

$$V[\boldsymbol{\xi}_t] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 30 & 21 \\ 21 & 23 \end{pmatrix} (\mu\text{g}/\text{m}^3)^2$$

- ▶ Assume that t corresponds to 09:00 today and we have measurements $64 \mu\text{g}/\text{m}^3 NO_2$ and $93 \mu\text{g}/\text{m}^3 NO$
- ▶ Forecast the concentrations at 12:00 ($t + 1$)
- ▶ What is the variance-covariance of the forecast error?
- ▶ The best predictor is the conditional expectation

Air pollution in cities – model and forecast

The forecast:

$$\begin{aligned} E(\mathbf{X}_{t+1} | \mathbf{X}_t) &= E(\Phi \mathbf{X}_t + \boldsymbol{\xi}_{t+1} | \mathbf{X}_t) \\ &= \Phi \mathbf{X}_t \end{aligned}$$

Variance-covariance of the forecast error.

$$\begin{aligned} V(\mathbf{X}_{t+1} - E(\mathbf{X}_{t+1} | \mathbf{X}_t) | \mathbf{X}_t) &= V(\Phi \mathbf{X}_t + \boldsymbol{\xi}_{t+1} - \Phi \mathbf{X}_t | \mathbf{X}_t) \\ &= V(\boldsymbol{\xi}_{t+1} | \mathbf{X}_t) \\ &= \Sigma \end{aligned}$$

Air pollution in cities – forecast with R

```
## The system
mu <- matrix(c(48,79),nrow=2)
Phi <- matrix(c(.9,.4,-.1,.8),nrow=2)
Sigma <- matrix(c(20,21,21,23),nrow=2)
## The forecast of the concentrations
Xt <- matrix(c(64,93),nrow=2)-mu
Xtp1.hat <- Phi%*%Xt
Xtp1.hat + mu

##      [,1]
## [1,] 61.0
## [2,] 96.6

## Variance of the error is trivial
```

Air pollution in cities – linear projection

- ▶ At 12:00 ($t + 1$) we now assume that NO_2 is measured with $67 \mu g/m^3$ as the result, **but** NO cannot be measured due to some trouble with the equipment.
- ▶ Estimate the missing NO measurement.
- ▶ What is the variance of the error of the estimation?

Air pollution in cities – linear projection

$$\begin{aligned} E[X_{2,t+1}|X_{1,t+1}, \mathbf{X}_t] &= \overbrace{E[(X_{2,t+1}|X_{1,t+1})|\mathbf{X}_t]}^{\text{compare with (2.65)}} = \\ E[X_{2,t+1}|\mathbf{X}_t] + \text{Cov}(X_{1,t+1}, X_{2,t+1}|\mathbf{X}_t) V[X_{1,t+1}|\mathbf{X}_t]^{-1} (X_{1,t+1} - E(X_{1,t+1}|\mathbf{X}_t)) \\ &= (\Phi_{21} X_{1,t} + \Phi_{22} X_{2,t}) + \Sigma_{12} \Sigma_{11}^{-1} (X_{1,t+1} - (\Phi_{11} X_{1,t} + \Phi_{12} X_{2,t})) \end{aligned}$$

The variance of the projection error is (2.66)

$$\begin{aligned} E(V(X_{2,t+1}|X_{1,t+1}, \mathbf{X}_t)) \\ = V(X_{2,t+1}|\mathbf{X}_t) - \text{Cov}(X_{2,t+1}, X_{1,t+1}|\mathbf{X}_t)^2 V(X_{1,t+1}|\mathbf{X}_t)^{-1} \\ = \Sigma_{22} - \Sigma_{12}^2 / \Sigma_{11} \end{aligned}$$

Air pollution in cities – linear projection with R

```
## The new observation of  $X_{1,t+1}$ 
X1tp1 <- 67 - mu[1]
## The projection
Xtp1.hat[2] + mu[2] +
  Sigma[1,2]/Sigma[1,1] * (X1tp1 - Xtp1.hat[1])
## [1] 102.9

## The variance of the projection error
Sigma[2,2] - Sigma[1,2]^2/Sigma[1,1]
## [1] 0.95
```

Highlights

- ▶ Covariance calculation rule

$$\text{Cov}[aX_1 + bX_2, cX_3 + dX_4] =$$

$$ac \text{Cov}[X_1, X_3] + ad \text{Cov}[X_1, X_4] + bc \text{Cov}[X_2, X_3] + bd \text{Cov}[X_2, X_4]$$

- ▶ The variance separation theorem:

$$V[\mathbf{Y}] = E[V[\mathbf{Y}|\mathbf{X}]] + V[E[\mathbf{Y}|\mathbf{X}]]$$

$$C[\mathbf{Y}, \mathbf{Z}] = E[C[\mathbf{Y}, \mathbf{Z}|\mathbf{X}]] + C[E[\mathbf{Y}|\mathbf{X}], E[\mathbf{Z}|\mathbf{X}]]$$

- ▶ Linear projection:

$$(E[\mathbf{Y}|\mathbf{X}] =) \rho_{\mathbf{X}}(\mathbf{Y}) \stackrel{\text{def}}{=} \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{Y}\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$$

- ▶ Second order moment representation:
All moments up to second order
(mean, variance and covariance).

Exercises

Exercises 2.1, 2.2, 2.3

Correction in exercise 2.2: X and ε are mutually independent.

Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

September 15, 2017

Outline of the lecture

- ▶ Regression based methods, 1st part:
 - ▶ Introduction (Sec. 3.1)
 - ▶ The General Linear Model, including OLS-, WLS-, and ML-estimates (Sec. 3.2)
 - ▶ Prediction in the General Linear Model (Sec. 3.3)
 - ▶ Examples...

General form of the regression model

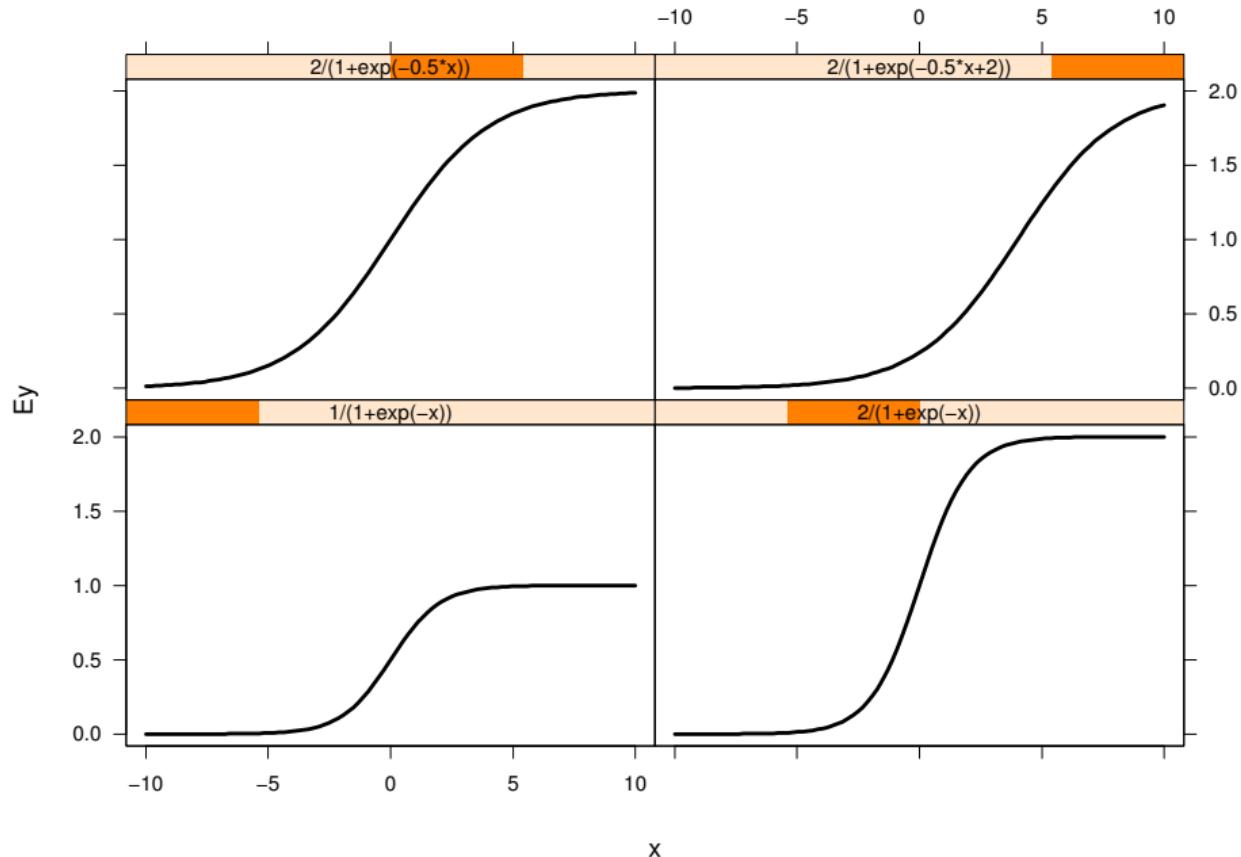
$$Y_t = f(\mathbf{X}_t, t; \boldsymbol{\theta}) + \varepsilon_t$$

Where:

- ▶ Y_t is the output we aim to model
- ▶ \mathbf{X}_t indicates the p independent variables $\mathbf{X}_t = (X_{1t}, \dots, X_{pt})^T$
- ▶ t is the time index
- ▶ $\boldsymbol{\theta}$ indicates m unknown parameters $(\theta_1, \dots, \theta_m)^T$
- ▶ ε_t is a sequence of random variables with mean zero, variance σ_t^2 , and $\text{Cov}[\varepsilon_{t_i}, \varepsilon_{t_j}] = \sigma_t^2 \Sigma_{ij}$

For now we restrict the discussion to the case where \mathbf{X}_t is non-random and thus we write \mathbf{x}_t instead of \mathbf{X}_t .

$$Y_t = \theta_1 / (1 + \exp(-\theta_2 x_t + \theta_3)) + \varepsilon_t$$



Ordinary least squares (OLS) estimates

Observations:

$$(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \dots, (y_n, \mathbf{x}_n)$$

Ordinary Least Square (unweighted) estimates are found from

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} S(\boldsymbol{\theta})$$

where

$$S(\boldsymbol{\theta}) = \sum_{t=1}^n [y_t - f(\mathbf{x}_t; \boldsymbol{\theta})]^2 = \sum_{t=1}^n \varepsilon_t^2(\boldsymbol{\theta})$$

For the unweighted method to result in reliable estimates, the errors must be assumed to all have the same variance and be mutually uncorrelated.

OLS – Variance of error and estimates

If the model errors ε_t are i.i.d.

- ▶ The variance of the model errors is estimated as:

$$\hat{\sigma}^2 = \frac{S(\hat{\theta})}{n - p}$$

where p is the number of estimated parameters.

- ▶ The variance-covariance matrix of the estimates is approximately

$$V[\hat{\theta}] = 2\hat{\sigma}^2 \left[\frac{\partial^2}{\partial^2 \theta} S(\theta) \right]^{-1} \Big|_{\theta=\hat{\theta}}$$

The General Linear Model (GLM)

$$Y_t = \boldsymbol{x}_t^T \boldsymbol{\theta} + \varepsilon_t$$

Can this quadratic model in z_t be a GLM?

$$Y_t = \theta_0 + \theta_1 z_t + \theta_2 z_t^2 + \varepsilon_t$$

Yes, as it can be written as

$$y_t = \begin{pmatrix} 1 & z_t & z_t^2 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix} + \varepsilon_t$$

It is linearity in the parameters that matters!

General Linear Model - sub classes

(Multiple) regression analysis, ex: $Y = \alpha + \beta x + \varepsilon$

Ex: The height of a plant described by age (x_1), concentration of nutrients in soil (x_2), etc.

Analysis of variance, ex: $Y = \alpha_i + \varepsilon$

Ex: The height of plants described by species (i).

Analysis of covariance, ex: $Y = \alpha_i + \beta x + \varepsilon$

Ex: Height of plant described by species *and* age, nutrients...

Comments

- ▶ For ANOVA and ANCOVA the treatments must be coded into a number of x -variables.
- ▶ Some examples in the book

OLS-estimates

- ▶ Non-linear regression: Numerical optimization is required; see the book for a simple example (Newton-Raphson)
- ▶ For the general linear model a closed-form solution exists.
For all observations the model equations are written as:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \text{or} \quad \mathbf{Y} = \mathbf{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

i.e. we want to minimize $S(\boldsymbol{\theta}) = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$

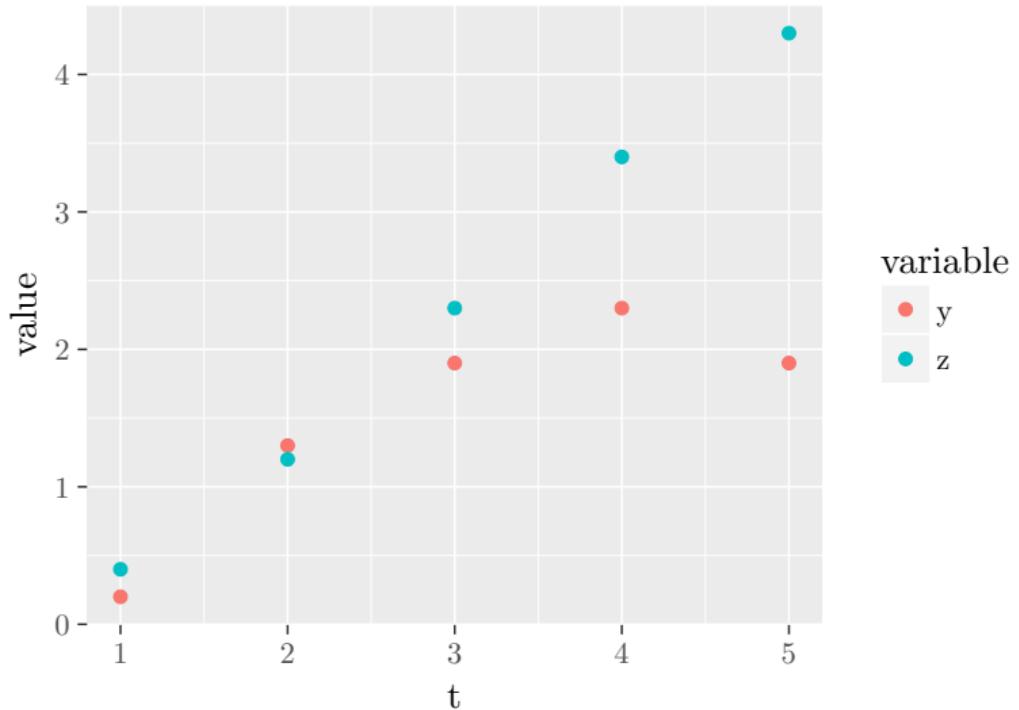
- ▶ The solution is $\widehat{\boldsymbol{\theta}} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{Y}$ (if \mathbf{x} has full rank)
- ▶ $\widehat{\sigma}^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} / (n - p)$ and $V[\widehat{\boldsymbol{\theta}}] = \widehat{\sigma}^2 (\mathbf{x}^T \mathbf{x})^{-1}$

Example

Data:

t	y	z
1	0.2	0.4
2	1.3	1.2
3	1.9	2.3
4	2.3	3.4
5	1.9	4.3

Model: $Y_t = \theta_0 + \theta_1 z_t + \theta_2 z_t^2 + \varepsilon_t$



Example

t	y	z
1	0.2	0.4
2	1.3	1.2
3	1.9	2.3
4	2.3	3.4
5	1.9	4.3

Data:

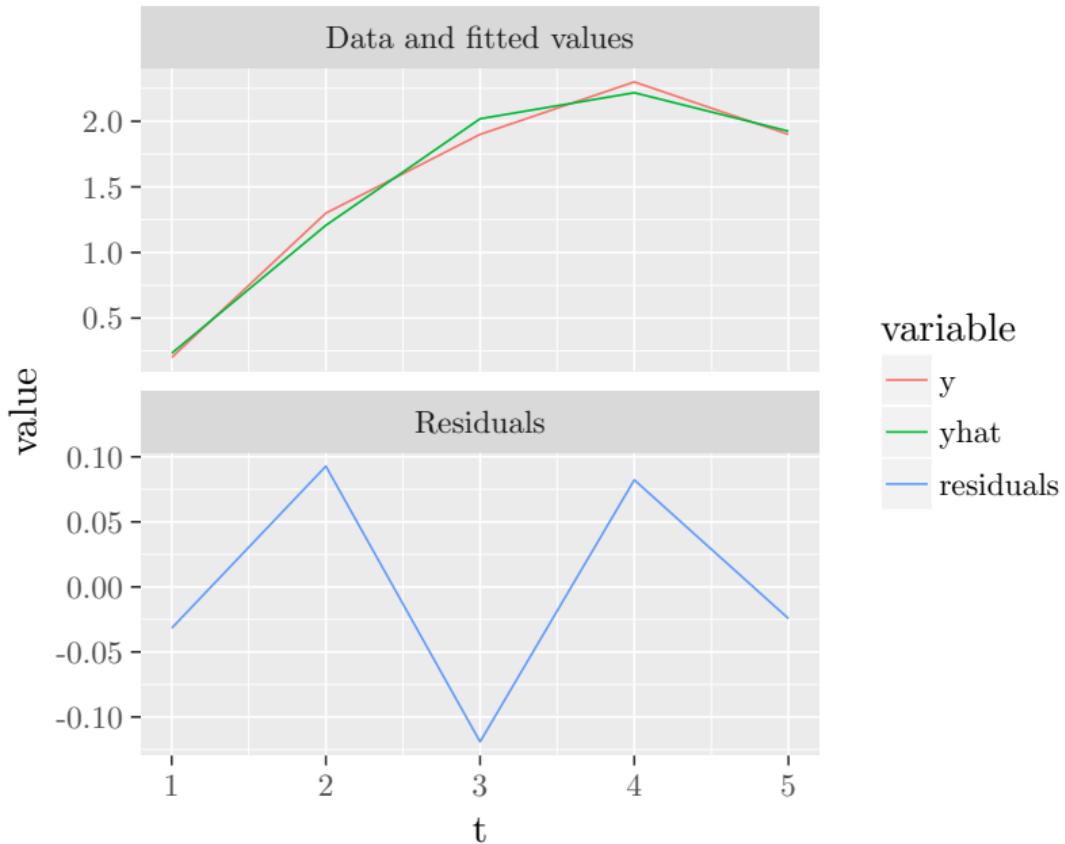
Model: $Y_t = \theta_0 + \theta_1 z_t + \theta_2 z_t^2 + \varepsilon_t$

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} 0.2 \\ 1.3 \\ 1.9 \\ 2.3 \\ 1.9 \end{bmatrix} = \begin{bmatrix} 1 & 0.4 & 0.16 \\ 1 & 1.2 & 1.44 \\ 1 & 2.3 & 5.29 \\ 1 & 3.4 & 11.56 \\ 1 & 4.3 & 18.49 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{Y} = \begin{bmatrix} -0.38 \\ 1.62 \\ -0.25 \end{bmatrix}$$

Plot of the model fit



Properties of the OLS-estimator of a GLM

- ▶ It is a linear function of the observations \mathbf{Y} (and $\widehat{\mathbf{Y}}$ is thus a linear function of the observations)
- ▶ It is unbiased, i.e. $E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$
- ▶ $V[\widehat{\boldsymbol{\theta}}] = E \left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \right] = \sigma^2(\mathbf{x}^T \mathbf{x})^{-1}$
- ▶ $\widehat{\boldsymbol{\theta}}$ is BLUE (Best Linear Unbiased Estimator), which means that it has the smallest variance among all estimators which are a linear function of the observations.

WLS-estimates

- ▶ Equation for all observations: $\mathbf{Y} = \mathbf{x}\boldsymbol{\theta} + \boldsymbol{\epsilon}$
- ▶ $E[\boldsymbol{\epsilon}] = \mathbf{0}$ and $V[\boldsymbol{\epsilon}] = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is known
- ▶ We want to minimize $(\mathbf{Y} - \mathbf{x}\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\boldsymbol{\theta})$
- ▶ The solution is

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

(if $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}$ is invertible)

- ▶ An estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})$$

Example WLS/OLS

- ▶ H. Madsen & P. Thyregod (1988). *Modelling the Time Correlation in Hourly Observations of Direct Radiation in Clear Skies*. Energy and Buildings, **11**, 201–211.
- ▶ See the examples in the book.

Example WLS/OLS: Clear sky radiation

$$I_N(h(t)) = a_N(1 - \exp(-b_N h(t))) + \varepsilon_N(t) \quad (1)$$

$$\varepsilon_N(t) \sim N(\mathbf{0}, \sigma^2 \Sigma) \quad (2)$$

Suggested variance structures

- ▶ i.i.d

$$\Sigma = I$$

- ▶ Only correlation

$$\Sigma_{ij} = \rho^{|t_i - t_j|}$$

- ▶ Only variance

$$\Sigma_{ij} = \frac{1}{\sin(h(t_i)) \sin(h(t_j))}$$

- ▶ Both correlation and variance

$$\Sigma_{ij} = \frac{\rho^{|t_i - t_j|}}{\sin(h(t_i)) \sin(h(t_j))}$$

Maximum Likelihood (ML) - estimates

- ▶ We now assume that the observations are Gaussian:

$$\mathbf{Y} \sim N_n(\mathbf{x}\boldsymbol{\theta}, \sigma^2 \boldsymbol{\Sigma})$$

- ▶ $\boldsymbol{\Sigma}$ is assumed known
- ▶ The ML-estimator is (here) the same as the WLS-estimator:

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x})^{-1} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y}$$

- ▶ The ML-estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})$$

Properties of the ML-estimator

- ▶ It is a linear function of the observations which now implies that it is normally distributed.
- ▶ It is unbiased, i.e. $E[\hat{\theta}] = \theta$ and
- ▶ The variance $V[\hat{\theta}] = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] = (\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x})^{-1} \sigma^2$;
- ▶ It is an efficient estimator (minimum variance of unbiased estimators).

Unknown Σ

Relaxation algorithm:

- a) Select a value for Σ (e.g. $\Sigma = I$).
- b) Find the estimates for this value of Σ e.g. by solving the normal equations.
- c) Consider the residuals $\{\hat{\epsilon}_t\}$ and calculate the correlation and variance structure of the residuals.
Then select a new value for Σ which reflects that correlation and variance structure.
- d) Stop if convergence - otherwise go to b).

See (Goodwin and Payne, 1977) for details.

Prediction

Theorem 3.8 and 3.9

- ▶ If the expected value of the squared prediction error is to be minimized, then
- ▶ the expected mean $E[Y|\mathbf{X} = \mathbf{x}]$ is the optimal predictor.

Prediction in the general linear model

- ▶ Known parameters:

$$\hat{Y}_{t+\ell} = E_{\theta}[Y_{t+\ell} | \mathbf{X}_{t+\ell} = \mathbf{x}_{t+\ell}] = \mathbf{x}_{t+\ell}^T \boldsymbol{\theta}$$

$$V_{\theta}[Y_{t+\ell} - \hat{Y}_{t+\ell}] = V_{\theta}[\varepsilon_{t+\ell}] = \sigma^2$$

- ▶ Estimated parameters:

$$\hat{Y}_{t+\ell} = E_{\hat{\theta}}[Y_{t+\ell} | \mathbf{X}_{t+\ell} = \mathbf{x}_{t+\ell}] = \mathbf{x}_{t+\ell}^T \hat{\boldsymbol{\theta}}$$

$$V_{\hat{\theta}}[Y_{t+\ell} - \hat{Y}_{t+\ell}] = V_{\hat{\theta}}[\varepsilon_{t+\ell} + \mathbf{x}_{t+\ell}^T (\theta - \hat{\theta})] = \hat{\sigma}^2 [1 + \mathbf{x}_{t+\ell}^T (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}_{t+\ell}]$$

Prediction in the general linear model – continued

- ▶ In practice we have to use an estimate of σ and therefore a $100(1 - \alpha)\%$ prediction interval of a future value is calculated as:

$$\hat{Y}_{t+\ell} \pm t_{\alpha/2}(n-p)\hat{\sigma}\sqrt{1 + \mathbf{x}_{t+\ell}^T(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}_{t+\ell}}$$

where $t_{\alpha/2}(n-p)$ refers to the $\alpha/2$ 'th quantile of the t -distribution with $n-p$ degrees of freedom.

- ▶ For $n-p$ large, percentiles from the normal distribution can be used.

Highlights

- When ε_t are i.i.d then the variance of the model errors is estimated as:

$$\hat{\sigma}^2 = \frac{S(\hat{\theta})}{n - p}$$

- OLS estimator:

$$\hat{\theta} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{Y}$$

- WLS estimator (When $V[\boldsymbol{\epsilon}] = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 \boldsymbol{\Sigma}$)

$$\hat{\theta} = (\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x})^{-1} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y}$$

- ML estimator for $\hat{\theta}$ is the same as the WLS estimator.
- but not for $\hat{\sigma}^2$
- The expected mean $E[Y|\mathbf{X} = \mathbf{x}]$ is the optimal predictor.
- Prediction in GLM

$$\hat{Y}_{t+\ell} \pm t_{\alpha/2}(n - p)\hat{\sigma} \sqrt{1 + \mathbf{x}_{t+\ell}^T (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}_{t+\ell}}$$

Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

September 22, 2017

Outline of the lecture

- ▶ Regression based methods, 2nd part:
 - ▶ Global trend models (Sec. 3.4)
- ▶ R examples

Trend models

- ▶ Linear regression model
- ▶ Functions of time are taken as the independent variables

$$Y_{N+j} = f^T(j)\boldsymbol{\theta} + \varepsilon_{N+j}$$

Linear trend - A motivation

- ▶ Observations for $t = 1, \dots, N$. Naive formulation of the model: $Y_t = \phi_0 + \phi_1 t + \varepsilon_t$
- ▶ If we want to forecast Y_{N+j} given information up to N we use $\hat{Y}_{N+j|N} = \hat{\phi}_0 + \hat{\phi}_1 (N + j)$
- ▶ However, for on-line applications $N + j$ can be arbitrary large
- ▶ The problem arise because ϕ_0 and ϕ_1 are defined w.r.t. the origin 0

- ▶ Defining the parameters w.r.t. the origin N we obtain the model: $Y_t = \theta_0 + \theta_1 (t - N) + \varepsilon_t$
- ▶ Using this formulation we get: $\hat{Y}_{N+j|N} = \hat{\theta}_0 + \hat{\theta}_1 j$
- ▶ Same model, different parameterisation.

Linear trend in a general setting

- ▶ The general trend model:

$$Y_{N+j} = \mathbf{f}^T(j) \boldsymbol{\theta} + \varepsilon_{N+j}$$

- ▶ The linear trend model is obtained when: $\mathbf{f}(j) = \begin{pmatrix} 1 \\ j \end{pmatrix}$
- ▶ It follows that for $N+1+j$:

$$Y_{N+1+j} = \begin{pmatrix} 1 \\ j+1 \end{pmatrix}^T \boldsymbol{\theta} + \varepsilon_{N+1+j} = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}}_L \begin{pmatrix} 1 \\ j \end{pmatrix}^T \boldsymbol{\theta} + \varepsilon_{N+1+j}$$

- ▶ The 2×2 matrix L defines the transition from $\mathbf{f}(j)$ to $\mathbf{f}(j+1)$

Trend models in general

- ▶ Model: $Y_{N+j} = \mathbf{f}^T(j)\boldsymbol{\theta} + \varepsilon_{N+j}$
- ▶ Requirement: $\mathbf{f}(j+1) = \mathbf{L}\mathbf{f}(j)$
- ▶ Initial value: $\mathbf{f}(0)$
- ▶ In Section 3.4 some trend models which fulfill the requirement above are listed.
 - ▶ Constant mean: $Y_{N+j} = \theta_0 + \varepsilon_{N+j}$
 - ▶ Linear trend: $Y_{N+j} = \theta_0 + \theta_1 j + \varepsilon_{N+j}$
 - ▶ Quadratic trend: $Y_{N+j} = \theta_0 + \theta_1 j + \theta_2 \frac{j^2}{2} + \varepsilon_{N+j}$
 - ▶ k 'th order polynomial trend: $Y_{N+j} = \theta_0 + \theta_1 j + \theta_2 \frac{j^2}{2} + \cdots + \theta_k \frac{j^k}{k!} + \varepsilon_{N+j}$
 - ▶ Harmonic model with the period p : $Y_{N+j} = \theta_0 + \theta_1 \sin \frac{2\pi}{p} j + \theta_2 \cos \frac{2\pi}{p} j + \varepsilon_{N+j}$

Estimation

- Model equations written for all observations Y_1, \dots, Y_N

$$\begin{aligned} \mathbf{Y}_N &= \mathbf{x}_N \boldsymbol{\theta}_N + \boldsymbol{\varepsilon} \\ \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} &= \begin{bmatrix} \mathbf{f}^T(-N+1) \\ \mathbf{f}^T(-N+2) \\ \vdots \\ \mathbf{f}^T(0) \end{bmatrix} \boldsymbol{\theta}_N + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \end{aligned}$$

- OLS-estimates:

$$\hat{\boldsymbol{\theta}}_N = (\mathbf{x}_N^T \mathbf{x}_N)^{-1} \mathbf{x}_N^T \mathbf{Y}_N = \mathbf{F}_N^{-1} \mathbf{h}_N$$

$$\mathbf{F}_N = \mathbf{x}_N^T \mathbf{x}_N = \sum_{j=0}^{N-1} \mathbf{f}(-j) \mathbf{f}^T(-j)$$

$$\mathbf{h}_N = \mathbf{x}_N^T \mathbf{Y} = \sum_{j=0}^{N-1} \mathbf{f}(-j) Y_{N-j}$$

ℓ -step prediction

- ▶ Prediction:

$$\hat{Y}_{N+\ell|N} = \mathbf{f}^T(\ell) \hat{\boldsymbol{\theta}}_N$$

- ▶ Variance of the prediction error:

$$V[Y_{N+\ell} - \hat{Y}_{N+\ell|N}] = \sigma^2 [1 + \mathbf{f}^T(\ell) \mathbf{F}_N^{-1} \mathbf{f}(\ell)]$$

- ▶ $100(1 - \alpha)\%$ prediction interval:

$$\begin{aligned} & \hat{Y}_{N+\ell|N} \pm t_{\alpha/2}(N-p) \sqrt{V[e_N(\ell)]} \\ &= \hat{Y}_{N+\ell|N} \pm t_{\alpha/2}(N-p) \hat{\sigma} \sqrt{1 + \mathbf{f}^T(\ell) \mathbf{F}_N^{-1} \mathbf{f}(\ell)} \end{aligned}$$

where $\hat{\sigma}^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} / (N - p)$ (p is the number of estimated parameters)

Updating the estimates when Y_{N+1} is available

► Task:

- Going from estimates based on $t = 1, \dots, N$, i.e. $\hat{\theta}_N$ to
- estimates based on $t = 1, \dots, N, N+1$, i.e. $\hat{\theta}_{N+1}$
- without redoing everything...

► Solution:

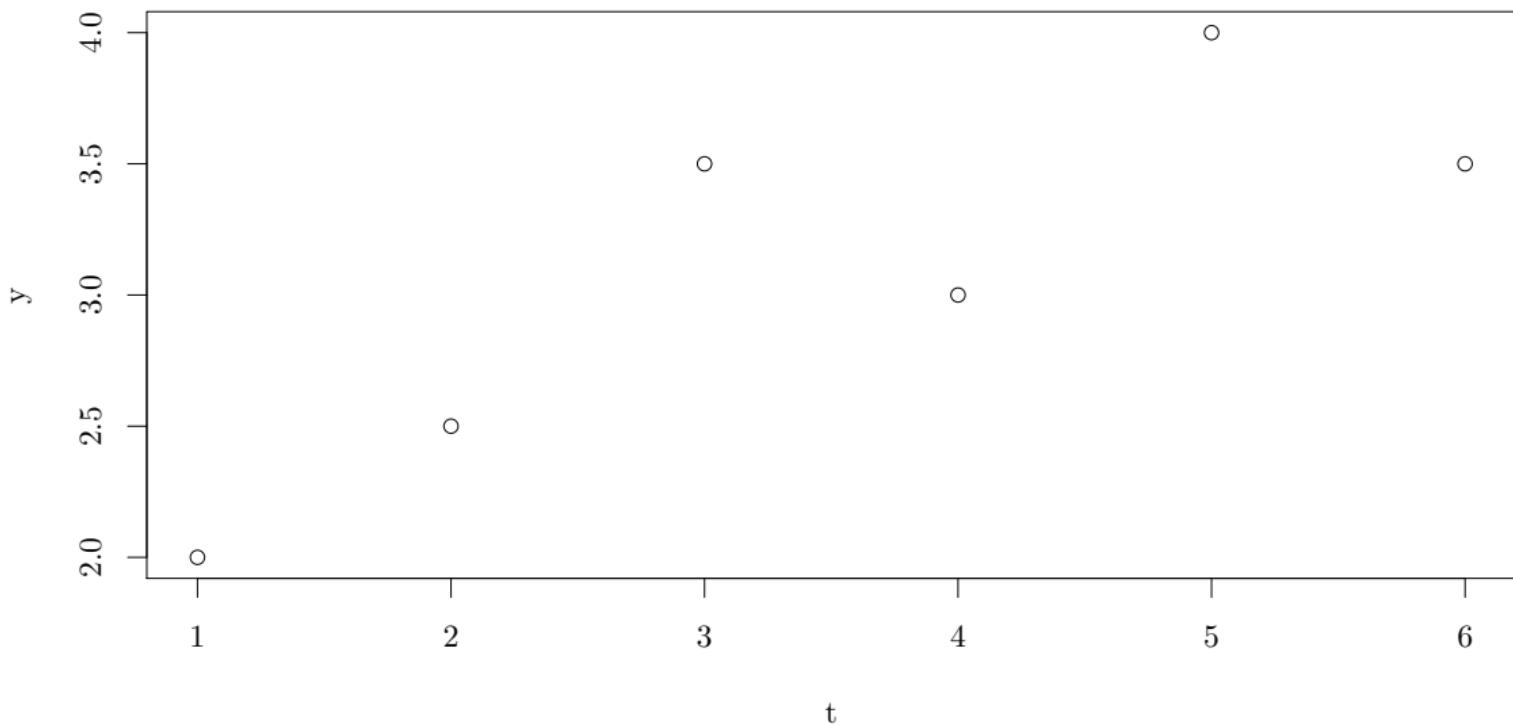
$$\mathbf{F}_{N+1} = \mathbf{F}_N + \mathbf{f}(-N) \mathbf{f}^T(-N)$$

$$\mathbf{h}_{N+1} = \mathbf{L}^{-1} \mathbf{h}_N + \mathbf{f}(0) Y_{N+1}$$

$$\hat{\theta}_{N+1} = \mathbf{F}_{N+1}^{-1} \mathbf{h}_{N+1}$$

Local Trend Models - an Example

6 observations ($N = 6$):



Global Linear Trend:

$$Y_{N+j} = \theta_0 + \theta_1 j + \varepsilon_{N+j} \Rightarrow \mathbf{f}(j) = (1 \quad j)^T$$

Linear Model form:

$$\begin{pmatrix} 2.0 \\ 2.5 \\ 3.5 \\ 3.0 \\ 4.0 \\ 3.5 \end{pmatrix} = \begin{pmatrix} 1 & -5 \\ 1 & -4 \\ 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix} \Leftrightarrow \mathbf{y} = \mathbf{x}_6 \boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

Global linear trend: Estimation

$$\begin{aligned}\mathbf{F}_6 &= \mathbf{x}_6^T \mathbf{x}_6 = \begin{pmatrix} 6 & -15 \\ -15 & 55 \end{pmatrix} \\ \mathbf{h}_6 &= \mathbf{x}_6^T \mathbf{y} = \begin{pmatrix} 18.5 \\ -40.5 \end{pmatrix} \\ \hat{\boldsymbol{\theta}}_6 &= \mathbf{F}_6^{-1} \mathbf{h}_6 = \begin{pmatrix} 0.5238 & 0.1429 \\ 0.1429 & 0.0571 \end{pmatrix} \begin{pmatrix} 18.5 \\ -40.5 \end{pmatrix} = \begin{pmatrix} 3.905 \\ 0.329 \end{pmatrix}\end{aligned}$$

Global linear trend: Estimation with R

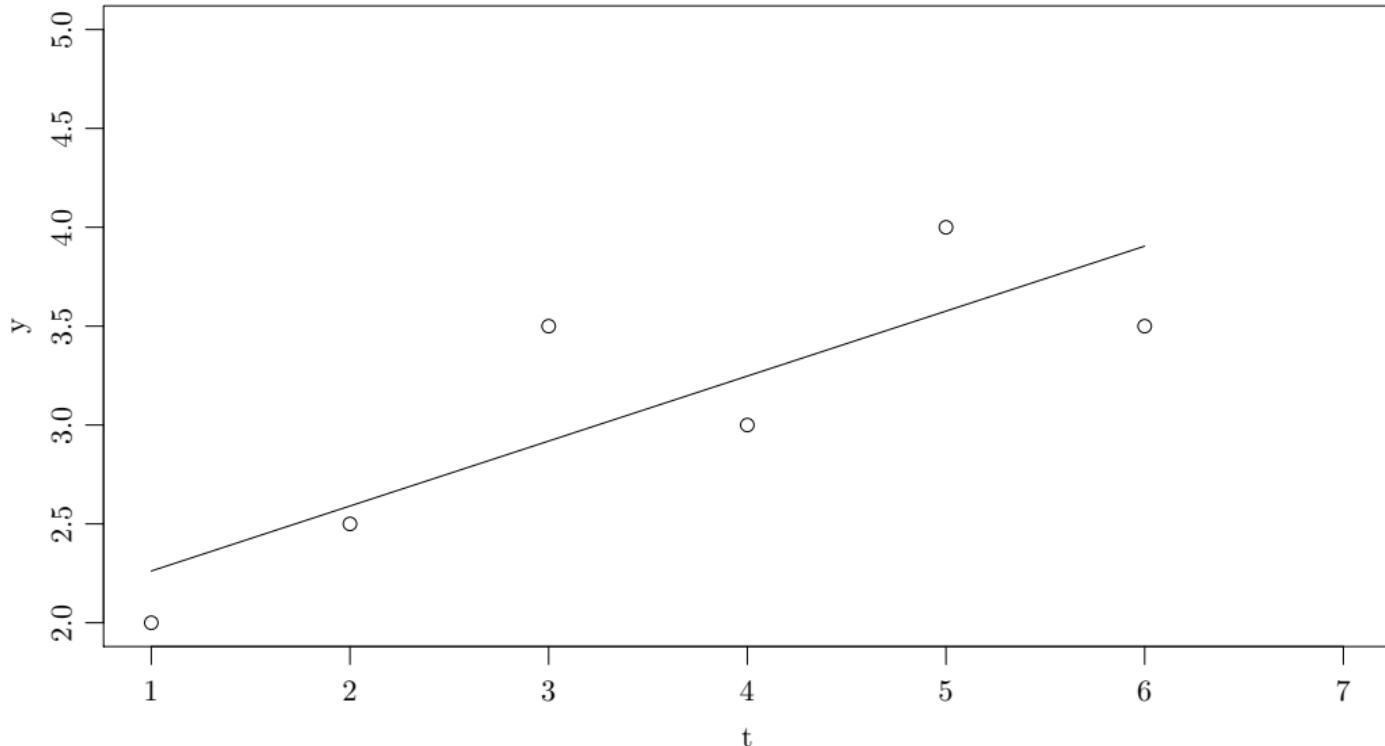
```
F6 <- t(x) %*% x
h6 <- t(x) %*% y
(th.hat6 <- solve(F6, h6))

##           [,1]
## [1,] 3.9047619
## [2,] 0.3285714

## check
(lm(y~0+x))

##
## Call:
## lm(formula = y ~ 0 + x)
##
## Coefficients:
##       x1       x2
## 3.9048  0.3286
```

Global linear trend: Estimation - global linear trend



Global linear trend: Prediction

Linear predictor:

$$\hat{Y}_{6+\ell|6} = f(\ell)^T \hat{\theta}_6 = 3.905 + 0.328\ell$$

LS-estimate for σ^2 :

$$\begin{aligned}\hat{\sigma}^2 &= (\mathbf{y} - \mathbf{x}_6 \hat{\theta}_6)^T (\mathbf{y} - \mathbf{x}_6 \hat{\theta}_6) / (6 - 2) \\ &= \frac{(-0.262)^2 + 0.090^2 + 0.581^2 + (-0.248)^2 + 0.424^2 + (-0.405)^2}{4} \\ &= 0.453^2\end{aligned}$$

Global linear trend: Prediction Error

$$\varepsilon_6(\ell) = Y_{6+\ell} - \hat{Y}_{6+\ell|6}$$

$$\begin{aligned}\widehat{\text{Var}}(\varepsilon_6(\ell)) &= \hat{\sigma}^2 (1 + \mathbf{f}^T(\ell) \mathbf{F}_6^{-1} \mathbf{f}(\ell)) \\ &= 0.453^2 \left(1 + (1 - \ell) \begin{pmatrix} 0.5238 & 0.1429 \\ 0.1429 & 0.0571 \end{pmatrix} \begin{pmatrix} 1 \\ \ell \end{pmatrix} \right) \\ &= 0.453^2 (1.5238 + 0.2858\ell + 0.0571\ell^2)\end{aligned}$$

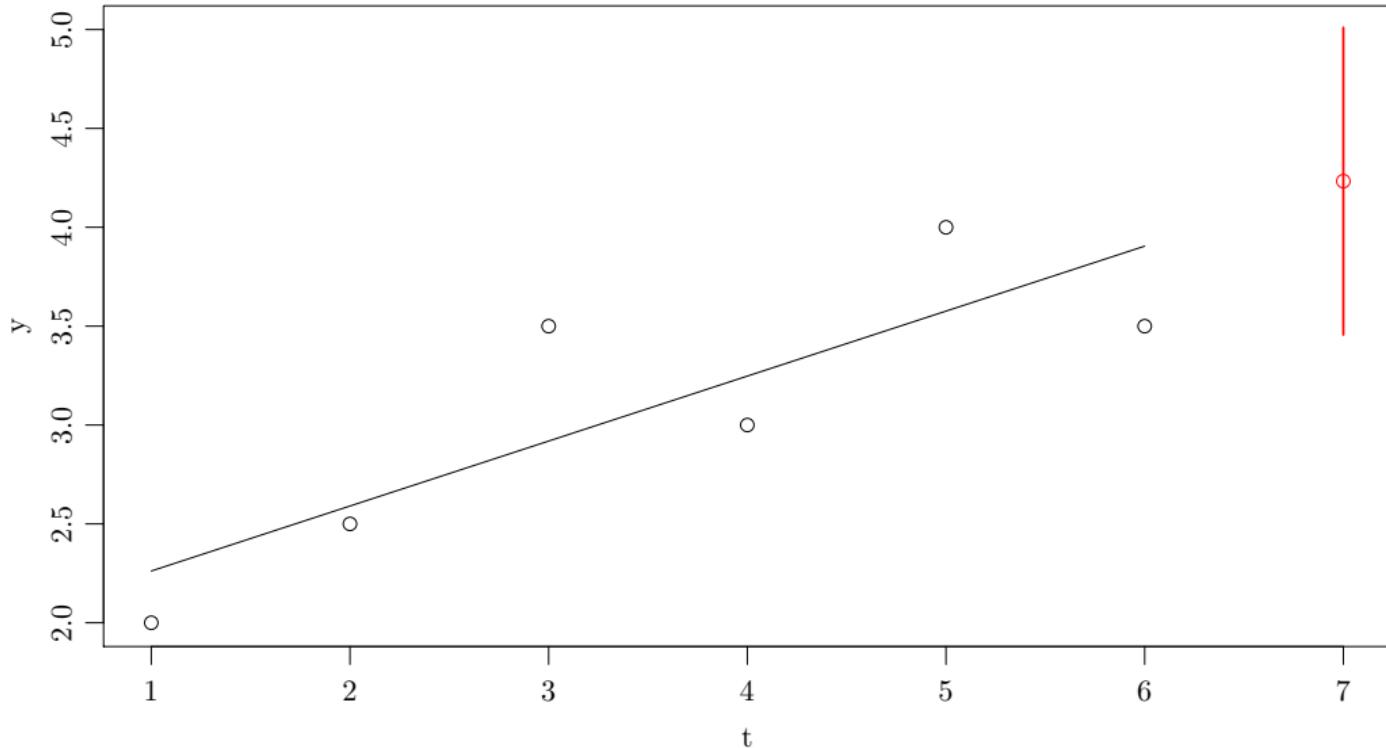
For example,

$$\hat{Y}_{7|6} = 4.234 \text{ with } \widehat{\text{Var}}(\varepsilon_6(1)) = 0.619^2.$$

90% prediction interval:

$$\hat{Y}_{7|6} \pm t_{0.05}(6 - 2) \sqrt{\widehat{\text{Var}}(\varepsilon_6(1))} = 4.234 \pm 1.320$$

Global linear trend: Estimation - global linear trend



Global linear trend: Updating the parameters

- ▶ New observation: $y_7 = 3.5$.

$$\begin{aligned}\mathbf{F}_7 &= \mathbf{F}_6 + \mathbf{f}(-6)\mathbf{f}^T(-6) \\&= \begin{pmatrix} 6 & -15 \\ -15 & 55 \end{pmatrix} + \begin{pmatrix} 1 \\ -6 \end{pmatrix} \begin{pmatrix} 1 & -6 \end{pmatrix} = \begin{pmatrix} 7 & -21 \\ -21 & 91 \end{pmatrix}, \\ \mathbf{h}_7 &= L^{-1}\mathbf{h}_6 + \mathbf{f}(0)\mathbf{y}_7 \\&= \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 18.5 \\ -40.5 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} 3.5 = \begin{pmatrix} 22 \\ -59 \end{pmatrix}, \\ \hat{\boldsymbol{\theta}}_7 &= \begin{pmatrix} 0.4643 & 0.1071 \\ 0.1071 & 0.0357 \end{pmatrix} \begin{pmatrix} 22 \\ -59 \end{pmatrix} = \begin{pmatrix} 3.896 \\ 0.250 \end{pmatrix}.\end{aligned}$$

Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

September 26, 2017

Outline of the lecture

- ▶ Regression based methods, 3rd part:
 - ▶ Regression and exponential smoothing (Sec. 3.4)
 - ▶ Global and local trend models - an example (Sec. 3.6)
- ▶ Operators; the backward shift operator; sec. 4.5.

Predictions In Time Series

- ▶ Are model-based - one for all data (so far).
- ▶ What if no model fits our need (and data)?
- ▶ Methods that aren't model based in the usual sense applies.

Exponential smoothing

Given a forgetting factor $\lambda \in]0; 1[$

$$\hat{\mu}_N = c \sum_{j=0}^{N-1} \lambda^j Y_{N-j} = c [Y_N + \lambda Y_{N-1} + \cdots + \lambda^{N-1} Y_1]$$

The constant c is chosen so that the weights sum to one, which implies that $c = (1 - \lambda)/(1 - \lambda^N)$.

When N is large $c \approx 1 - \lambda$:

$$\begin{aligned}\hat{\mu}_N &= (1 - \lambda)[Y_N + \lambda Y_{N-1} + \cdots + \lambda^{N-1} Y_1] \\ &= (1 - \lambda)Y_N + (1 - \lambda)[\lambda Y_{N-1} + \cdots + \lambda^{N-1} Y_1] \\ &= (1 - \lambda)Y_N + \lambda(1 - \lambda)[Y_{N-1} + \cdots + \lambda^{N-2} Y_1] \\ &= (1 - \lambda)Y_N + \lambda\hat{\mu}_{N-1}\end{aligned}$$

Exponential Smoothing and prediction

Used as a prediction model:

$$\hat{Y}_{N+\ell|N} = \hat{\mu}_N$$

Updating predictions with new observations:

$$\hat{Y}_{N+\ell+1|N+1} = (1 - \lambda) Y_{N+1} + \lambda \hat{Y}_{N+\ell|N}$$

Simple Exponential Smoothing

For large N :

$$\hat{\mu}_{N+1} = (1 - \lambda) Y_{N+1} + \lambda \hat{\mu}_N$$

Definition (Simple Exponential Smoothing):

The sequence S_N defined by

$$S_N = (1 - \lambda) Y_N + \lambda S_{N-1}$$

is called the *simple exponential smoothing* or first order exponential smoothing of the time series Y .

- ▶ The smoothing constant $\alpha = 1 - \lambda$ (or the forgetting factor λ) determines how much the latest observation influence the prediction.

Choice of smoothing constant $\alpha = 1 - \lambda$

- Given a data set $t = 1, \dots, N$ we construct

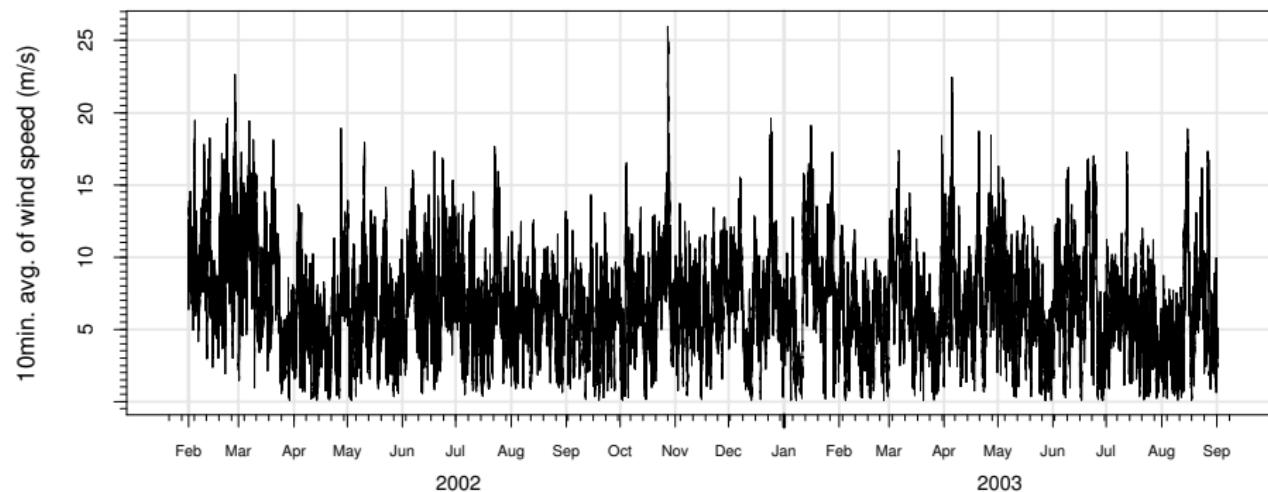
$$S(\alpha) = \sum_{t=1}^N (Y_t - \hat{Y}_{t|t-l}(\alpha))^2 = \sum_{t=1}^N (Y_t - \hat{\mu}_{t-l}(\alpha))^2$$

The value minimizing $S(\alpha)$ is chosen.

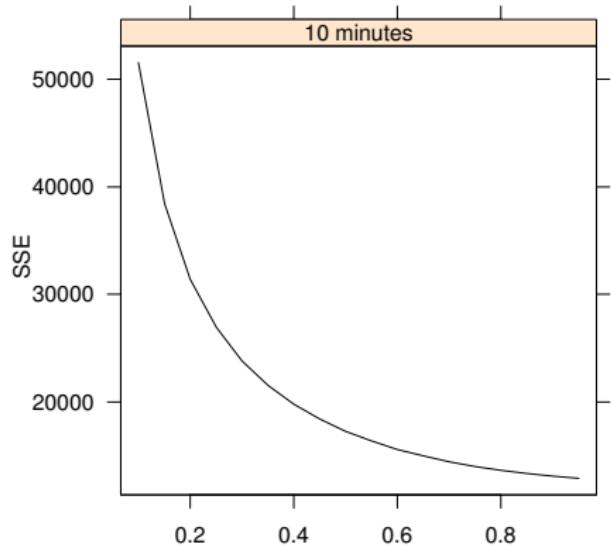
- If the data set is large we eliminate the influence of the initial estimate by dropping the first part of the errors when evaluating $S(\alpha)$
- Keep in mind however what the smoothing is used for, and modify the criteria accordingly. The next slides show an example of this.

Example – wind speed 76 m a.g.l. at DTU Risø

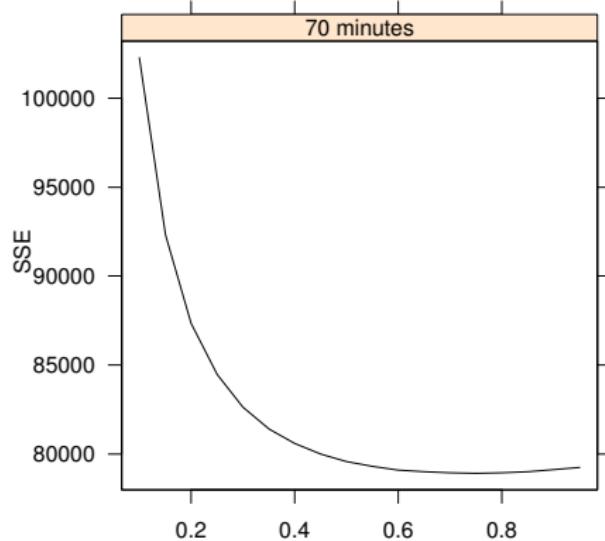
- ▶ Measurements of wind speed every 10th minute
- ▶ Task: Forecast up to approximately 3 hours ahead using exponential smoothing



$S(\alpha)$ for horizons 10 and 70 minutes



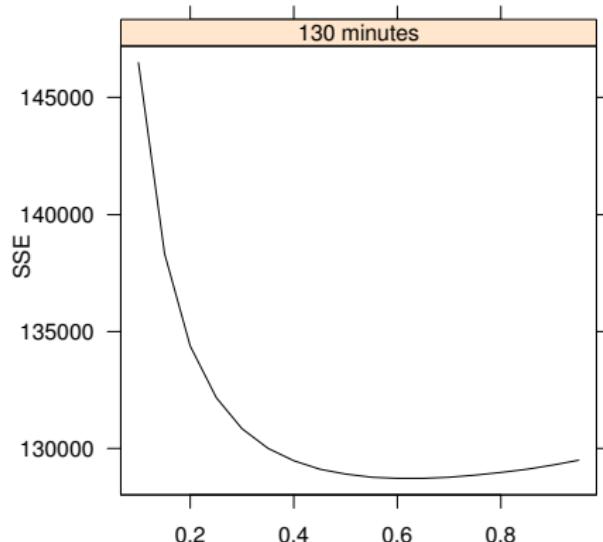
Weight on most recent observation



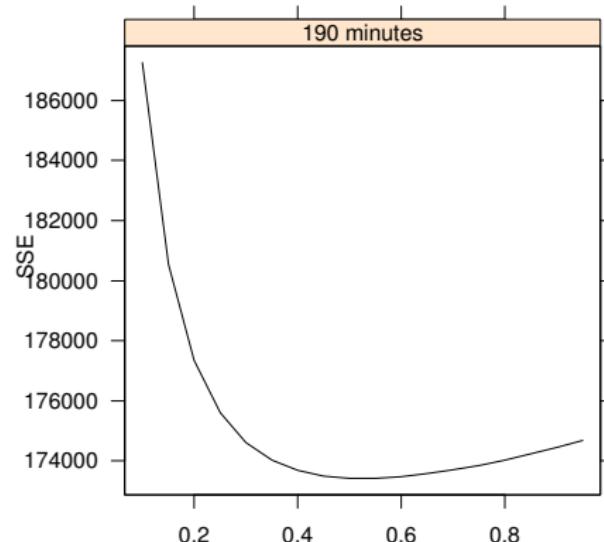
Weight on most recent observation

- ▶ 10 minutes (1-step): Use $\alpha = 0.95$ or higher
- ▶ 70 minutes (7-step): Use $\alpha \approx 0.7$

$S(\alpha)$ for horizons 130 and 190 minutes



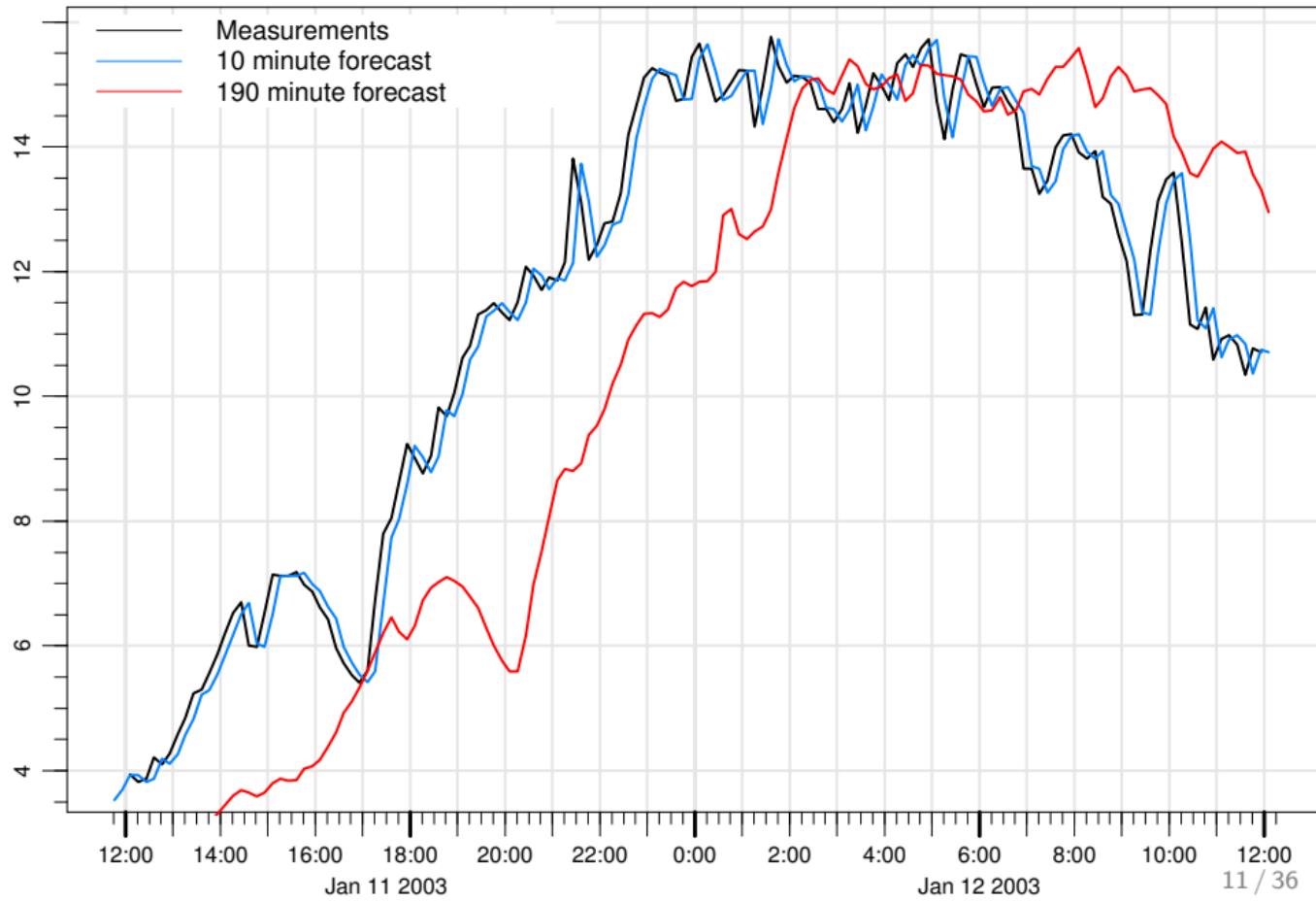
Weight on most recent observation



Weight on most recent observation

- ▶ 130 minutes (13-step): Use $\alpha \approx 0.6$
- ▶ 190 minutes (19-step): Use $\alpha \approx 0.5$

Example of forecasts with optimal α



From global to local trend models

Last week we worked with the global trend model

$$Y_{N+j} = \mathbf{f}^T(j) \boldsymbol{\theta} + \varepsilon_{N+j}$$

which was solved iteratively by

$$\begin{aligned}\mathbf{F}_{N+1} &= \mathbf{F}_N + \mathbf{f}(-N) \mathbf{f}^T(-N) \\ \mathbf{h}_{N+1} &= \mathbf{L}^{-1} \mathbf{h}_N + \mathbf{f}(0) Y_{N+1} \\ \hat{\boldsymbol{\theta}}_{N+1} &= \mathbf{F}_{N+1}^{-1} \mathbf{h}_{N+1}\end{aligned}$$

Could we do that locally?

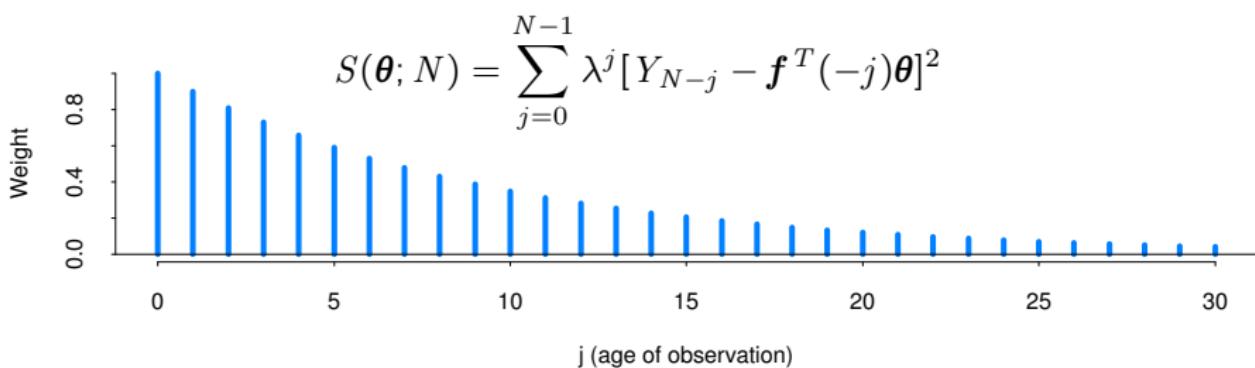
Local trend models

We forget old observations in an exponential manner:

$$\hat{\boldsymbol{\theta}}_N = \arg \min_{\boldsymbol{\theta}} S(\boldsymbol{\theta}; N)$$

where for $0 < \lambda < 1$

$$S(\boldsymbol{\theta}; N) = \sum_{j=0}^{N-1} \lambda^j [Y_{N-j} - \mathbf{f}^T(-j)\boldsymbol{\theta}]^2$$



WLS formulation

The criterion:

$$S(\boldsymbol{\theta}; N) = \sum_{j=0}^{N-1} \lambda^j [Y_{N-j} - \mathbf{f}^T(-j)\boldsymbol{\theta}]^2$$

can be written as:

$$\begin{bmatrix} Y_1 - \mathbf{f}^T(N-1)\boldsymbol{\theta} \\ Y_2 - \mathbf{f}^T(N-2)\boldsymbol{\theta} \\ \vdots \\ Y_N - \mathbf{f}^T(0)\boldsymbol{\theta} \end{bmatrix}^T \begin{bmatrix} \lambda^{N-1} & 0 & \cdots & 0 \\ 0 & \lambda^{N-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Y_1 - \mathbf{f}^T(N-1)\boldsymbol{\theta} \\ Y_2 - \mathbf{f}^T(N-2)\boldsymbol{\theta} \\ \vdots \\ Y_N - \mathbf{f}^T(0)\boldsymbol{\theta} \end{bmatrix}$$

which is a WLS criterion with $\boldsymbol{\Sigma} = \text{diag}[1/\lambda^{N-1}, \dots, 1/\lambda, 1]$

WLS solution

$$\widehat{\boldsymbol{\theta}}_N = (\mathbf{x}_N^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_N)^{-1} \mathbf{x}_N^T \boldsymbol{\Sigma}^{-1} \mathbf{Y}$$

or

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_N &= \mathbf{F}_N^{-1} \mathbf{h}_N \\ \mathbf{F}_N &= \sum_{j=0}^{N-1} \lambda^j \mathbf{f}(-j) \mathbf{f}^T(-j) \\ \mathbf{h}_N &= \sum_{j=0}^{N-1} \lambda^j \mathbf{f}(-j) Y_{N-j}\end{aligned}$$

Updating the estimates when Y_{N+1} is available

$$\begin{aligned}\mathbf{F}_{N+1} &= \mathbf{F}_N + \lambda^N \mathbf{f}(-N) \mathbf{f}^T(-N) \\ \mathbf{h}_{N+1} &= \lambda \mathbf{L}^{-1} \mathbf{h}_N + \mathbf{f}(0) Y_{N+1} \\ \widehat{\boldsymbol{\theta}}_{N+1} &= \mathbf{F}_{N+1}^{-1} \mathbf{h}_{N+1}\end{aligned}$$

As initial values we can use $\mathbf{h}_0 = \mathbf{0}$ and $\mathbf{F}_0 = \mathbf{0}$

For many functions $\lambda^N \mathbf{f}(-N) \mathbf{f}^T(-N) \rightarrow 0$ for $N \rightarrow \infty$ and we get the stationary result $\mathbf{F}_{N+1} = \mathbf{F}_N = \mathbf{F}$. Hence:

$$\widehat{\boldsymbol{\theta}}_{N+1} = \mathbf{L}^T \widehat{\boldsymbol{\theta}}_N + \mathbf{F}^{-1} \mathbf{f}(0) [Y_{N+1} - \widehat{Y}_{N+1|N}]$$

Variance estimation in local trend models

Define the *total memory*

$$T = \sum_{j=0}^{N-1} \lambda^j = \frac{1 - \lambda^N}{1 - \lambda}$$

T is a measure of how many observations estimation is essentially based upon.

A variance estimator is therefore

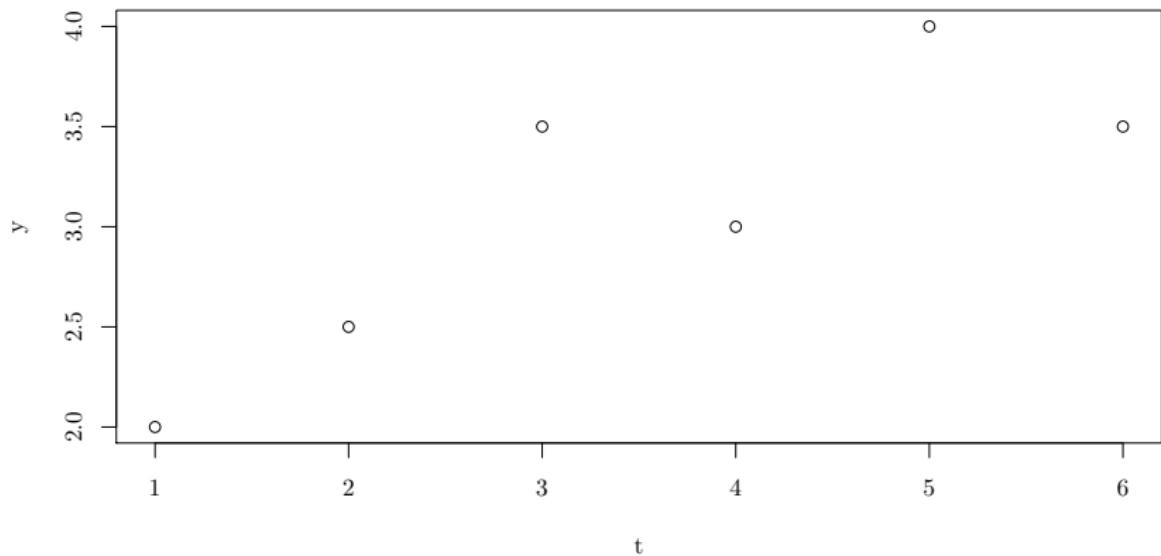
$$\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{x}_N \hat{\boldsymbol{\theta}}_N)^T \Sigma^{-1} (\mathbf{Y} - \mathbf{x}_N \hat{\boldsymbol{\theta}}_N) / (T - p), \quad T > p$$

Notice that the restriction on T is a restriction on λ . How do you interpret this?

(Note: This estimator is not in the book.)

Global and Local Trend Models - an Example

6 observations ($N = 6$):



Global Linear Trend:

$$Y_{N+j} = \theta_0 + \theta_1 j + \varepsilon_{N+j} \Rightarrow \mathbf{f}(j) = (1 \quad j)^T$$

Linear Model form:

$$\begin{pmatrix} 2.0 \\ 2.5 \\ 3.5 \\ 3.0 \\ 4.0 \\ 3.5 \end{pmatrix} = \begin{pmatrix} 1 & -5 \\ 1 & -4 \\ 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix} \Leftrightarrow \mathbf{y} = \mathbf{x}_6 \boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

Global linear trend: Estimation

$$\mathbf{F}_6 = \mathbf{x}_6^T \mathbf{x}_6 = \begin{pmatrix} 6 & -15 \\ -15 & 55 \end{pmatrix}$$

$$\mathbf{h}_6 = \mathbf{x}_6^T \mathbf{y} = \begin{pmatrix} 18.5 \\ -40.5 \end{pmatrix}$$

$$\hat{\boldsymbol{\theta}}_6 = \mathbf{F}_6^{-1} \mathbf{h}_6 = \begin{pmatrix} 0.5238 & 0.1429 \\ 0.1429 & 0.0571 \end{pmatrix} \begin{pmatrix} 18.5 \\ -40.5 \end{pmatrix} = \begin{pmatrix} 3.905 \\ 0.329 \end{pmatrix}$$

Global linear trend: Prediction

Linear predictor:

$$\hat{Y}_{6+\ell|6} = \mathbf{f}(\ell)^T \hat{\theta}_6 = 3.905 + 0.328\ell$$

LS-estimate for σ^2 :

$$\hat{\sigma}^2 = (\mathbf{y} - \mathbf{x}_6 \hat{\theta}_6)^T (\mathbf{y} - \mathbf{x}_6 \hat{\theta}_6) / (6 - 2) = 0.453^2$$

Prediction error:

$$\begin{aligned}\varepsilon_6(\ell) &= Y_{6+\ell} - \hat{Y}_{6+\ell|6} \\ \widehat{\text{Var}}(\varepsilon_6(\ell)) &= \hat{\sigma}^2 (1 + \mathbf{f}^T(\ell) \mathbf{F}_6^{-1} \mathbf{f}(\ell))\end{aligned}$$

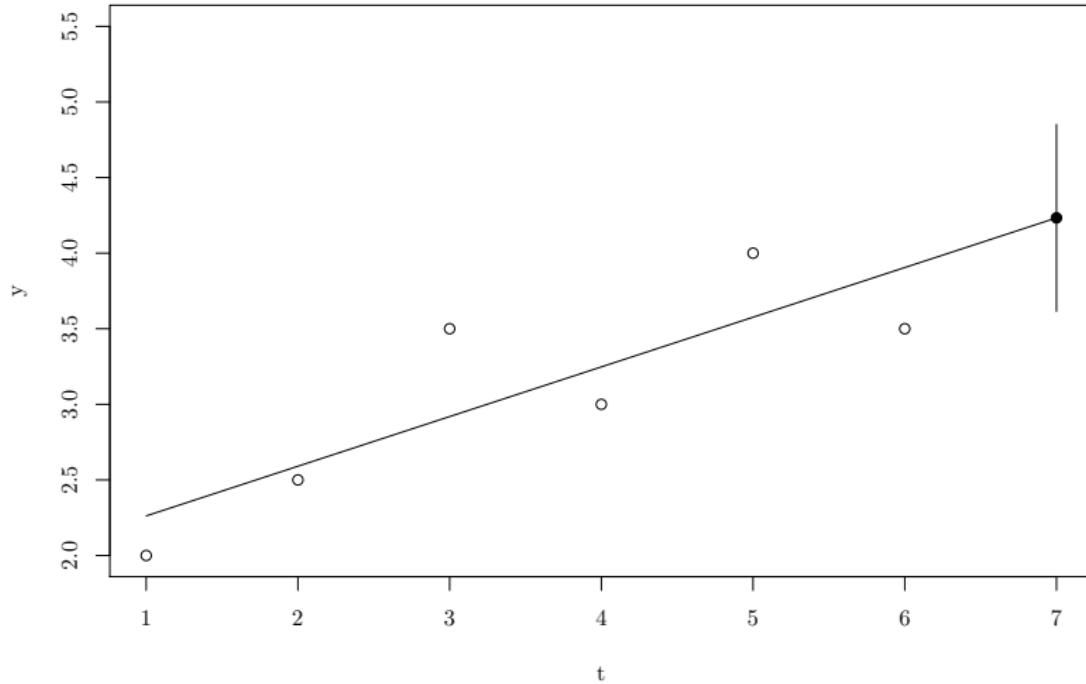
For example,

$$\hat{Y}_{7|6} = 4.234 \text{ with } \widehat{\text{Var}}(\varepsilon_6(1)) = 0.619^2.$$

90% prediction interval:

$$\hat{Y}_{7|6} \pm t_{0.05}(6 - 2) \sqrt{\widehat{\text{Var}}(\varepsilon_6(1))} = 4.234 \pm 1.320$$

Global linear trend: Estimation - global linear trend

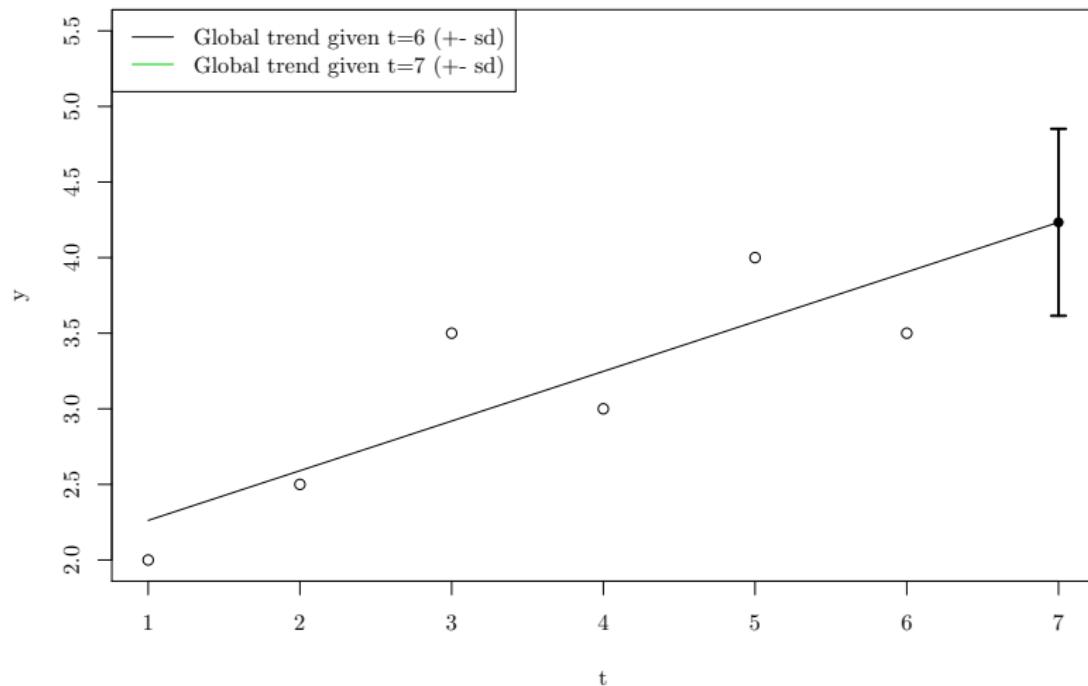


Global linear trend: Updating the parameters

- ▶ New observation: $y_7 = 3.5$.

$$\begin{aligned}\mathbf{F}_7 &= \mathbf{F}_6 + \mathbf{f}^T(-6)\mathbf{f}(-6) \\&= \begin{pmatrix} 6 & -15 \\ -15 & 55 \end{pmatrix} + (1 \quad -6) \begin{pmatrix} 1 \\ -6 \end{pmatrix} = \begin{pmatrix} 7 & -21 \\ -21 & 91 \end{pmatrix}, \\ \mathbf{h}_7 &= L^{-1}\mathbf{h}_6 + \mathbf{f}(0)\mathbf{y}_7 \\&= \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 18.5 \\ -40.5 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} 3.5 = \begin{pmatrix} 22 \\ -59 \end{pmatrix}, \\ \hat{\boldsymbol{\theta}}_7 &= \begin{pmatrix} 0.4643 & 0.1071 \\ 0.1071 & 0.0357 \end{pmatrix} \begin{pmatrix} 22 \\ -59 \end{pmatrix} = \begin{pmatrix} 3.896 \\ 0.250 \end{pmatrix}.\end{aligned}$$

Global linear trend: Updating - global linear trend



Local linear trend: Estimation

- ▶ Forgetting factor $\lambda = 0.9$. Linear model unchanged.

$$\mathbf{F}_6 = \sum_{j=0}^5 \lambda^j \mathbf{f}(-j) \mathbf{f}^T(-j) = \begin{pmatrix} 4.6856 & -10.284 \\ -10.284 & 35.961 \end{pmatrix}$$

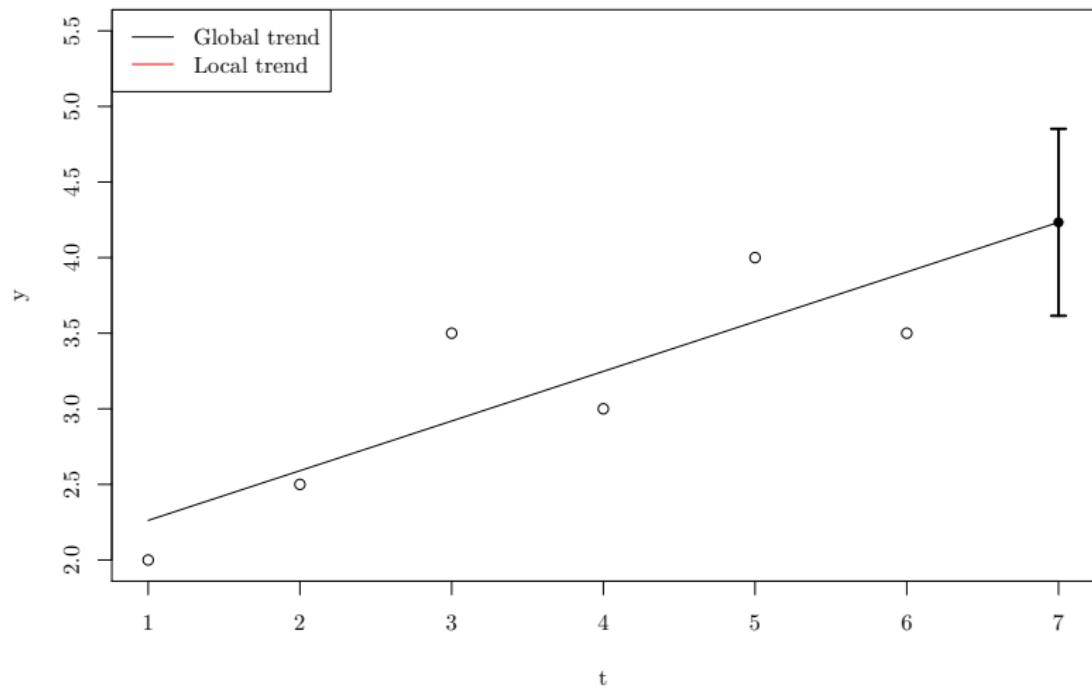
$$\mathbf{h}_6 = \sum_{j=0}^5 \lambda^j \mathbf{f}(-j) Y_{6-j} = \begin{pmatrix} 14.902 \\ -28.580 \end{pmatrix}$$

$$\hat{\boldsymbol{\theta}}_6 = \mathbf{F}_6^{-1} \mathbf{h}_6 = \begin{pmatrix} 0.573 & 0.164 \\ 0.164 & 0.075 \end{pmatrix} \begin{pmatrix} 14.902 \\ -28.580 \end{pmatrix} = \begin{pmatrix} 3.85 \\ 0.308 \end{pmatrix}$$

$$\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{x}_6 \hat{\boldsymbol{\theta}}_6)^T \Sigma^{-1} (\mathbf{Y} - \mathbf{x}_6 \hat{\boldsymbol{\theta}}_6) / (T - 2) = 0.496^2$$

$$\widehat{\text{Var}}(\varepsilon_6(1)) = \hat{\sigma}^2 (1 + \mathbf{f}^T(1) \mathbf{F}_6^{-1} \mathbf{f}(1)) = 0.697^2$$

Local linear trend: Predicting for $t = 7$



Local linear trend: Estimating $\hat{\sigma}^2$

- ▶ We can use the WLS estimator for $\hat{\theta}_N$
- ▶ but not for $\hat{\sigma}^2$!?!
- ▶ Reason: Local trend models assume that ϵ_t are i.i.d.
- ▶ The proposed estimator:

$$\hat{\sigma}_N^2 = (\mathbf{Y} - \mathbf{x}_N \hat{\boldsymbol{\theta}}_N)^T \Sigma^{-1} (\mathbf{Y} - \mathbf{x}_N \hat{\boldsymbol{\theta}}_N) / (T - p), \quad T > p$$

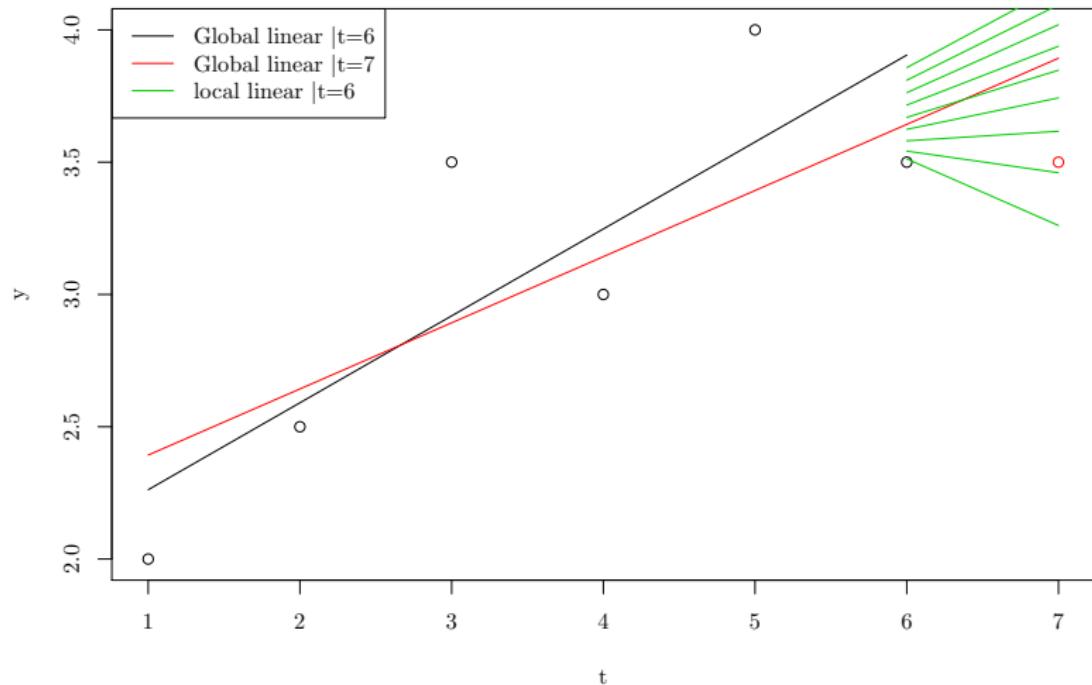
provides a local estimate.

- ▶ A global estimator is given as:

$$\hat{\sigma}_N^2 = \frac{1}{N - n} \sum_{j=n+1}^N \frac{(Y_j - \hat{Y}_{j|j-1})^2}{1 + \mathbf{f}^T(1) \mathbf{F}_{j-1}^{-1} \mathbf{f}(1)}$$

- ▶ Note that the first n predictions are ignored to stabilize the estimator
- ▶ Note that the prediction errors are normed.

Local linear trend: Estimation



Which of the (green) local trend models have the highest λ ?

Operators; The backwards shift operator B

- ▶ An operator A is (here) a function of a time series $\{x_t\}$ (or a stochastic process $\{X_t\}$).
- ▶ Application of an operator on a time series $\{x_t\}$ yields a new time series $\{Ax_t\}$. Likewise of a stochastic process $\{AX_t\}$.
- ▶ Most important operator for us: The backwards shift operator B :
 $Bx_t = x_{t-1}$. Obviously, $B^j x_t = x_{t-j}$.
- ▶ All other operators we shall consider in this lecture may be expressed in terms of B .

The forward shift F and difference ∇

The forward shift operator

- ▶ $Fx_t = x_{t+1}; F^j x_t = x_{t+j}$;
- ▶ Obviously, combining a forward and backward shift yields the identity operator 1, ie. F and B are each others inverse: $B^{-1} = F$ and $F^{-1} = B$.

The difference operator

- ▶ $\nabla x_t = x_t - x_{t-1} = \mathbf{1}x_t - Bx_t = (1 - B)x_t$.
- ▶ Thus: $\nabla = 1 - B$.

The summation S

$$\begin{aligned} Sx_t &= x_t + x_{t-1} + x_{t-2} + \dots \\ &= x_t + Bx_t + B^2 x_t \dots \\ &= (1 + B + B^2 + \dots) x_t \end{aligned}$$

- ▶ Summation, then difference (remember $Sx_t = x_t + Sx_{t-1}$)

$$\nabla Sx_t = Sx_t - Sx_{t-1} = x_t + Sx_{t-1} - Sx_{t-1} = x_t$$

- ▶ Difference, then summation

$$\begin{aligned} S\nabla x_t &= (1 + B + B^2 \dots) x_t - (1 + B + B^2 \dots) x_{t-1} \\ &= (1 + B + B^2 \dots) x_t - (B + B^2 \dots) x_t = x_t \end{aligned}$$

- ▶ So ∇ and S are each others inverse:

$$\nabla^{-1} = \frac{1}{1 - B} = 1 + B + B^2 + \dots = S$$

Properties of B , F , ∇ and S

- ▶ The operators are all linear, ie.

$$H[\lambda x_t + (1 - \lambda)y_t] = \lambda H[x_t] + (1 - \lambda)H[y_t]$$

- ▶ The operators may be combined into new operators:

The power series

$$a(z) = \sum_{i=0}^{\infty} a_i z^i$$

defines a new operator from an operator H by linear combinations:

$$a(H) = \sum_{i=0}^{\infty} a_i H^i$$

Examples of combined operators

- ▶ ∇^{-1} :

$$\frac{1}{1-z} = \sum_{i=0}^{\infty} z^i \quad \text{so} \quad \nabla^{-1} = \frac{1}{1-B} = \sum_{i=0}^{\infty} B^i = S$$

- ▶ Operator polynomial of order q :

$$\theta(z) = \sum_{i=0}^q \theta_i z^i$$

ie. $\theta_i = 0$ for $i > q$.

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

where θ_0 is chosen to be 1

The Cauchy product (discrete convolution)

The equation

$$\{\lambda_i\} * \{\psi_i\} = \{\pi_i\}$$

means that

$$\pi_0 = \lambda_0 \psi_0$$

$$\pi_1 = \lambda_1 \psi_0 + \lambda_0 \psi_1$$

⋮

$$\pi_i = \lambda_i \psi_0 + \lambda_{i-1} \psi_1 + \dots + \lambda_0 \psi_i$$

⋮

Multiplying combined operators

Theorem 4.13

- ▶ For the operator H the following operators are given:

$$\lambda(H) = \sum_{i=0}^{\infty} \lambda_i H^i, \quad \psi(H) = \sum_{i=0}^{\infty} \psi_i H^i, \quad \pi(H) = \sum_{i=0}^{\infty} \pi_i H^i$$

such that $\lambda(H)\psi(H) = \pi(H)$.

- ▶ Then λ, ψ, π satisfies the equation

$$\{\lambda_i\} * \{\psi_i\} = \{\pi_i\}.$$

Highlights

- ▶ Local trend model: $S(\boldsymbol{\theta}; N) = \sum_{j=0}^{N-1} \lambda^j [Y_{N-j} - \mathbf{f}^T(-j)\boldsymbol{\theta}]^2$
- ▶ Iterative updates

$$\begin{aligned}\mathbf{F}_N &= \sum_{j=0}^{N-1} \lambda^j \mathbf{f}(-j) \mathbf{f}^T(-j) \\ \mathbf{h}_N &= \sum_{j=0}^{N-1} \lambda^j \mathbf{f}(-j) Y_{N-j} \\ \widehat{\boldsymbol{\theta}}_N &= \mathbf{F}_N^{-1} \mathbf{h}_N\end{aligned}$$

- ▶ Backwards shift operator: $BX_t = X_{t-1}$
- ▶ Difference operator: $\nabla X_t = X_t - X_{t-1}$

Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

September 28, 2017

Outline of the lecture

- ▶ Stochastic processes, 1st part:
 - ▶ Stochastic processes in general: Sec 5.1, 5.2, 5.3 [except 5.3.2], 5.4.
 - ▶ MA, AR, and ARMA-processes, Sec. 5.5
 - ▶ Non-stationary models, Sec. 5.6
 - ▶ Optimal Prediction, Sec. 5.7

Stochastic Processes – in general

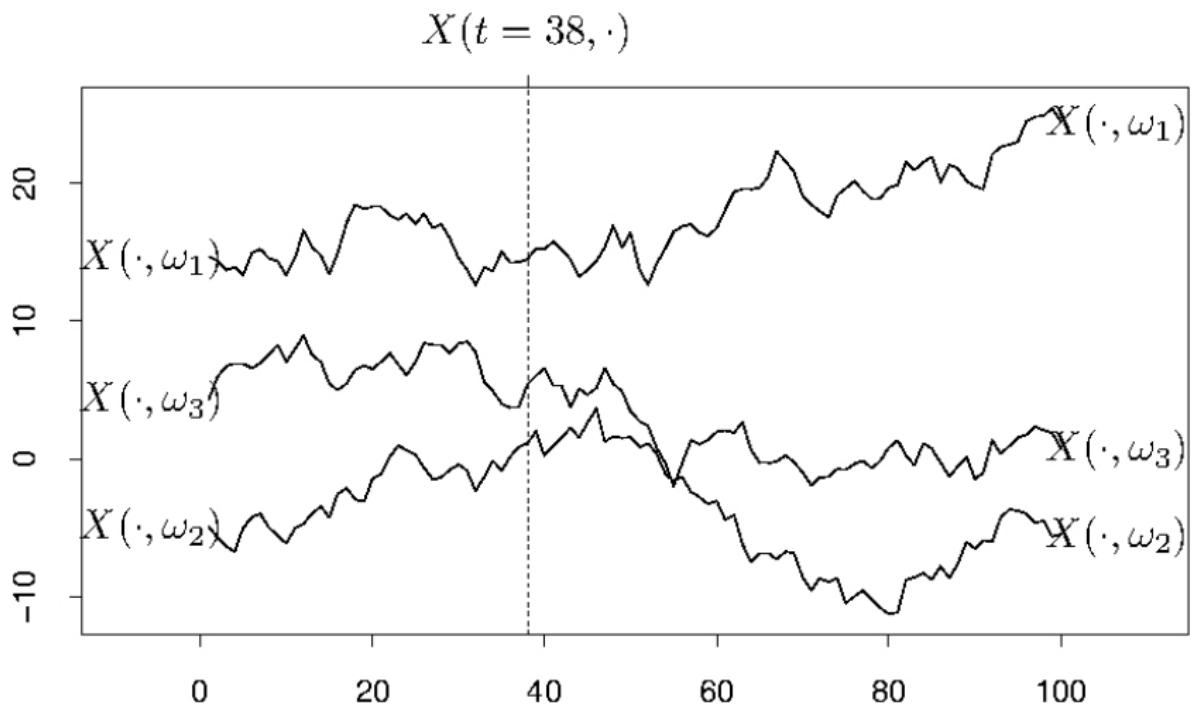
- ▶ Function: $X(t, \omega)$
- ▶ Time: $t \in T$
- ▶ Realization: $\omega \in \Omega$

- ▶ Index set: T
- ▶ Sample Space: Ω

- ▶ $X(t = t_0, \cdot)$ is a random variable
- ▶ $X(\cdot, \omega)$ is a time series (i.e. *one realization*). This is what we often denote $\{x_t\}$.

- ▶ In this course we consider the case where time is discrete, and the realizations take values on the real numbers (continuous range).

Stochastic Processes – illustration



Complete Characterization

n -dimensional probability density:

$$f_{X(t_1), \dots, X(t_n)}(x_1, \dots, x_n)$$

Family of probability density functions, i.e.:

- ▶ For all $n = 1, 2, 3, \dots$
- ▶ and all t

is called the *family of finite-dimensional probability density functions (pdf's) for the process*. This family completely characterize the stochastic process.

2nd order moment representation

Mean function:

$$\mu(t) = E[X(t)] = \int_{-\infty}^{\infty} x f_{X(t)}(x) dx,$$

Autocovariance function:

$$\begin{aligned}\gamma_{XX}(t_1, t_2) &= \gamma(t_1, t_2) = \text{Cov}[X(t_1), X(t_2)] \\ &= E[(X(t_1) - \mu(t_1))(X(t_2) - \mu(t_2))]\end{aligned}$$

The variance function is obtained from $\gamma(t_1, t_2)$ when $t_1 = t_2 = t$:

$$\sigma^2(t) = V[X(t)] = E[(X(t) - \mu(t))^2]$$

Stationarity

- ▶ A process $\{X(t)\}$ is said to be *strongly stationary* if all finite-dimensional distributions are invariant for changes in time, i.e. for every n , and for any set (t_1, t_2, \dots, t_n) and for any h it holds

$$f_{X(t_1), \dots, X(t_n)}(x_1, \dots, x_n) = f_{X(t_1+h), \dots, X(t_n+h)}(x_1, \dots, x_n)$$

- ▶ A process $\{X(t)\}$ is said to be *weakly stationary of order k* if all the first k moments are invariant to changes in time
- ▶ A weakly stationary process of order 2 is simply called *weakly stationary* or just *stationary*:

$$\mu(t) = \mu \quad \sigma^2(t) = \sigma^2 \quad \gamma(t_1, t_2) = \gamma(t_1 - t_2)$$

Ergodicity

- ▶ In time series analysis we normally assume that we have access to one realization only
- ▶ We therefore need to be able to determine characteristics of the random variable X_t from one realization x_t
- ▶ It is often enough to require the process to be mean-ergodic:

$$E[X(t)] = \int_{\Omega} x(t, \omega) f(\omega) d\omega = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t, \omega) dt$$

i.e. if the *mean of the ensemble* equals the *mean over time*

Some intuitive examples, not directly related to time series:

<http://news.softpedia.com/news/What-is-ergodicity-15686.shtml>

Special processes

- ▶ *Normal processes* (also called *Gaussian processes*): All finite-dimensional distribution functions are (multivariate) normal distributions
- ▶ *Markov processes*: The conditional distribution depends only on the latest state of the process:

$$P\{X(t_n) \leq x | X(t_{n-1}), \dots, X(t_1)\} = P\{X(t_n) \leq x | X(t_{n-1})\}$$

- ▶ *Deterministic processes*: Can be predicted without uncertainty from past observations
- ▶ *Pure stochastic processes*: Can be written as a (infinite) linear combination of uncorrelated random variables
- ▶ Decomposition: $X_t = S_t + D_t$

Autocovariance and autocorrelation

- ▶ For stationary processes: Only dependent on the time difference
 $\tau = t_2 - t_1$
- ▶ Autocovariance:

$$\gamma(\tau) = \gamma_{XX}(\tau) = \text{Cov}[X(t), X(t + \tau)] = E[X(t)X(t + \tau)]$$

- ▶ Autocorrelation:

$$\rho(\tau) = \rho_{XX}(\tau) = \gamma_{XX}(\tau)/\gamma_{XX}(0) = \gamma_{XX}(\tau)/\sigma_X^2$$

- ▶ Some properties of the autocovariance function:
 - ▶ $\gamma(\tau) = \gamma(-\tau)$
 - ▶ $|\gamma(\tau)| \leq \gamma(0)$

Linear processes

- ▶ A linear process $\{Y_t\}$ is a process that can be written on the form

$$Y_t - \mu = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$$

where μ is the mean value of the process and

- ▶ $\{\varepsilon_t\}$ is white noise, i.e. a sequence of uncorrelated, identically distributed random variables.
- ▶ $\{\varepsilon_t\}$ can be scaled so that $\psi_0 = 1$
- ▶ Without loss of generality we assume $\mu = 0$

ψ - and π -weights

- ▶ Transfer function and linear process:

$$\psi(B) = 1 + \sum_{i=1}^{\infty} \psi_i B^i \quad Y_t = \psi(B)\varepsilon_t$$

- ▶ Inverse operator (if it exists) and the linear process:

$$\pi(B) = 1 + \sum_{i=1}^{\infty} \pi_i B^i \quad \pi(B)Y_t = \varepsilon_t,$$

- ▶ Autocovariance using ψ -weights:

$$\gamma(k) = \text{Cov} \left[\sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}, \sum_{i=0}^{\infty} \psi_i \varepsilon_{t+k-i} \right] = \sigma_{\varepsilon}^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}$$

Stationarity and invertibility

- ▶ The linear process $Y_t = \psi(B)\varepsilon_t$ is *stationary* if

$$\psi(z) = \sum_{i=0}^{\infty} \psi_i z^{-i}$$

converges for $|z| \geq 1$ (i.e. old values of ε_t are down-weighted)

- ▶ The linear process $\pi(B)Y_t = \varepsilon_t$ is said to be *invertible* if

$$\pi(z) = \sum_{i=0}^{\infty} \pi_i z^{-i}$$

converges for $|z| \geq 1$ (i.e. ε_t can be calculated from recent values of Y_t)

Linear process as a statistical model?

$$Y_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \psi_3 \varepsilon_{t-3} + \dots$$

- ▶ Observations: $Y_1, Y_2, Y_3, \dots, Y_N$
- ▶ Task: Find an infinite number of parameters from N observations!
- ▶ Solution: Restrict the sequence $1, \psi_1, \psi_2, \psi_3, \dots$

$MA(q)$, $AR(p)$, and $ARMA(p, q)$ processes

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

$$Y_t + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} = \varepsilon_t$$

$$Y_t + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

$\{\varepsilon_t\}$ is white noise

$$Y_t = \theta(B) \varepsilon_t$$

$$\phi(B) Y_t = \varepsilon_t$$

$$\phi(B) Y_t = \theta(B) \varepsilon_t$$

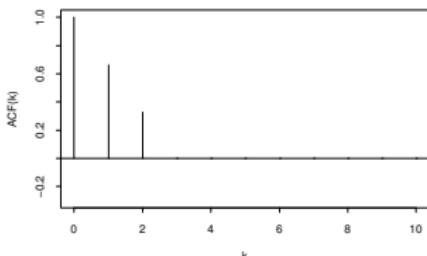
$\phi(B)$ and $\theta(B)$ are polynomials in the backward shift operator B ,
 $(BX_t = X_{t-1}, B^2X_t = X_{t-2})$

Stationarity and invertibility

- ▶ $MA(q)$
 - ▶ Always stationary
 - ▶ Invertible if the roots in $\theta(z^{-1}) = 0$ with respect to z all are within the unit circle
- ▶ $AR(p)$
 - ▶ Always invertible
 - ▶ Stationary if the roots of $\phi(z^{-1}) = 0$ with respect to z all lie within the unit circle
- ▶ $ARMA(p, q)$
 - ▶ Stationary if the roots of $\phi(z^{-1}) = 0$ with respect to z all lie within the unit circle
 - ▶ Invertible if the roots in $\theta(z^{-1}) = 0$ with respect to z all are within the unit circle

Autocorrelations

MA(2)

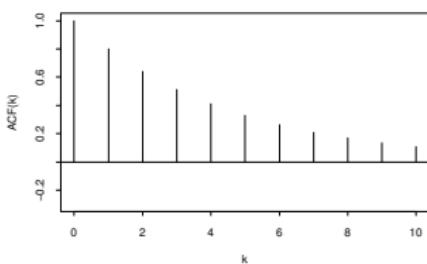


MA(2):

$$Y_t = (1 + 0.9B + 0.8B^2)\varepsilon_t$$

zero after lag 2

AR(1)

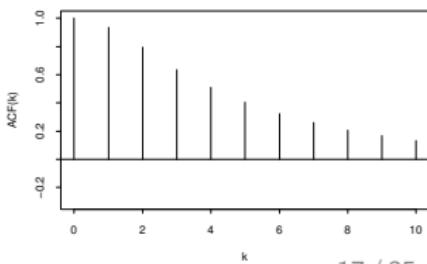


AR(1):

$$(1 - 0.8B) Y_t = \varepsilon_t$$

exponential decay (damped sine in case of complex roots)

ARMA(1,2)



ARMA(1,2):

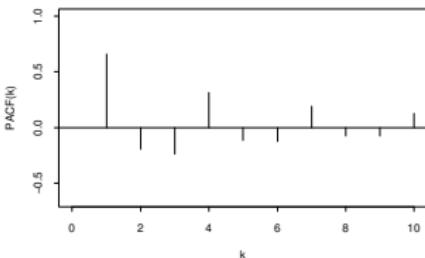
$$(1 - 0.8B) Y_t = (1 + 0.9B + 0.8B^2)\varepsilon_t$$

exponential decay from lag $q+1-p = 2+1-1 = 2$

(damped sine in case of complex roots)

Partial autocorrelations (Appendix B)

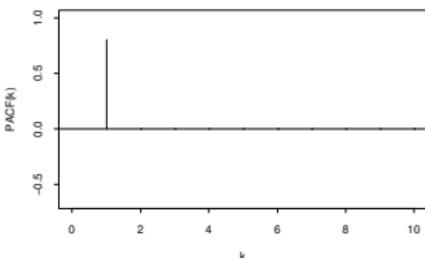
MA(2)



MA(2):

$$Y_t = (1 + 0.9B + 0.8B^2)\varepsilon_t$$

AR(1)

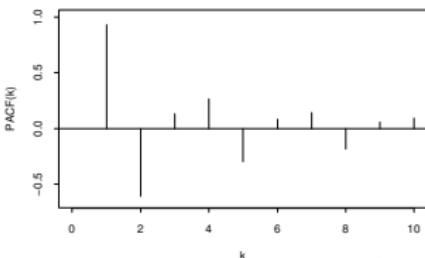


AR(1):

$$(1 - 0.8B)Y_t = \varepsilon_t$$

zero after lag 1

ARMA(1,2)



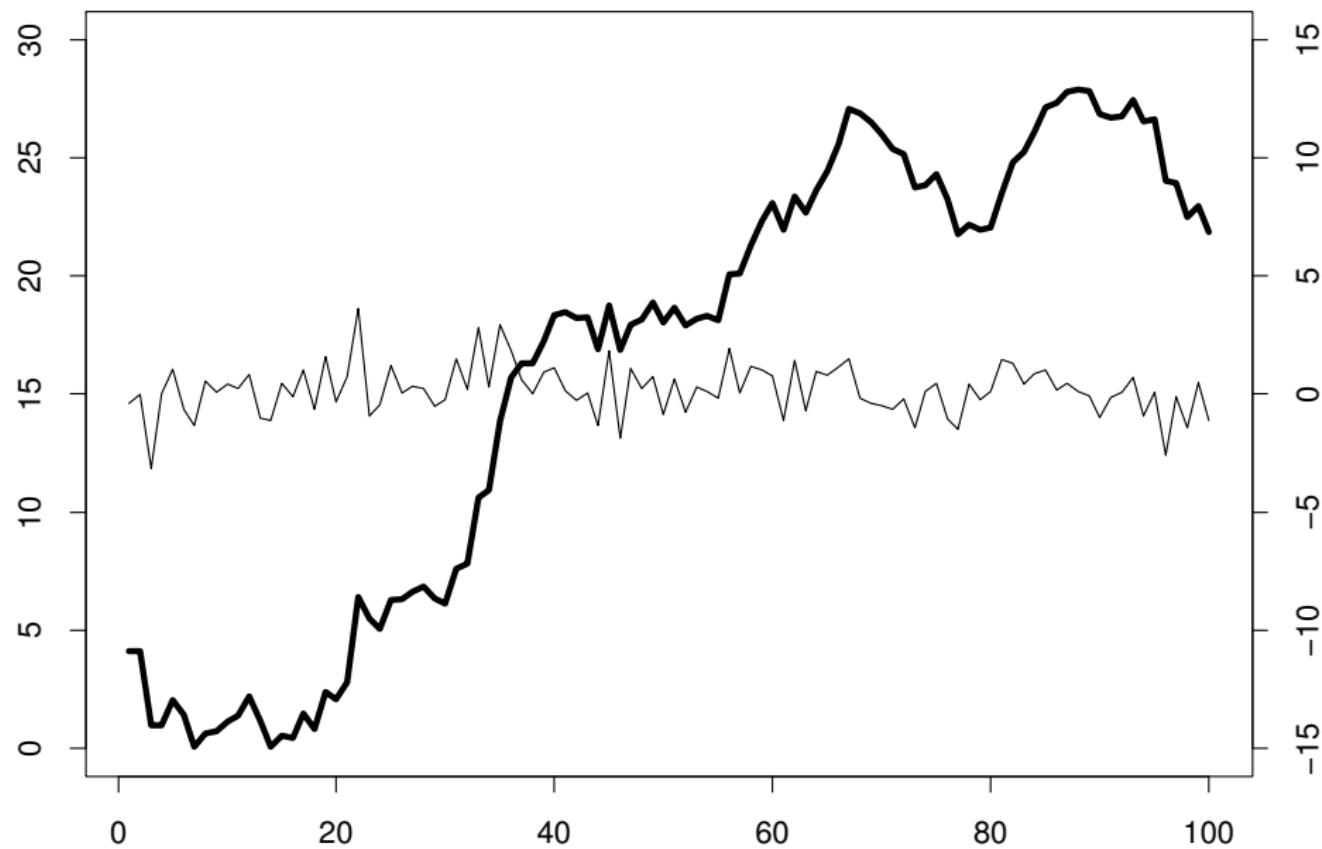
ARMA(1,2):

$$(1 - 0.8B)Y_t = (1 + 0.9B + 0.8B^2)\varepsilon_t$$

Inverse autocorrelation

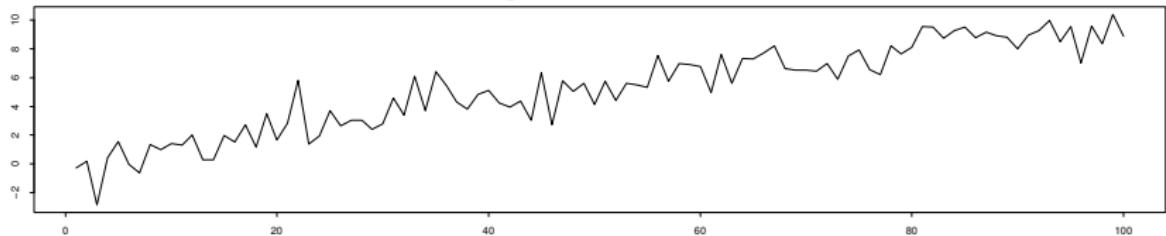
- ▶ The process: $\phi(B)Y_t = \theta(B)\varepsilon_t$
- ▶ The dual process: $\theta(B)Z_t = \phi(B)\varepsilon_t$
- ▶ The inverse autocorrelation is the autocorrelation for the dual process
- ▶ Thus, the IACF can be used in a similar way as the PACF.

Non-stationary series

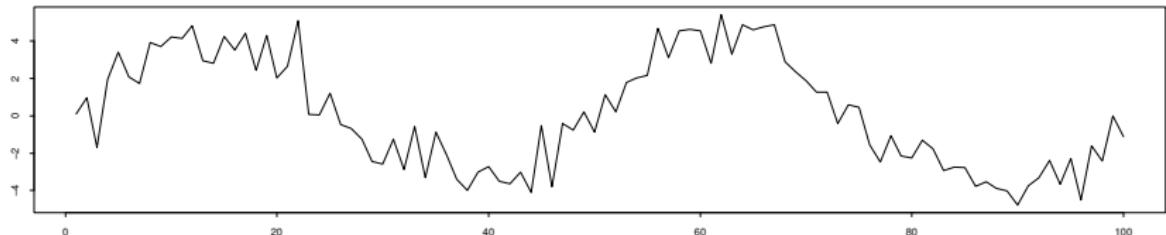


Some types of non-stationarity

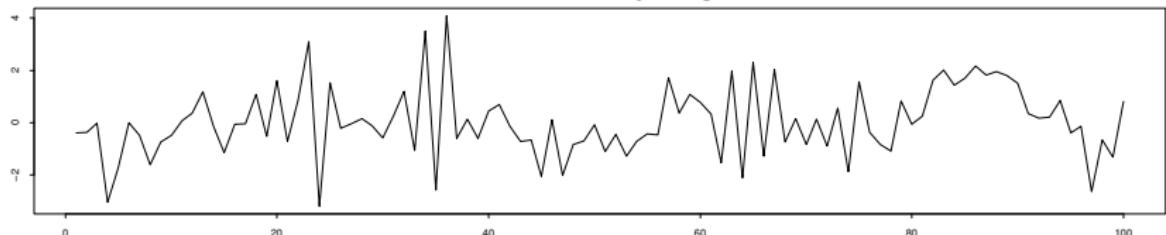
Long term trends



Periodic trends



General time varying behavior



The $ARIMA(p, d, q)$ -process

- ▶ An $ARMA(p, q)$ model for:

$$W_t = \nabla^d Y_t = (1 - B)^d Y_t$$

where $\{Y_t\}$ is the series

- ▶ That is:

$$\phi(B)\nabla^d Y_t = \theta(B)\varepsilon_t$$

- ▶ If we consider stationarity:

$$\phi(z^{-1})(1 - z^{-1})^d = 0$$

i.e. d roots in $z = 1 + 0i$, and the rest inside the unit circle

The $(p, d, q) \times (P, D, Q)_s$ seasonal process

- ▶ A multiplicative (stationary) $ARMA(p, q)$ model for:

$$W_t = \nabla^d \nabla_s^D Y_t = (1 - B)^d (1 - B^s)^D Y_t$$

where $\{Y_t\}$ is the series

- ▶ That is:

$$\phi(B)\Phi(B^s)\nabla^d \nabla_s^D Y_t = \theta(B)\Theta(B^s)\varepsilon_t$$

- ▶ If we consider stationarity:

$$\phi(z^{-1})\Phi(z^{-s})(1 - z^{-1})^d(1 - z^{-s})^D = 0$$

i.e. d roots in $z = 1 + 0i$, $D \times s$ roots on the unit circle, and the rest inside the unit circle

The case $d = D = 0$; stationary seasonal process

- ▶ General:

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)\varepsilon_t$$

- ▶ Example:

$$(1 - \Phi B^{12}) Y_t = \varepsilon_t$$

- ▶ Which can also be written:

$$Y_t = \Phi Y_{t-12} + \varepsilon_t$$

i.e. Y_t depend on Y_{t-12} , Y_{t-24} , ... (thereof the name)

- ▶ How would you think that the auto correlation function looks?
- ▶ Take a look at Example 5.10 also.

Prediction

- ▶ At time t we have observations $Y_t, Y_{t-1}, Y_{t-2}, Y_{t-3}, \dots$
- ▶ We want a prediction of Y_{t+k} , where $k \geq 1$
- ▶ If we want to minimize the expected squared error the optimal prediction is the conditional expectation:

$$\hat{Y}_{t+k|t} = E[Y_{t+k} | Y_t, Y_{t-1}, Y_{t-2}, \dots]$$

Example – prediction in the $AR(1)$ model

- ▶ We write the model like $Y_{t+1} = \phi Y_t + \varepsilon_{t+1}$ (note the sign on ϕ)
- ▶ 1-step prediction:

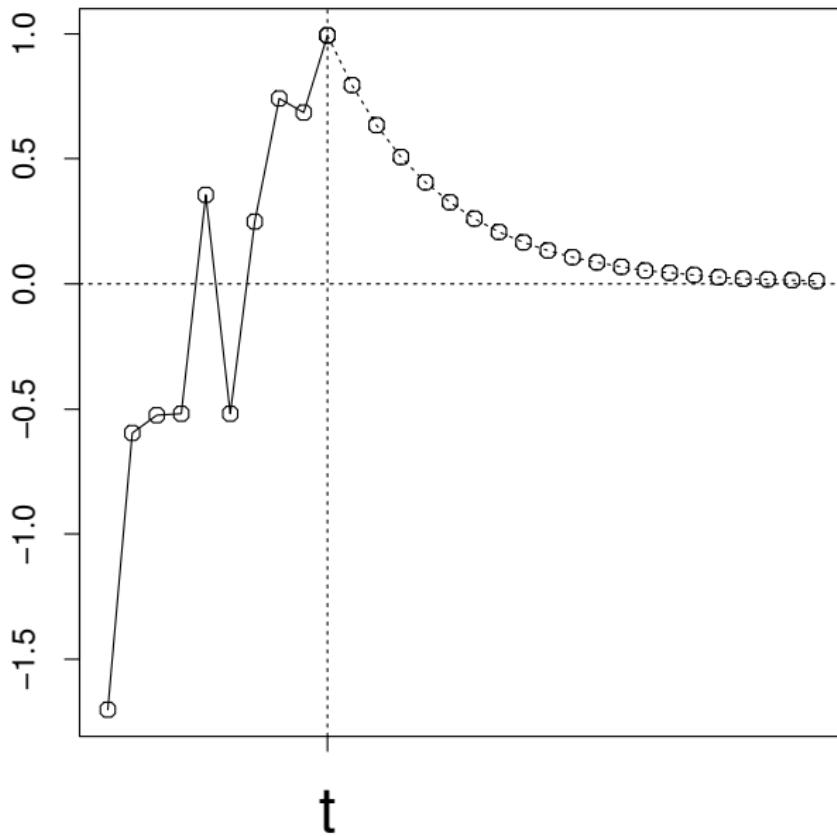
$$\begin{aligned}\hat{Y}_{t+1|t} &= E[Y_{t+1} | Y_t, Y_{t-1}, \dots] = E[\phi Y_t + \varepsilon_{t+1} | Y_t, Y_{t-1}, \dots] \\ &= \phi Y_t + 0 = \phi Y_t\end{aligned}$$

- ▶ 2-step prediction:

$$\begin{aligned}\hat{Y}_{t+2|t} &= E[Y_{t+2} | Y_t, Y_{t-1}, \dots] = E[\phi Y_{t+1} + \varepsilon_{t+2} | Y_t, Y_{t-1}, \dots] \\ &= \phi \hat{Y}_{t+1|t} + 0 = \phi^2 Y_t\end{aligned}$$

- ▶ k-step prediction: $\boxed{\hat{Y}_{t+k|t} = \phi^k Y_t}$

Example – prediction in $Y_t = 0.8 Y_{t-1} + \varepsilon_t$



Variance of prediction error for the $AR(1)$ -process

Prediction error:

$$e_{t+k|t} = Y_{t+k} - \hat{Y}_{t+k|t} = Y_{t+k} - \phi^k Y_t$$

$$\begin{aligned} Y_{t+k} &= \phi Y_{t+k-1} + \varepsilon_{t+k} \\ &= \phi(\phi Y_{t+k-2} + \varepsilon_{t+k-1}) + \varepsilon_{t+k} \\ &= \phi^2 Y_{t+k-2} + \phi \varepsilon_{t+k-1} + \varepsilon_{t+k} \\ &= \phi^2(\phi Y_{t+k-3} + \varepsilon_{t+k-2}) + \phi \varepsilon_{t+k-1} + \varepsilon_{t+k} \\ &= \phi^3 Y_{t+k-3} + \phi^2 \varepsilon_{t+k-2} + \phi \varepsilon_{t+k-1} + \varepsilon_{t+k} \\ &\vdots \\ &= \phi^k Y_t + \phi^{k-1} \varepsilon_{t+1} + \phi^{k-2} \varepsilon_{t+2} + \dots + \phi \varepsilon_{t+k-1} + \varepsilon_{t+k} \end{aligned}$$

Variance of prediction error for the $AR(1)$ -process

Variance of prediction error:

$$\begin{aligned} V[e_{t+k|t}] &= V[\phi^{k-1}\varepsilon_{t+1} + \phi^{k-2}\varepsilon_{t+2} + \dots + \phi\varepsilon_{t+k-1} + \varepsilon_{t+k}] \\ &= (\phi^{2(k-1)} + \phi^{2(k-2)} + \dots + \phi^2 + 1)\sigma_\varepsilon^2 \end{aligned}$$

$(1 - \alpha) \times 100\%$ prediction interval:

$$\hat{Y}_{t+k|t} \pm u_{\alpha/2} \sqrt{V[e_{t+k|t}]}$$

$u_{\alpha/2}$ is the $\alpha/2$ -quantile in the standard normal distribution

k -step prediction in $ARMA(p, q)$ -models

We assume that $k > \max(p, q)$. The process:

$$Y_{t+k} + \phi_1 Y_{t+k-1} + \cdots + \phi_p Y_{t+k-p} = \\ \varepsilon_{t+k} + \theta_1 \varepsilon_{t+k-1} + \cdots + \theta_q \varepsilon_{t+k-q}$$

Using conditional expectation on both sides we get:

$$\widehat{Y}_{t+k|t} = -\phi_1 \widehat{Y}_{t+k-1|t} - \cdots - \phi_p \widehat{Y}_{t+k-p|t} \\ + \widehat{\varepsilon}_{t+k|t} + \theta_1 \widehat{\varepsilon}_{t+k-1|t} + \cdots + \theta_q \widehat{\varepsilon}_{t+k-q|t}$$

This results in a recursive method for calculating the predictions – how would you find $\widehat{\varepsilon}_{t+k-q|t}$?

Inverse form

For an invertible process, the π -weights

$$\varepsilon_t = Y_t + \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \dots$$

goes to zero sufficiently fast and only recent values of the process is needed.

Variance of prediction error

Process written with ψ -weights:

$$Y_{t+k} = \varepsilon_{t+k} + \psi_1 \varepsilon_{t+k-1} + \cdots + \psi_k \varepsilon_t + \psi_{k+1} \varepsilon_{t-1} + \cdots$$

k -step prediction:

$$\hat{Y}_{t+k|t} = \psi_k \varepsilon_t + \psi_{k+1} \varepsilon_{t-1} + \cdots$$

k -step prediction error:

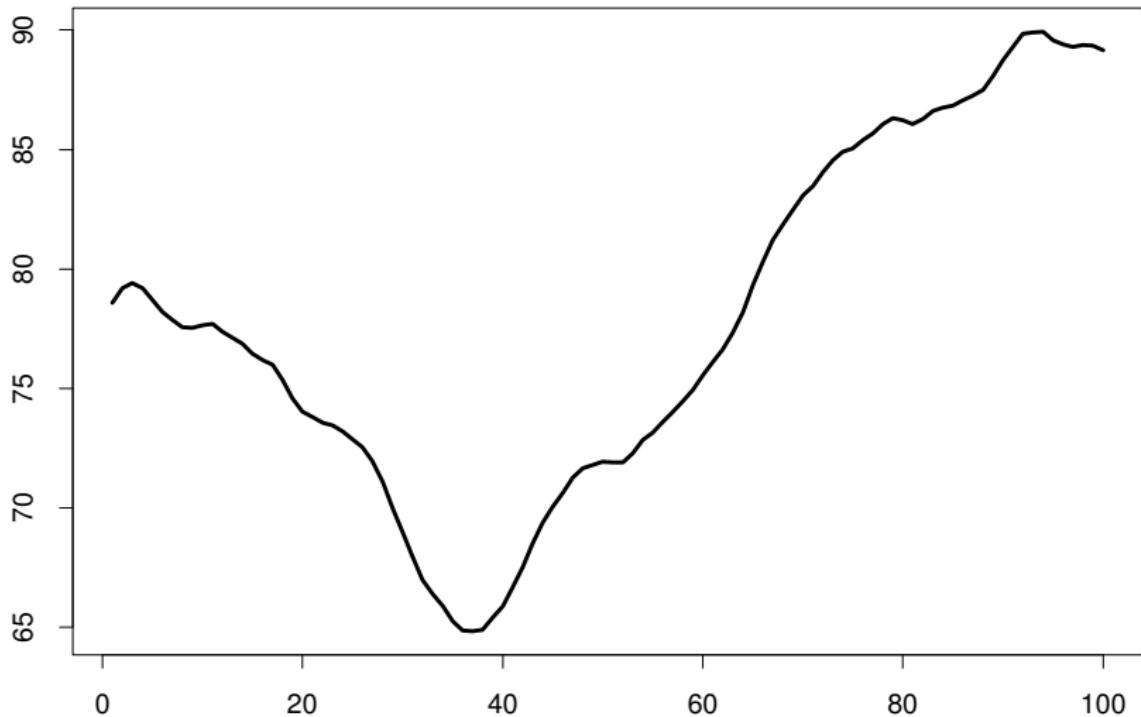
$$e_{t+k|t} = Y_{t+k} - \hat{Y}_{t+k|t} = \varepsilon_{t+k} + \psi_1 \varepsilon_{t+k-1} + \cdots + \psi_{k-1} \varepsilon_{t+1}$$

Variance of k -step prediction error:

$$\sigma_k^2 = (1 + \psi_1^2 + \cdots + \psi_{k-1}^2) \sigma_\varepsilon^2$$

Prediction of bond prices

$$(1 - 1.274B + 0.3867B^2)\nabla Y_t = \varepsilon_t; \quad \sigma_\varepsilon^2 = 0.201^2; \quad \mu_Y = 84.3$$



Prediction of bond prices

- ▶ Price last 6 days: . . . , 90.79, 89.90, 88.88, 87.98, 87.41, 87.16
- ▶ Prediction of price in two days:

$$\begin{aligned}\varphi(B) &= \phi(B)\nabla = (1 + \phi_1 B + \phi_2 B^2)(1 - B) \\ &= 1 + (\phi_1 - 1)B + (\phi_2 - \phi_1)B^2 - \phi_2 B^3\end{aligned}$$

$$\hat{Y}_{t+1|t} = (1 - \phi_1) Y_t + (\phi_1 - \phi_2) Y_{t-1} + \phi_2 Y_{t-2} = 87.06$$

$$\hat{Y}_{t+2|t} = (1 - \phi_1) \hat{Y}_{t+1|t} + (\phi_1 - \phi_2) Y_t + \phi_2 Y_{t-1} = \underline{\underline{87.03}}$$

- ▶ Variance of prediction error:

$$V[\varepsilon_{t+2} + (1 - \phi_1)\varepsilon_{t+1}] = (1 + (1 - \phi_1)^2)\sigma_\varepsilon^2 = 0.499^2$$

- ▶ 95% prediction interval: $87.03 \pm 1.96 \cdot 0.50 = [86.05; 88.01]$

Highlights

- ▶ Stochastic process $X(t, \omega)$
 - ▶ $X(t = t_0, \cdot)$ is a random variable
 - ▶ $X(\cdot, \omega)$ is a time series (i.e. *one* realization).
- ▶ Stationarity
- ▶ Autocovariance:

$$\begin{aligned}\gamma_{XX}(t_1, t_2) &= \gamma(t_1, t_2) = \text{Cov}[X(t_1), X(t_2)] \\ &= E[(X(t_1) - \mu(t_1))(X(t_2) - \mu(t_2))]\end{aligned}$$

- ▶ $MA(q)$, $AR(p)$, and $ARMA(p, q)$ processes.
- ▶ $ARIMA(p, d, q) \times (P, D, Q)_s$ for seasonal and non stationary processes.
- ▶ The optimal prediction is the conditional expectation:

$$\hat{Y}_{t+k|t} = E[Y_{t+k} | Y_t, Y_{t-1}, Y_{t-2}, \dots]$$

Time Series Analysis

Lasse Engbo Christiansen

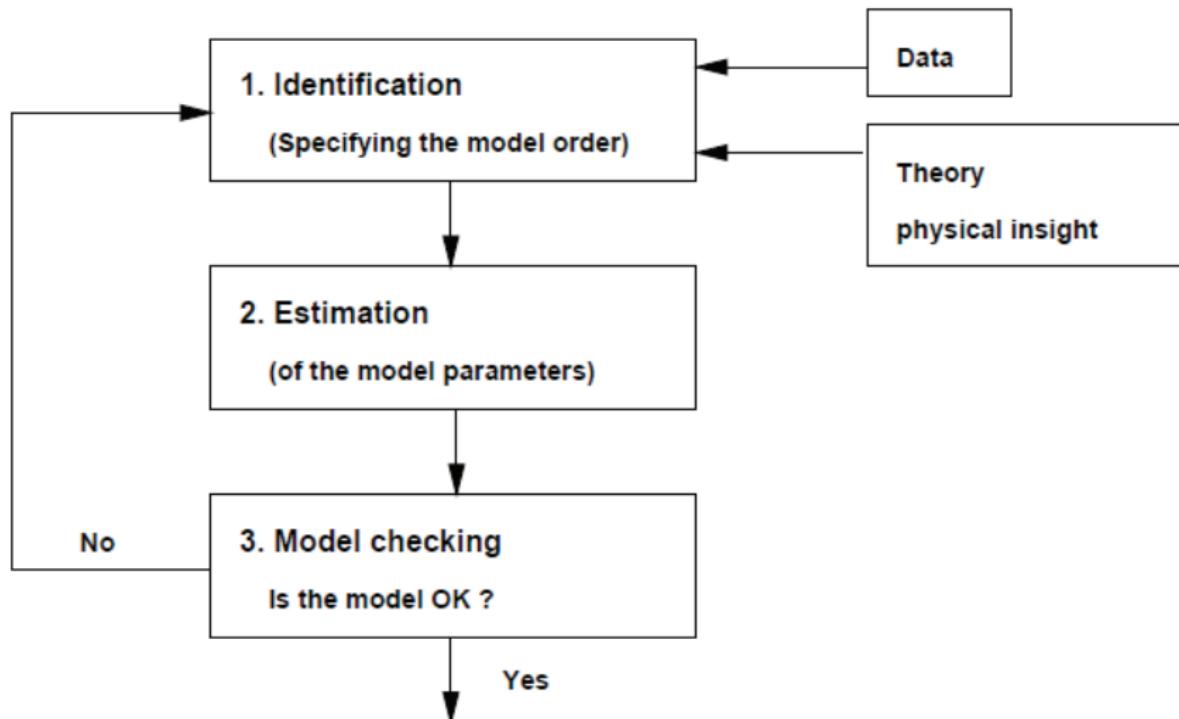
Department of Applied Mathematics and Computer Science
Technical University of Denmark

October 3, 2017

Outline of the lecture

- ▶ Identification of univariate time series models, 1st part:
 - ▶ Introduction, Sec. 6.1
 - ▶ Estimation of auto-covariance and -correlation, Sec. 6.2.1 (and the intro. to 6.2)
 - ▶ Using the SACF and SPACF for model order selection
 - ▶ Model order selection, Sec. 6.5
 - ▶ Model validation, Sec. 6.6

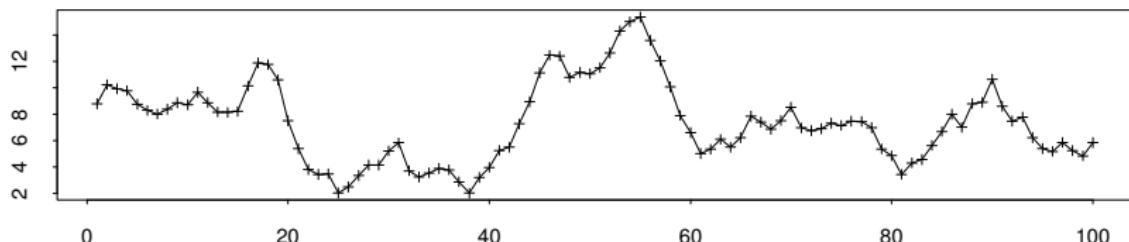
Model building in general



Applications using the model
(Prediction, simulation, etc.)

Identification of univariate time series models

- ▶ What ARIMA structure would be appropriate for the data at hand?
(If any)



- ▶ Given the structure we will then consider how to estimate the parameters (later)
- ▶ What do we know about ARIMA models which could help us?

Estimation of the autocovariance function

- ▶ Estimate of $\gamma(k)$

$$C_{YY}(k) = C(k) = \hat{\gamma}(k) = \frac{1}{N} \sum_{t=1}^{N-|k|} (Y_t - \bar{Y})(Y_{t+|k|} - \bar{Y})$$

- ▶ It is enough to consider $k > 0$
- ▶ R: `acf(x, type = "covariance")`

Some properties of $C(k)$

- ▶ The estimator is *non-central*:

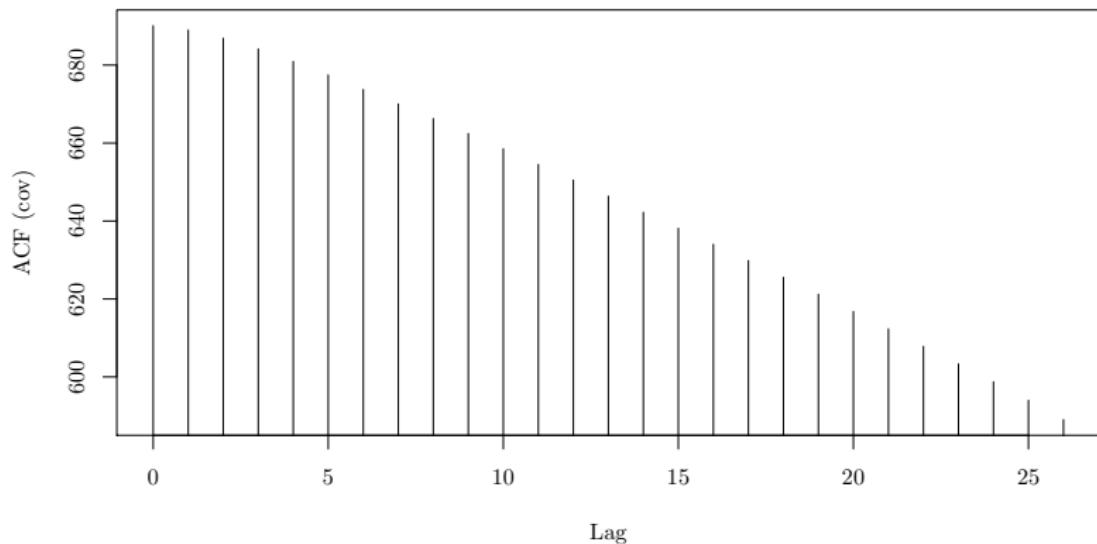
$$E[C(k)] = \frac{1}{N} \sum_{t=1}^{N-|k|} \gamma(k) = \left(1 - \frac{|k|}{N}\right) \gamma(k)$$

- ▶ Asymptotically central (consistent) for fixed k :
 $E[C(k)] \rightarrow \gamma(k)$ for $N \rightarrow \infty$
- ▶ The estimates are correlated themselves (don't trust apparent correlation at random high lags too much)

How does $C(k)$ behave for non-stationary series?

$$C(k) = \frac{1}{N} \sum_{t=1}^{N-|k|} (Y_t - \bar{Y})(Y_{t+|k|} - \bar{Y})$$

Series arima.sim(model = list(ar = 0.6, order = c(1, 1, 0)), n = 500)



Autocorrelation and Partial Autocorrelation

Autocorrelation

- ▶ Sample autocorrelation function (SACF): $\hat{\rho}(k) = r_k = C(k)/C(0)$
- ▶ For white noise and $k \neq 0$ it holds that $E[\hat{\rho}(k)] \simeq 0$ and $V[\hat{\rho}(k)] \simeq 1/N$, this gives the bounds $\pm 2/\sqrt{N}$ for deciding when it is not possible to distinguish a value from zero.
- ▶ R: `acf(x)`

Partial autocorrelation

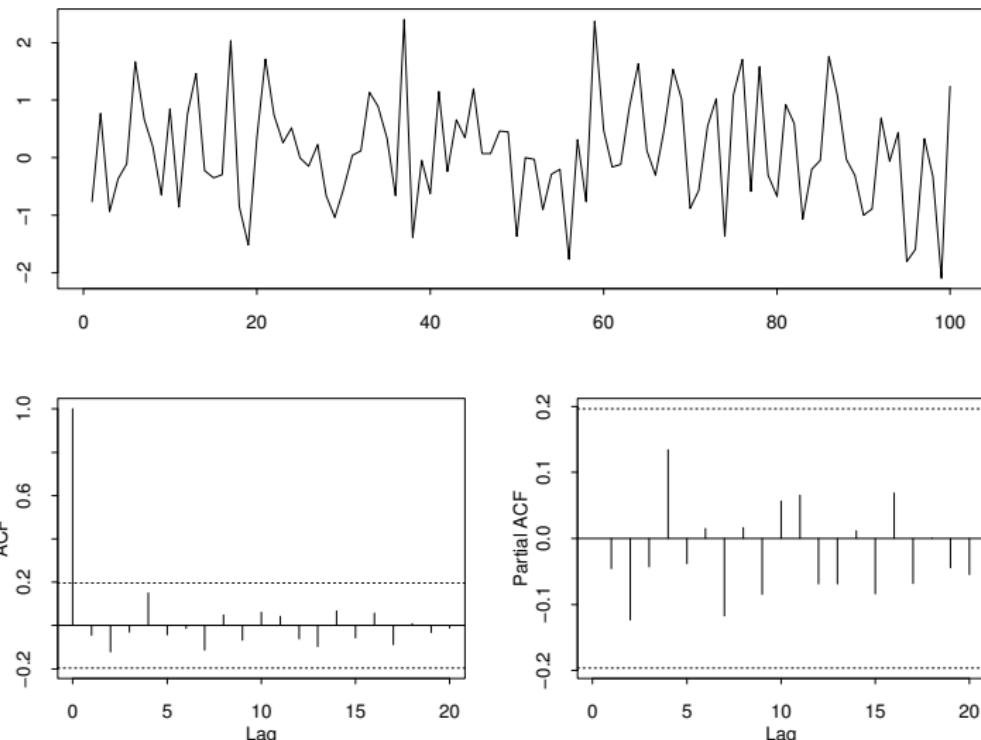
- ▶ Sample partial autocorrelation function (SPACF): Use the Yule-Walker equations on $\hat{\rho}(k)$ (exactly as for the theoretical relations Eq.(5.81))
- ▶ It turns out that $\pm 2/\sqrt{N}$ is also appropriate for deciding when the SPACF is zero (more in the next lecture)
- ▶ R: `acf(x, type="partial")` or `pacf(x, type="partial")`

The golden table for ARMA identification

(Table 6.1)

	ACF $\rho(k)$	PACF ϕ_{kk}
AR(p)	Damped exponential and/or sine functions	$\phi_{kk} = 0$ for $k > p$
MA(q)	$\rho(k) = 0$ for $k > q$	Dominated by damped exponential and or/sine functions
ARMA(p, q)	Damped exponential and/or sine functions after lag $\max(0, q - p)$	Dominated by damped exponential and/or sine functions after lag $\max(0, p - q)$

What would be an appropriate structure?



1: White noise

4: MA(1)

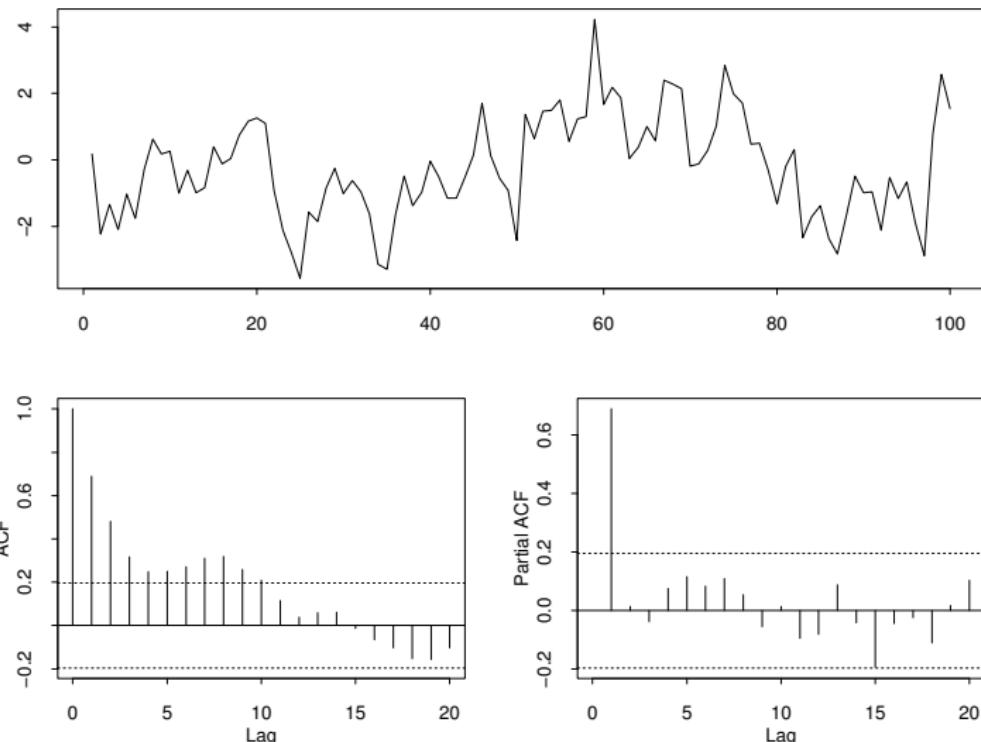
2: AR(1)

5: MA(2)

3: AR(2)

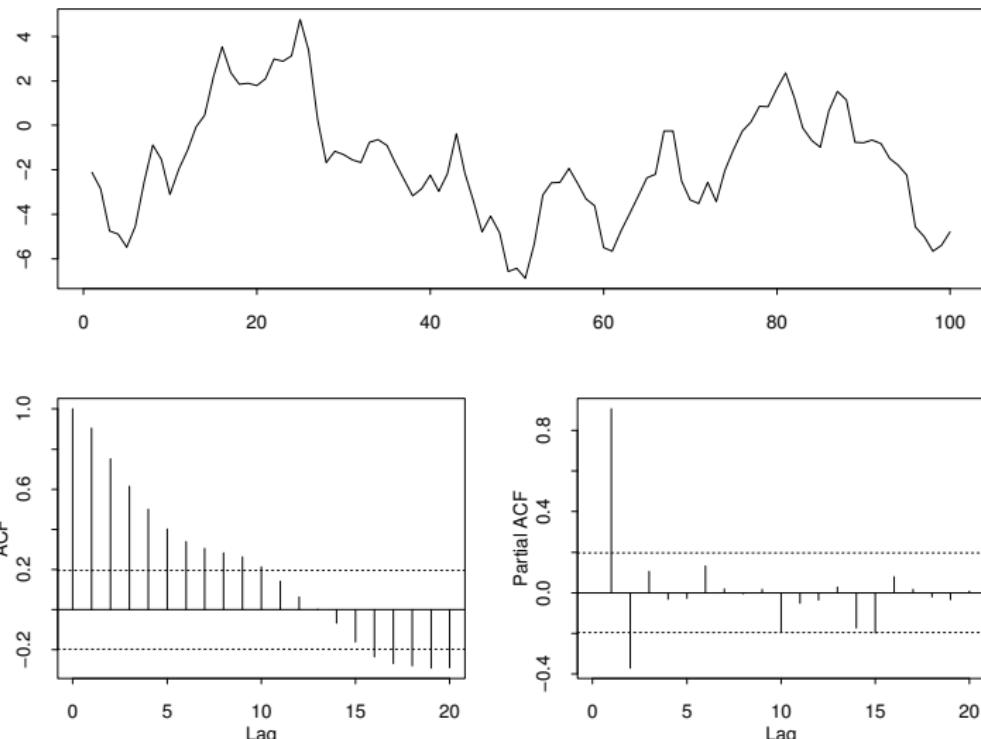
6: ARMA(1,1)

What would be an appropriate structure?



- 1: White noise
- 2: AR(1)
- 3: AR(2)
- 4: MA(1)
- 5: AR(8)
- 6: ARMA(1,1)

What would be an appropriate structure?



1: White noise

4: MA(1)

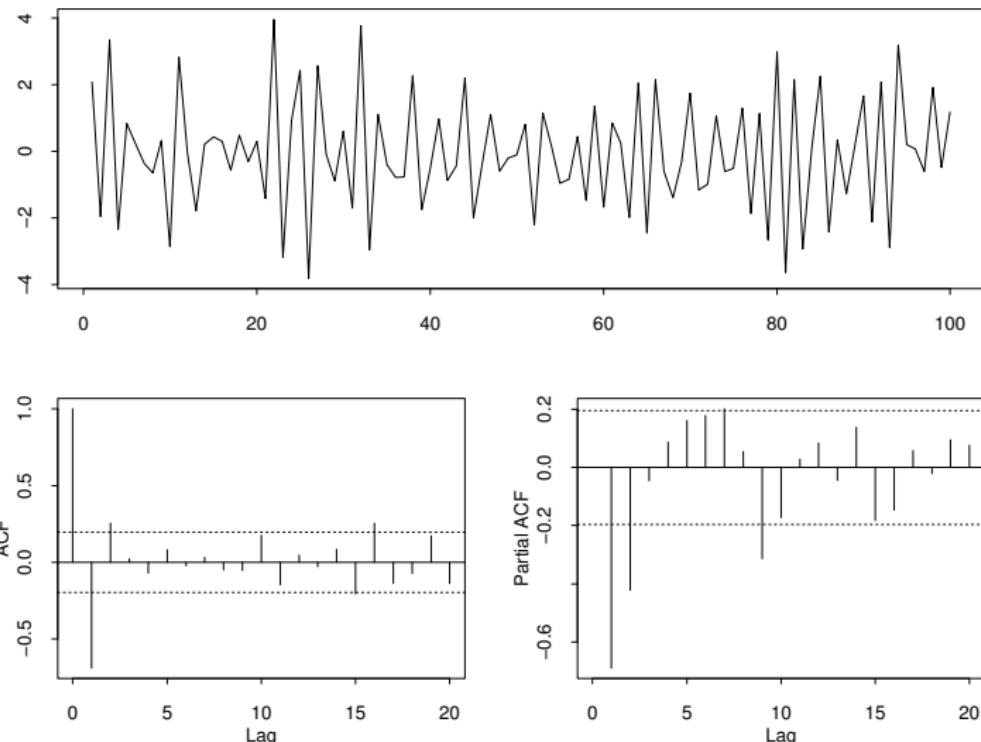
2: AR(1)

5: MA(2)

3: AR(2)

6: ARMA(1,1)

What would be an appropriate structure?



1: White noise

4: MA(1)

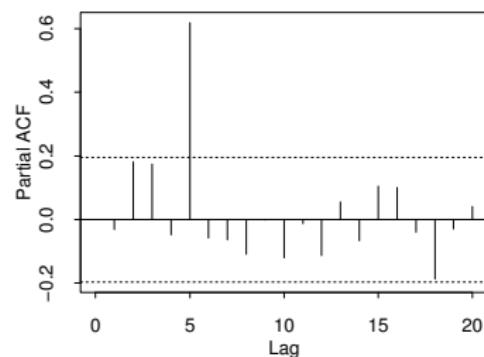
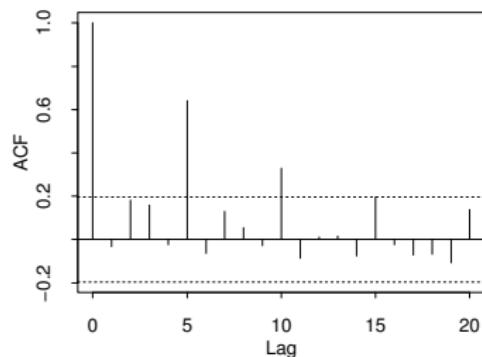
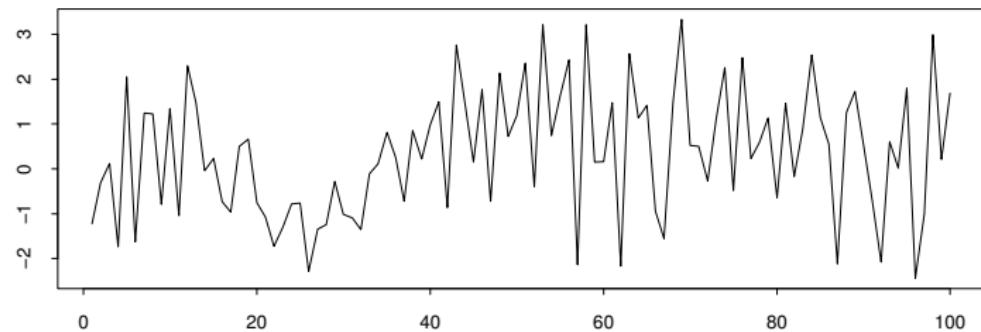
2: AR(1)

5: MA(2)

3: AR(2)

6: ARMA(1,1)

What would be an appropriate structure?



1: AR(1)

4: Seasonal AR(1) S=4

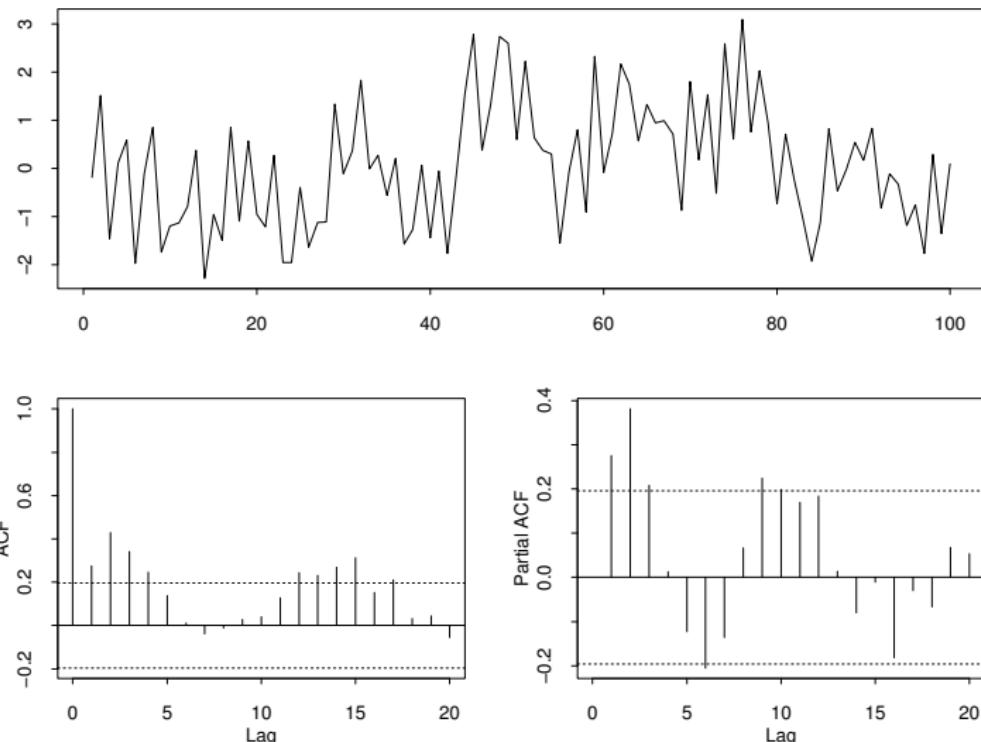
2: AR(5)

5: Seasonal AR(1) S=5

3: Seasonal AR(2) S=5

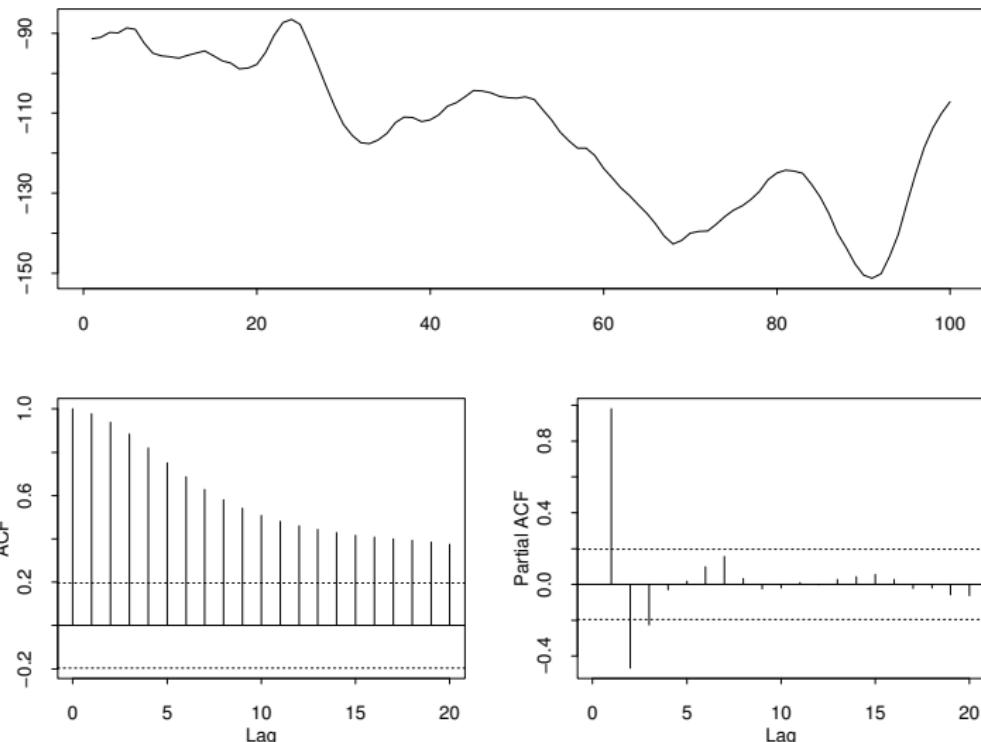
6: AR(10)

What would be an appropriate structure?



- 1: AR(4)
- 2: AR(5)
- 3: MA(4)
- 4: ARMA(1,1)
- 5: AR(2)
- 6: ARMA(2,2)

What would be an appropriate structure?



1: AR(3)

4: ARMA(1,1)

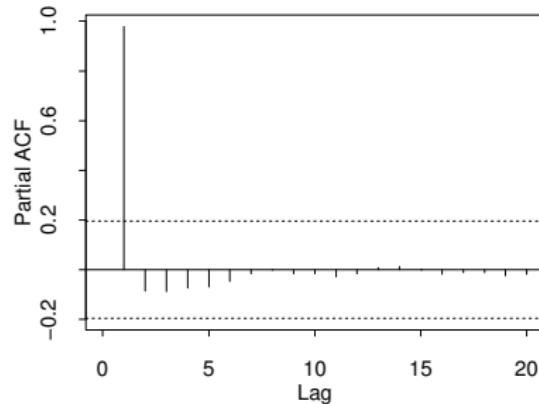
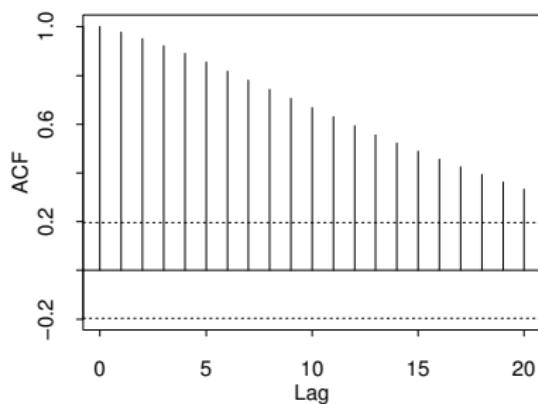
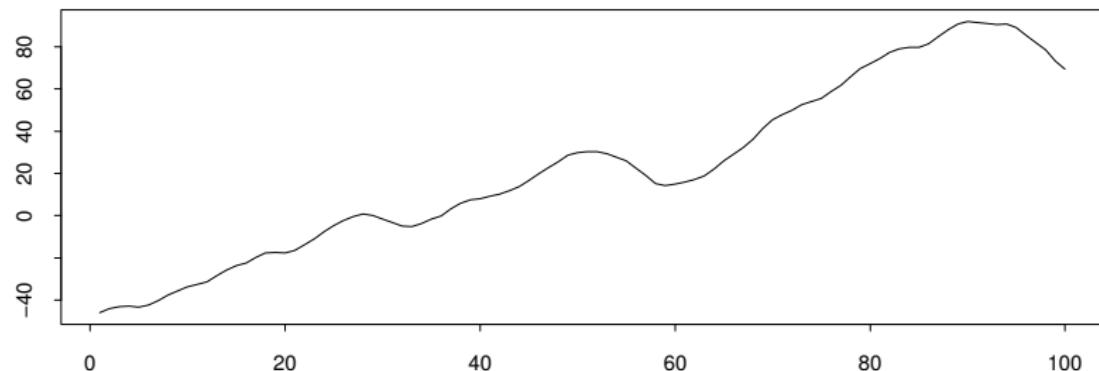
2: AR(2)

5: ARIMA(2,1,0)

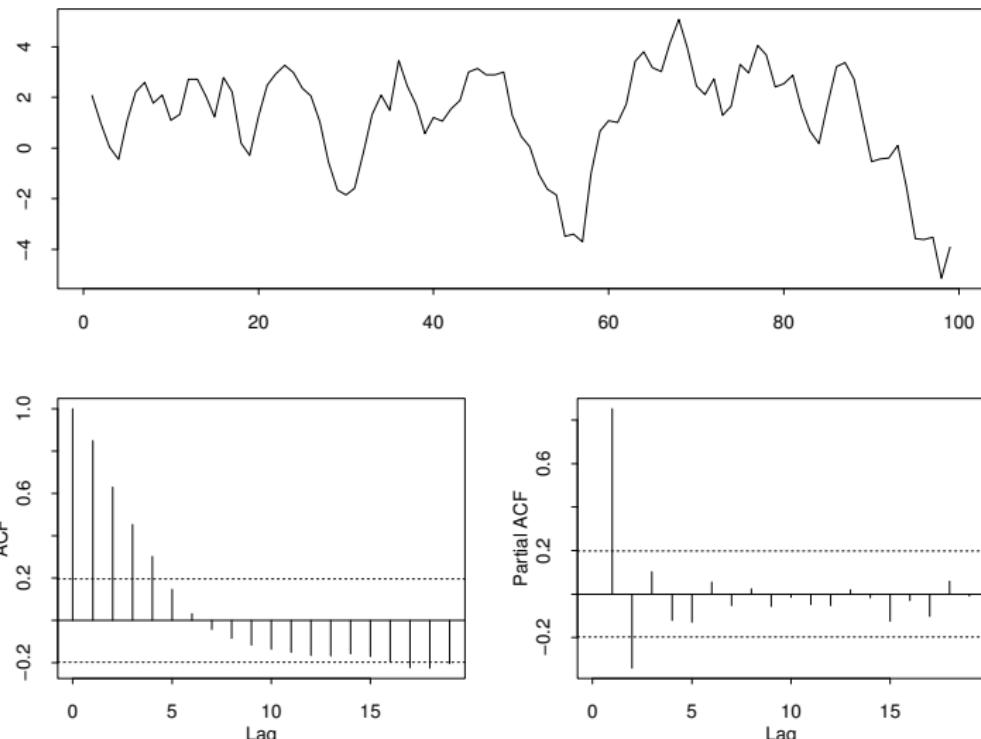
3: ARIMA(0,1,0)

6: ARIMA(0,2,2)

Example of data from a non-stationary process



Same series; analysing $\nabla Y_t = (1 - B) Y_t = Y_t - Y_{t-1}$



1: AR(3)

4: ARMA(1,1)

2: AR(2)

5: ARIMA(2,1,0)

3: ARIMA(0,1,0)

6: ARIMA(0,2,2)

Identification of the order of differencing

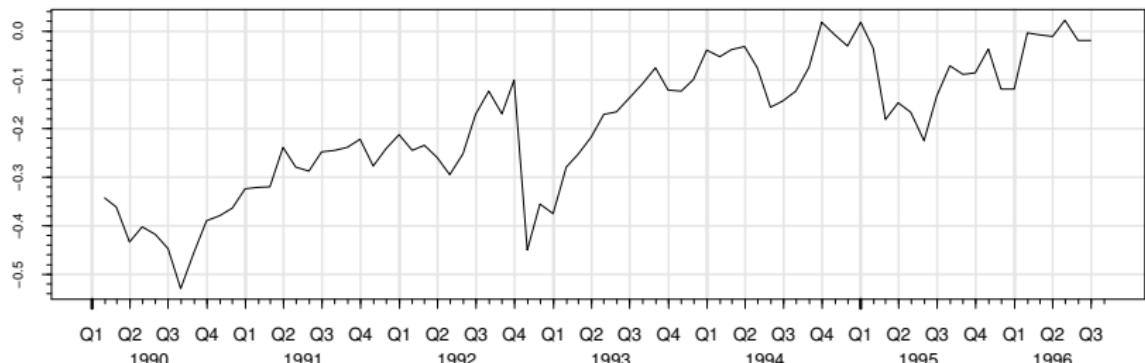
- ▶ Select the order of differencing d as the first order for which the autocorrelation decreases sufficiently fast towards 0
- ▶ In practice d is 0, 1, or maybe 2
- ▶ Sometimes a periodic difference is required, e.g. $Y_t - Y_{t-12}$
- ▶ Remember to consider the practical application . . . it may be that the system is stationary, but you measured over a too short period

Stationarity vs. length of measuring period

US/CA 30 day interest rate differential



US/CA 30 day interest rate differential



Selection of the Model Order

- ▶ The model order of an ARMA process model:
The number of parameters for the AR and MA part; (p, q).
- ▶ The autocorrelation functions can be used - as we just did
- ▶ If that method fails to identify (p, q) because the process:
 - ▶ Is not a standard AR-proces (the table should work directly);
 - ▶ is not a standard MA-proces (the table should work directly);
 - ▶ is not a directly identifiable ARMA proces;
- ▶ Then one must do something else...
- ▶ Try a small model and reconsider
- ▶ Consider transformations ...
Typically sqrt, log, square, inverse ...

Iterative model building

1. (Identification step): Construct a model for your data:

$$\phi(B) Y_t = \theta(B) W_t$$

2. (Estimation step): Estimate the parameters and calculate the model residuals $W(\hat{\phi}, \hat{\theta})$
3. (Model checking step):
 - ▶ Are the estimated parameters significant?
 - ▶ Does $W(\hat{\phi}, \hat{\theta})$ resemble white noise?
 - ▶ If so, the model can be described by the ϕ and θ polynomials.
 - ▶ If the model residuals do not resemble white noise, then what do they look like?

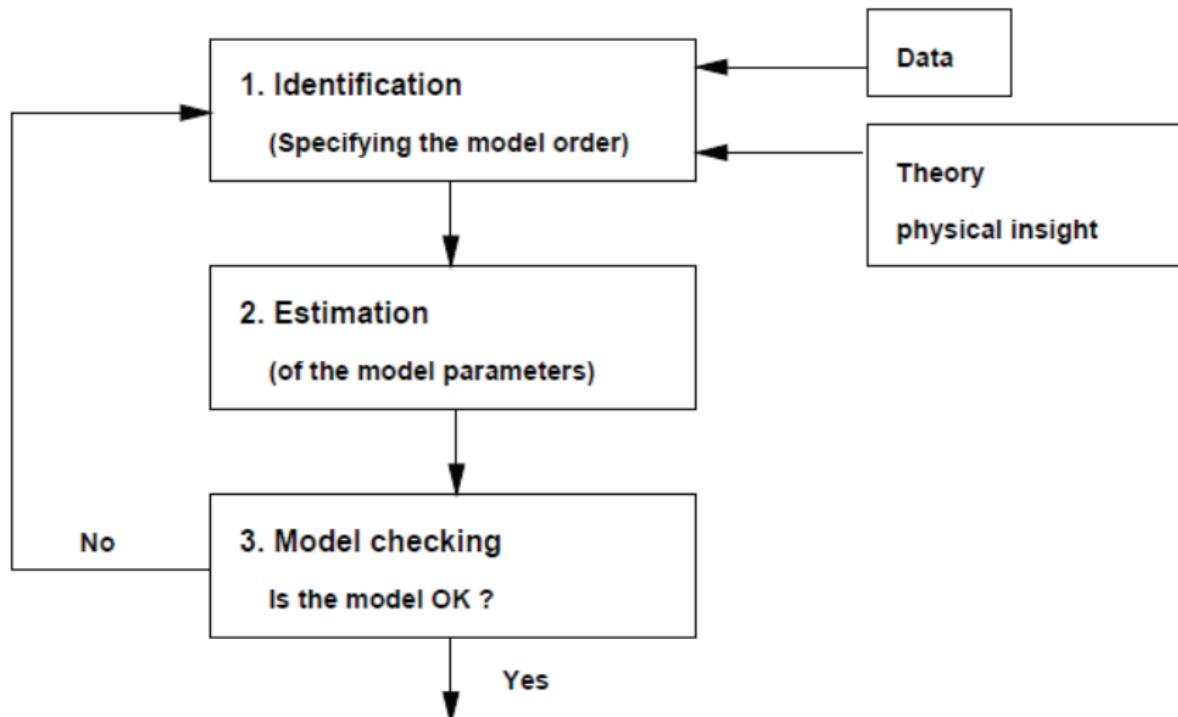
Iterative model building II

- ▶ $W(\hat{\phi}, \hat{\theta})$ will often have a simpler behavior than Y , if the original model $\phi(B)Y_t = \theta(B)W_t$ captures the essential terms of Y 's behavior.
1. Construct an ARMA description for $W(\hat{\phi}, \hat{\theta})$: $\phi^*(B)W_t = \theta^*(B)\varepsilon_t$.
 2. Insert $W_t = \phi^{*-1}(B)\theta^*(B)\varepsilon_t$ into the original model to obtain the model

$$\phi^*(B)\phi(B)Y_t = \theta(B)\theta^*(B)\varepsilon_t$$

3. Estimate the parameters in the model above with coefficients in $\phi^* \cdot \phi, \theta \cdot \theta^*$ varying freely, and proceed to model check.

Model building in general



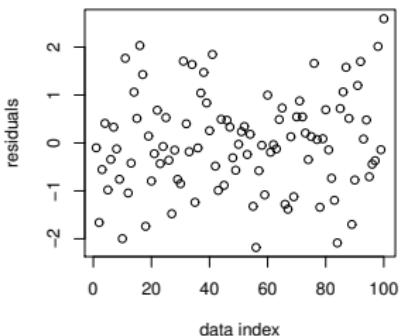
Applications using the model
(Prediction, simulation, etc.)

Residual Analysis

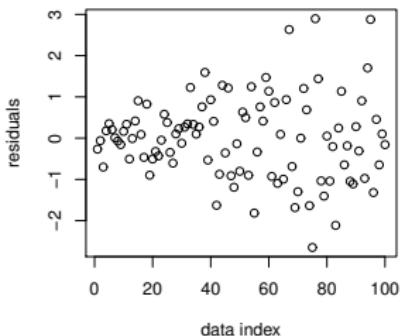
- ▶ The order of the model is decided when the model errors resemble white noise.
- ▶ Is the model order a uniquely determined set of numbers (p,q) ?
NO !
- ▶ How can we check that the model errors resemble white noise?
- ▶ First and most important - plot the data.

Residual analysis – Plot the data

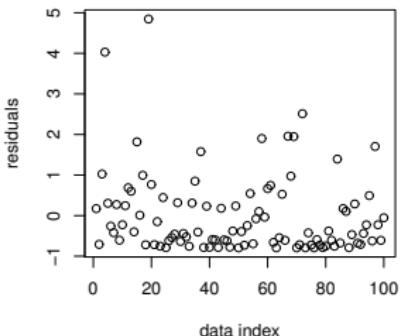
White noise



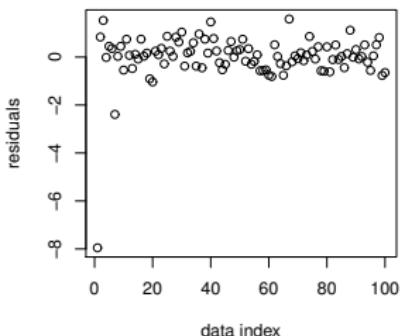
Not white noise



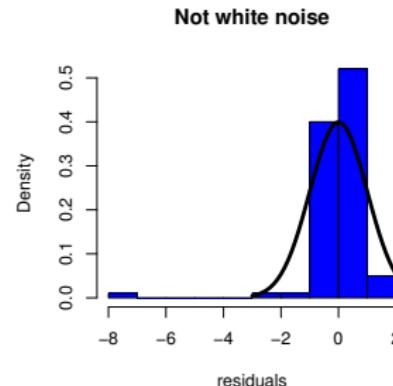
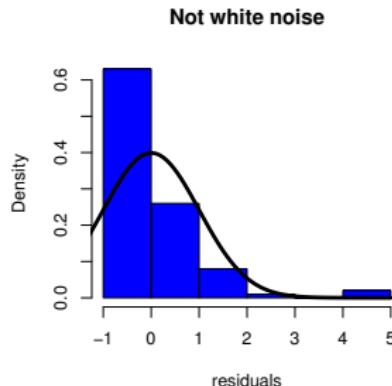
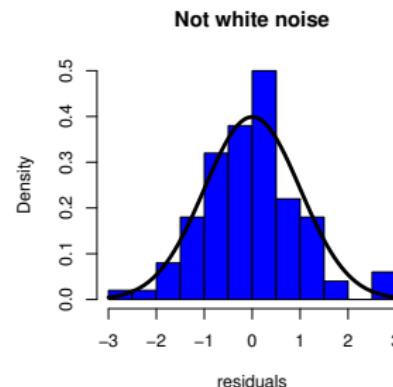
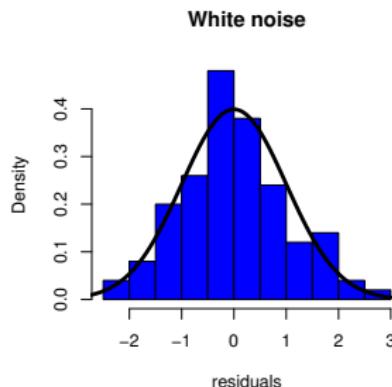
Not white noise



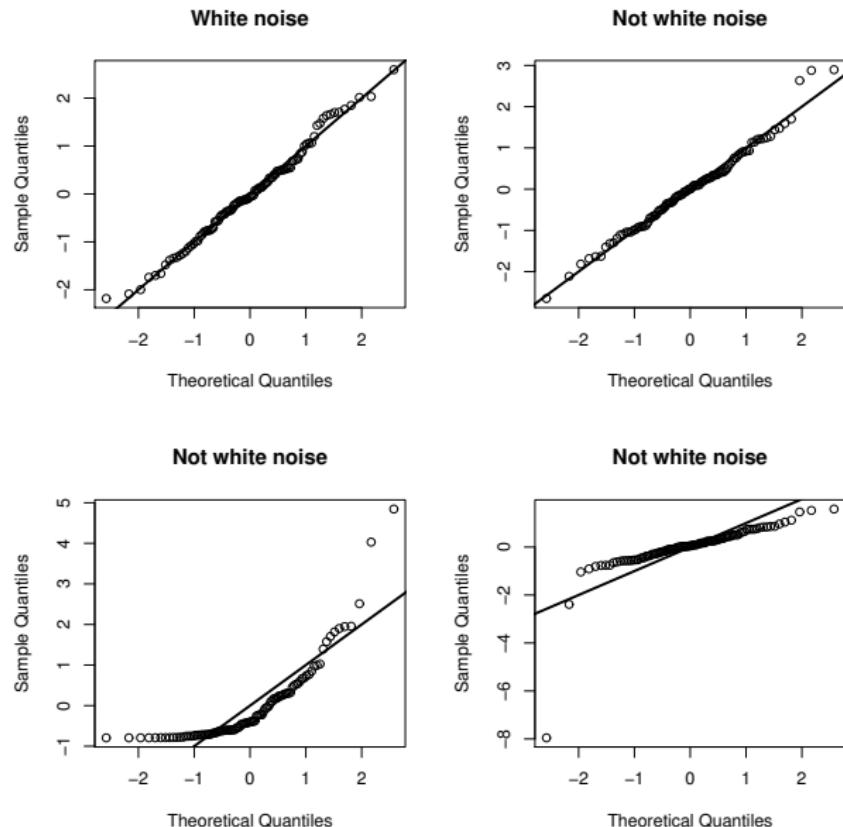
Not white noise



Residual analysis – Plot the data II

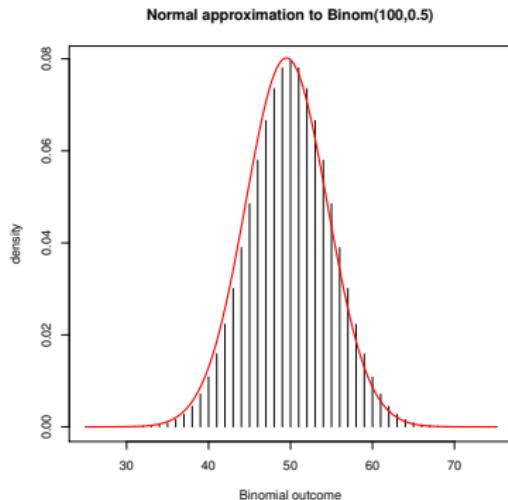


Residual analysis – Plot the data III

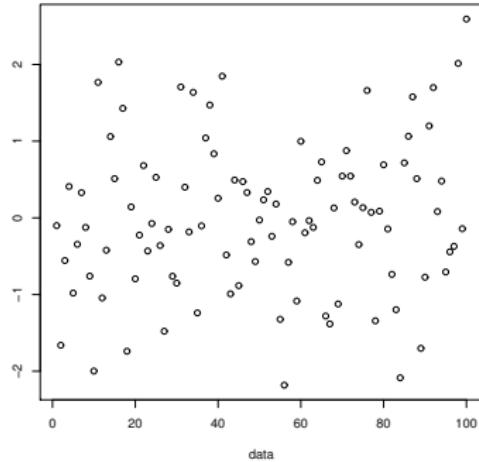


Residual analysis – sign test

- ▶ If (ε_t) is white noise, the probability that a new value has a different sign than the previous is $\frac{1}{2}$.
- ▶ Number of sign changes: $\text{Binom}(N - 1, \frac{1}{2})$.
- ▶ Approx. normal distribution; $N((N - 1)/2, (N - 1)/4)$:



Residual analysis – sign test II



- ▶ 95% confidence interval for sign changes within 100 white noise residuals: [40; 59]. Actual sign changes from the 100 data: 47.

Residual analysis – sign test III

Sign tests detects both asymmetry and correlation.

- ▶ Too few may indicate positive one-step correlation;
- ▶ Too many may indicate negative one-step correlation;
- ▶ Too few or too many may indicate that $P(\text{being above the mean}) \neq \frac{1}{2}$ with no correlation.

Residual analysis – autocorrelation test

- ▶ If (ε_t) is white noise, we have seen that $\hat{\rho}_\varepsilon(k) \sim N(0, \frac{1}{N})$ for all k (approx).
So : $Q^2 = \sum_{i=1}^m (\sqrt{N} \hat{\rho}_\varepsilon(i))^2 \sim \chi^2(m)$ (approx).
- ▶ If we instead consider the model errors $\varepsilon_t(\hat{\theta})$, $\frac{1}{N}$ is still an upper limit for the variance. However, we obtain less degrees of freedom:

$$Q^2 = \sum_{k=1}^m \left(\sqrt{N} \rho_{\varepsilon(\hat{\theta})}(k) \right)^2$$

is approximately distributed as $\chi^2(m - n)$, where n is the number of parameters - IF the residuals are white noise.

Residual analysis – summary

- ▶ Plot $\{\varepsilon_t(\hat{\theta})\}$; do the residuals look stationary?
- ▶ Tests in the autocorrelation. If $\{\varepsilon_t(\hat{\theta})\}$ is white noise then $(\rho_\varepsilon(k))$ is approximately Gaussian distributed with mean 0 and variance $1/N$.
If the model fails, calculate the SPACF also and see if an ARMA-structure for the residuals can be derived (Sec. 6.5.1)
- ▶ Since $\hat{\rho}_\varepsilon(k_1)$ and $\hat{\rho}_\varepsilon(k_2)$ are approximately independent for $k_1 \neq k_2$ (Eq. 6.4), the test statistic $Q^2 = \sum_{k=1}^m \left(\sqrt{N} \hat{\rho}_\varepsilon(\hat{\theta})(k) \right)^2$ is approximately distributed as $\chi^2(m - n)$, where n is the number of parameters.
- ▶ R: `tsdiag(output.from.arima)`

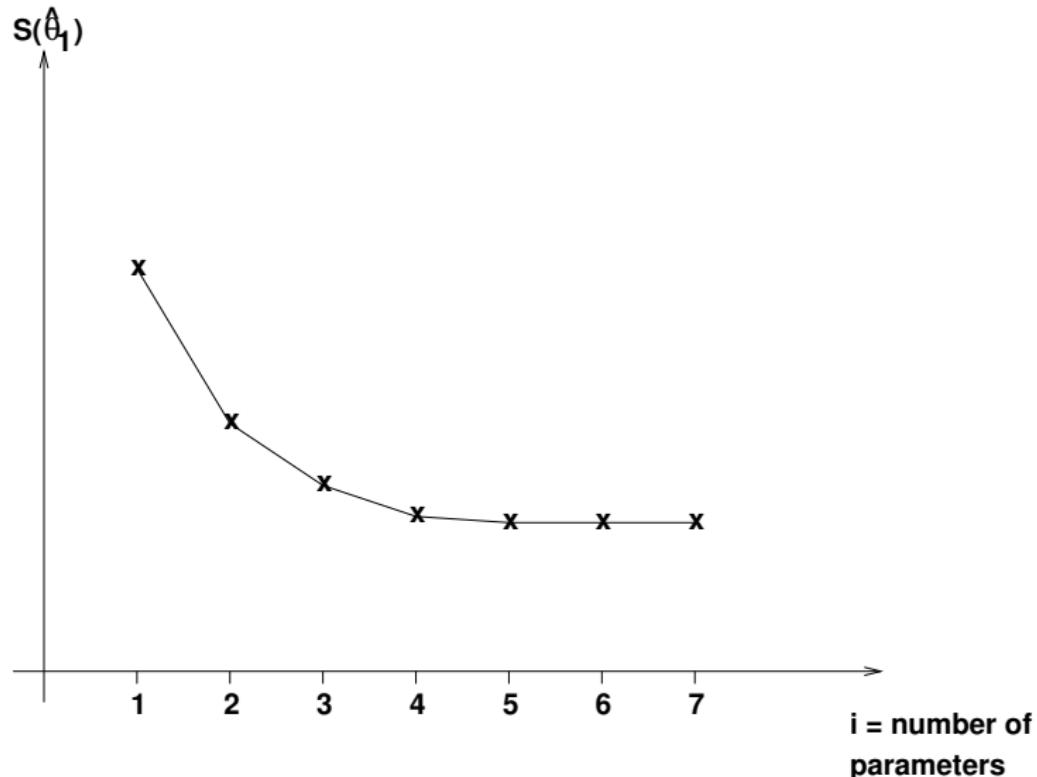
Residual analysis – summary II

- ▶ Test for the number of changes in sign.
 - ▶ In a series of length N there is $N - 1$ possibilities for changes in sign.
 - ▶ If the series is white noise (with mean zero) the probability of change is $1/2$ and the changes will be independent.
 - ▶ Therefore the number of changes is distributed as $\text{Bin}(N - 1, 1/2)$
 - ▶ R: `binom.test(No. of changes, N-1)`.
- ▶ Test in the scaled cumulated periodogram of the residuals is done by plotting it and adding lines at $\pm K_\alpha / \sqrt{q}$, where $q = (N - 2)/2$ for N even and $q = (N - 1)/2$ for N odd.
 - ▶ For $1 - \alpha$ confidence limits, K_α can be found in Table 6.2.
 - ▶ R (95% confidence interval):
 - > `cpgram('residuals')`

Model validation summary: Extensions/Reductions

- ▶ Residual analysis (Sec. 6.6.2): Is it possible to detect problems with residuals? (the 1-step prediction errors using the estimates, i.e. $(\varepsilon_t(\hat{\theta}))$ should be white noise).
- ▶ If the SACF or the SPACF of $(\varepsilon_t(\hat{\theta}))$ points towards a particular ARMA-structure we can derive how the original model should be extended (Sec. 6.5.1)
- ▶ If the model pass the residual analysis it makes sense to test null hypotheses about the parameters (Sec. 6.5.2)

Sum of squared residuals and model size



(It is assumed that the models are nested)

Test for model extension/reduction

- ▶ The test essentially checks if the reduction in SSE ($S_1 - S_2$) is large enough to justify the extra parameters in model 2 (n_2 parameters) as compared to model 1 (n_1 parameters). The number of observations used is called N .
- ▶ If vector θ_{extra} is used to denote the extra parameters in model 2 as compared to model 1, then the test is formally:

$$H_0 : \theta_{\text{extra}} = 0 \text{ vs. } H_1 : \theta_{\text{extra}} \neq 0$$

- ▶ If H_0 is true, it (approximately) holds that

$$\frac{(S_1 - S_2)/(n_2 - n_1)}{S_2/(N - n_2)} \sim F(n_2 - n_1, N - n_2)$$

The likelihood ratio test is also a possibility, which may coincide with the above.

Testing one parameter for significance

$$H_0 : \theta_i = 0 \quad \text{against} \quad H_1 : \theta_i \neq 0$$

- ▶ Can be done as described on the previous frame.
- ▶ Alternatively we can use a t-test based on the estimate and its standard error: $\hat{\theta}_i / \sqrt{V(\hat{\theta}_i)}$
- ▶ Under H_0 and for an $ARMA(p, q)$ -model this follows a $t(N - p - q)$ distribution (or $t(N - 1 - p - q)$ if we estimated an overall mean of the series)
- ▶ Often N is so large compared to the number of parameters that we can just use the standard normal distribution

Information criteria

For models that are not nested, the significance of a model extension cannot be tested.

- ▶ Select the model which minimizes some information criterion.
- ▶ Akaike's Information Criterion:

$$AIC = -2\log(L(Y_N; \hat{\theta}, \hat{\sigma}_\epsilon^2)) + 2n_{par}$$

- ▶ Bayesian Information Criterion:

$$BIC = -2\log(L(Y_N; \hat{\theta}, \hat{\sigma}_\epsilon^2)) + \log(N)n_{par}$$

- ▶ Except for an additive constant this can also be expressed as

$$AIC = N\log(\hat{\sigma}_\epsilon^2) + 2n_{par}$$

$$BIC = N\log(\hat{\sigma}_\epsilon^2) + \log(N)n_{par}$$

- ▶ AIC is most commonly used, but BIC yields a consistent estimate of the model order.

Highlights

- ▶ Estimating the sample ACF
- ▶ Table for identification of ARMA models.
- ▶ Iterative model building by making model for residuals.
- ▶ Residual analysis - several methods
- ▶ Testing significance of individual parameters
- ▶ Use information criteria when models are not nested. (Typically AIC)

Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

October 27, 2017

Outline of the lecture

- ▶ Estimation of parameters in linear dynamic models, Sec. 6.4
- ▶ Example in R

Identification of the ARMA-part

Characteristics for the autocorrelation functions:

	ACF $\rho(k)$	PACF ϕ_{kk}
AR(p)	Damped exponential and/or sine functions	$\phi_{kk} = 0$ for $k > p$
MA(q)	$\rho(k) = 0$ for $k > q$	Dominated by damped exponential and or/sine functions
ARMA(p, q)	Damped exponential and/or sine functions after lag $\max(0, q - p)$	Dominated by damped exponential and/or sine functions after lag $\max(0, p - q)$

The IACF is similar to the PACF; see the book page 133

Estimation

- ▶ We have an appropriate model structure $AR(p)$, $MA(q)$,
 $ARMA(p, q)$, $ARIMA(p, d, q)$ with p , d , and q known
- ▶ **Task:** Based on the observations find appropriate values of the parameters
- ▶ The book describes many methods:
 - ▶ Moment estimates
 - ▶ LS-estimates
 - ▶ Prediction error estimates
 - ▶ Conditioned
 - ▶ Unconditioned
 - ▶ ML-estimates
 - ▶ Conditioned
 - ▶ Unconditioned (exact)

Example



Using the autocorrelation functions we agreed that

$$\hat{y}_{t+1|t} = a_1 y_t + a_2 y_{t-1}$$

and we should select a_1 and a_2 so that the sum of the squared prediction errors is minimized.

To comply with the notation of the book we will write the 1-step forecasts as $\hat{y}_{t+1|t} = -\phi_1 y_t - \phi_2 y_{t-1}$

The errors given the parameters (ϕ_1 and ϕ_2)

- ▶ Observations: y_1, y_2, \dots, y_N
- ▶ Errors: $e_{t+1|t} = y_{t+1} - \hat{y}_{t+1|t} = y_{t+1} - (-\phi_1 y_t - \phi_2 y_{t-1})$

$$e_{3|2} = y_3 + \phi_1 y_2 + \phi_2 y_1$$

$$e_{4|3} = y_4 + \phi_1 y_3 + \phi_2 y_2$$

$$e_{5|4} = y_5 + \phi_1 y_4 + \phi_2 y_3$$

$$\vdots$$

$$e_{N|N-1} = y_N + \phi_1 y_{N-1} + \phi_2 y_{N-2}$$

$$\begin{bmatrix} y_3 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} -y_2 & -y_1 \\ \vdots & \vdots \\ -y_{N-1} & -y_{N-2} \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} + \begin{bmatrix} e_{3|2} \\ \vdots \\ e_{N|N-1} \end{bmatrix}$$

Or just:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

Solution

To minimize the sum of the squared 1-step prediction errors $\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}$ we use the result for the General Linear Model from Chapter 3:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

With

$$\mathbf{X} = \begin{bmatrix} -y_2 & -y_1 \\ \vdots & \vdots \\ -y_{N-1} & -y_{N-2} \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} y_3 \\ \vdots \\ y_N \end{bmatrix}$$

- ▶ Asymptotically: $V(\hat{\boldsymbol{\theta}}) = \sigma_\epsilon^2 (\mathbf{X}^T \mathbf{X})^{-1}$
- ▶ The method is called the LS-estimator for dynamical systems
- ▶ The method is also in the class of prediction error methods since it minimize the sum of the squared 1-step prediction errors
- ▶ **How does it generalize to AR(p)-models?**

Small illustrative example using R

```
obs <- c(-3.51, -3.81, -1.85, -2.02, -1.91, -0.88) ;
N <- length(obs)
(Y <- obs[3:N])

## [1] -1.85 -2.02 -1.91 -0.88

(X <- cbind(-obs[2:(N-1)], -obs[1:(N-2)]))

##      [,1] [,2]
## [1,] 3.81 3.51
## [2,] 1.85 3.81
## [3,] 2.02 1.85
## [4,] 1.91 2.02

solve(t(X) %*% X, t(X) %*% Y) # Estimates

##      [,1]
## [1,] -0.1474288
## [2,] -0.4476040
```

Maximum Likelihood estimates

- ▶ ARMA(p, q)-process:

$$Y_t + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

- ▶ Notation:

$$\boldsymbol{\theta}^T = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)$$

$$\mathbf{Y}_t^T = (Y_t, Y_{t-1}, \dots, Y_1)$$

- ▶ The Likelihood function is the joint probability distribution function for all observations for given values of $\boldsymbol{\theta}$ and σ_ε^2 :

$$L(\mathbf{Y}_N; \boldsymbol{\theta}, \sigma_\varepsilon^2) = f(\mathbf{Y}_N | \boldsymbol{\theta}, \sigma_\varepsilon^2)$$

- ▶ Given the observations \mathbf{Y}_N we estimate $\boldsymbol{\theta}$ and σ_ε^2 as the values for which the likelihood is maximized.

The likelihood function for ARMA(p, q)-models

- ▶ The random variable $Y_N | \mathbf{Y}_{N-1}$ only contains ε_N as a random component
- ▶ ε_N is a white noise process at time N and does therefore not depend on anything
- ▶ We therefore know that the random variables $Y_N | \mathbf{Y}_{N-1}$ and \mathbf{Y}_{N-1} are independent, hence:

$$f(\mathbf{Y}_N | \boldsymbol{\theta}, \sigma_\varepsilon^2) = f(Y_N | \mathbf{Y}_{N-1}, \boldsymbol{\theta}, \sigma_\varepsilon^2) f(\mathbf{Y}_{N-1} | \boldsymbol{\theta}, \sigma_\varepsilon^2)$$

- ▶ Repeating these arguments:

$$L(\mathbf{Y}_N; \boldsymbol{\theta}, \sigma_\varepsilon^2) = \left(\prod_{t=p+1}^N f(Y_t | \mathbf{Y}_{t-1}, \boldsymbol{\theta}, \sigma_\varepsilon^2) \right) f(\mathbf{Y}_p | \boldsymbol{\theta}, \sigma_\varepsilon^2)$$

The conditional likelihood function

- ▶ It turns out that the estimates obtained using the *conditional likelihood function*:

$$L(\mathbf{Y}_N; \boldsymbol{\theta}, \sigma_{\varepsilon}^2) = \prod_{t=p+1}^N f(Y_t | \mathbf{Y}_{t-1}, \boldsymbol{\theta}, \sigma_{\varepsilon}^2)$$

results in (almost) the same estimates as the *exact likelihood function* when many observations are available

- ▶ For small samples there can be some differences
- ▶ Software:
 - ▶ The R function arima calculate exact estimates per default

Evaluating the conditional likelihood function

- ▶ **Task:** Find the conditional densities given specified values of the parameters θ and σ_ε^2
- ▶ The mean of the random variable $Y_t | \mathbf{Y}_{t-1}$ is the the 1-step forecast $\hat{Y}_{t|t-1}$
- ▶ The prediction error $\varepsilon_t = Y_t - \hat{Y}_{t|t-1}$ has variance σ_ε^2
- ▶ We assume that the process is Gaussian:

$$f(Y_t | \mathbf{Y}_{t-1}, \theta, \sigma_\varepsilon^2) = \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp \left(-\frac{1}{2\sigma_\varepsilon^2} (Y_t - \hat{Y}_{t|t-1}(\theta))^2 \right)$$

- ▶ And therefore:

$$L(\mathbf{Y}_N; \theta, \sigma_\varepsilon^2) = (\sigma_\varepsilon^2 2\pi)^{-\frac{N-p}{2}} \exp \left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{t=p+1}^N \varepsilon_t^2(\theta) \right)$$

ML-estimates

- ▶ The (conditional) ML-estimate $\hat{\theta}$ is a prediction error estimate since it is obtained by minimizing

$$S(\theta) = \sum_{t=p+1}^N \varepsilon_t^2(\theta)$$

- ▶ By differentiating w.r.t. σ_ε^2 it can be shown that the ML-estimate of σ_ε^2 is (remember that p is the order of the AR part):

$$\hat{\sigma}_\varepsilon^2 = S(\hat{\theta})/(N - p)$$

- ▶ The estimate $\hat{\theta}$ is asymptotically unbiased and efficient, and the variance-covariance matrix is approximately

$$2\sigma_\varepsilon^2 \mathbf{H}^{-1}$$

where \mathbf{H} contains the 2nd order partial derivatives of $S(\theta)$ at the minimum

Finding the ML-estimates using the PE-method

- ▶ 1-step predictions:

$$\hat{Y}_{t|t-1} = -\phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

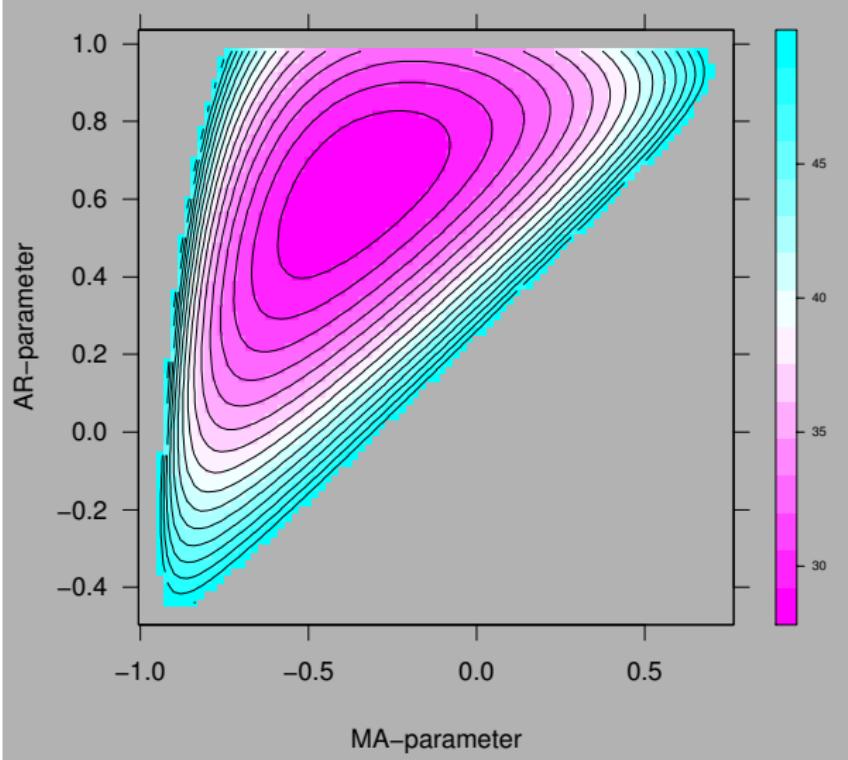
- ▶ If we use (Condition on) $\varepsilon_p = \varepsilon_{p-1} = \cdots = \varepsilon_{p+1-q} = 0$ we can find:

$$\hat{Y}_{p+1|p} = -\phi_1 Y_p - \cdots - \phi_p Y_1 + \theta_1 \varepsilon_p + \cdots + \theta_q \varepsilon_{p+1-q}$$

- ▶ Which will give us $\varepsilon_{p+1} = Y_{p+1} - \hat{Y}_{p+1|p}$ and we can then calculate $\hat{Y}_{p+2|p+1}$ and ε_{p+2} ... and so on until we have all the 1-step prediction errors we need.
- ▶ We use numerical optimization to find the parameters which minimize the sum of squared prediction errors

$S(\theta)$ for $(1 + 0.7B)Y_t = (1 - 0.4B)\varepsilon_t$ with $\sigma_\varepsilon^2 = 0.25^2$

Data: arima.sim(model=list(ar=-0.7,ma=0.4), n=500, sd=0.25)



Moment estimates

- ▶ Given the model structure: Find formulas for the theoretical autocorrelation or autocovariance as function of the parameters in the model
- ▶ Estimate, e.g. calculate the SACF
- ▶ Solve the equations by using the lowest lags necessary
- ▶ **Complicated**
- ▶ **General properties of the estimator are unknown**

Moment estimates for $AR(p)$ -processes

In this case moment estimates are simple to find due to the Yule-Walker equations. We simply plug in the estimated autocorrelation function in lags 1 to p :

$$\begin{bmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \\ \vdots \\ \hat{\rho}(p) \end{bmatrix} = \begin{bmatrix} 1 & \hat{\rho}(1) & \cdots & \hat{\rho}(p-1) \\ \hat{\rho}(1) & 1 & \cdots & \hat{\rho}(p-2) \\ \vdots & \vdots & & \vdots \\ \hat{\rho}(p-1) & \hat{\rho}(p-2) & \cdots & 1 \end{bmatrix} \begin{bmatrix} -\phi_1 \\ -\phi_2 \\ \vdots \\ -\phi_p \end{bmatrix}$$

and solve w.r.t. the ϕ 's

The function `ar` in R does this

Highlights

- ▶ Maximum likelihood estimation by looking at independent one step prediction errors.
- ▶ “Essentially, all models are wrong, but some are useful.”
(George E. P. Box)

Time Series Analysis

Lasse Engbo Christiansen

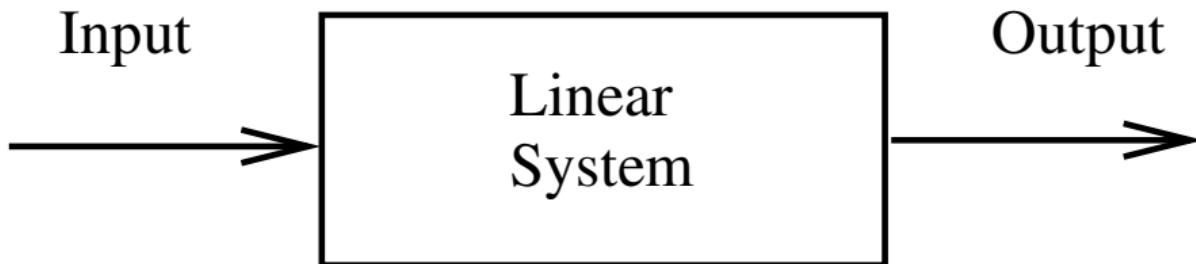
Department of Applied Mathematics and Computer Science
Technical University of Denmark

October 28, 2016

Outline of the lecture

- ▶ Input-Output systems, sec. 4 introduction and 4.1
- ▶ Linear system notation
- ▶ The z -transform, section 4.4
- ▶ Cross Correlation Functions – from Sec. 6.2.2
- ▶ Transfer function models; identification, estimation, validation, prediction, Chap. 8

Linear Dynamic Systems



- ▶ We are going to study the case where we measure the input and the output to/from a system
- ▶ Here we will discuss some theory and descriptions for such systems

Dynamic response

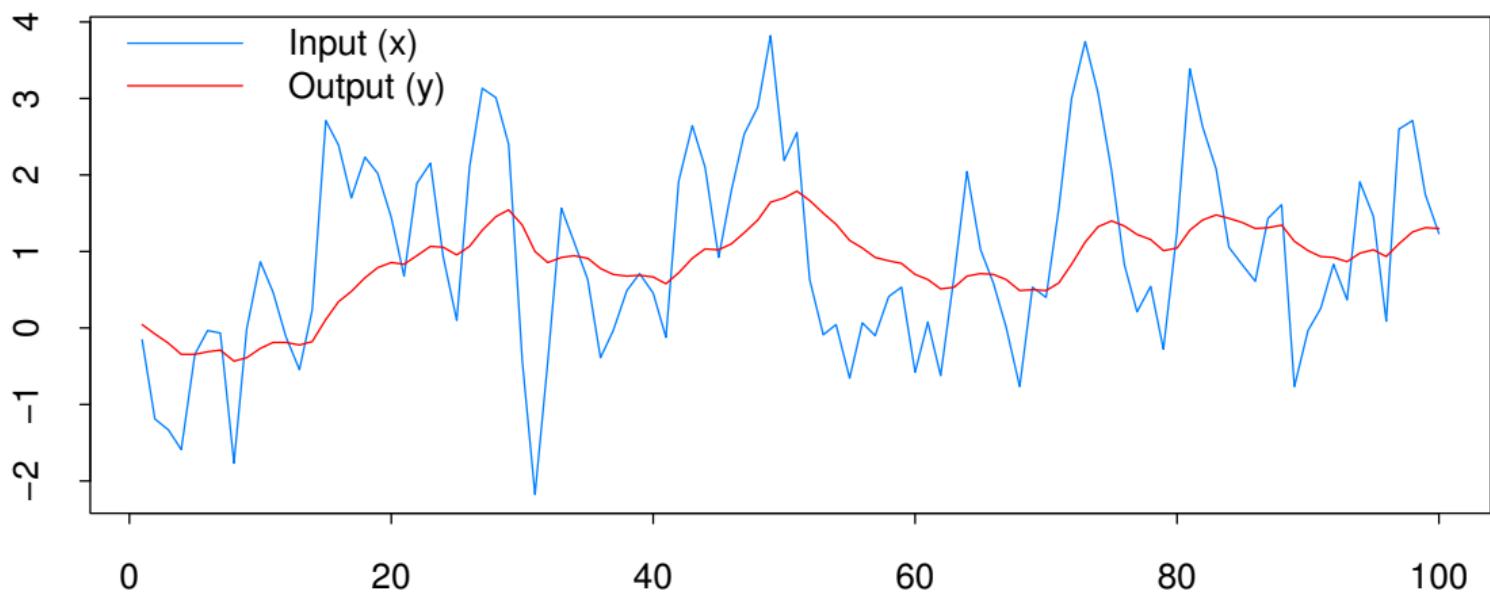
What would happen to the temperature inside a hollow, insulated, concrete block, if you

- ▶ place it in a controlled temperature environment at 20°C ,
- ▶ wait until everything is settled (all temperatures are equal), and then
- ▶ suddenly raise the temperature by 100°C outside the block?

Sketch the temporal development of the temperature outside and inside the block

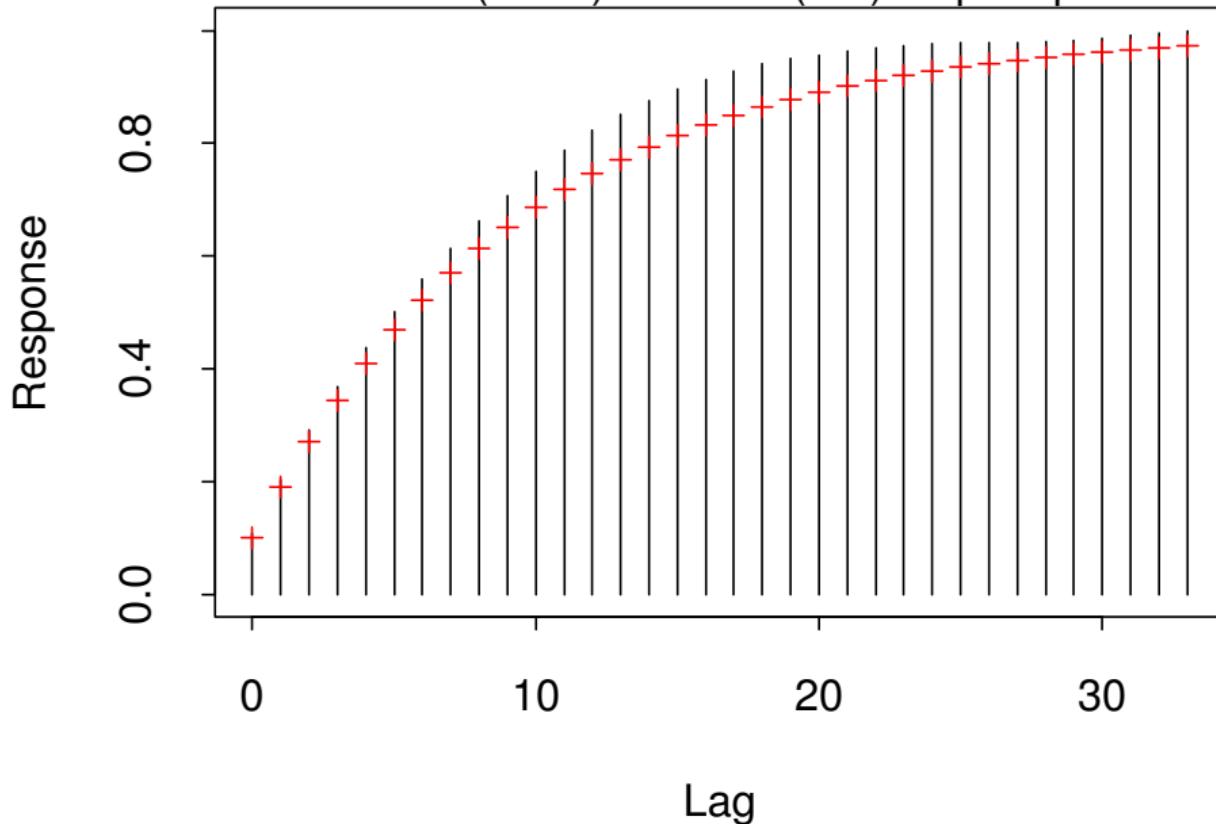
Dynamic response characteristics from data

- An important aspect of what we aim at later on is to identify the characteristics of the dynamic response based on measurements of input and output signals

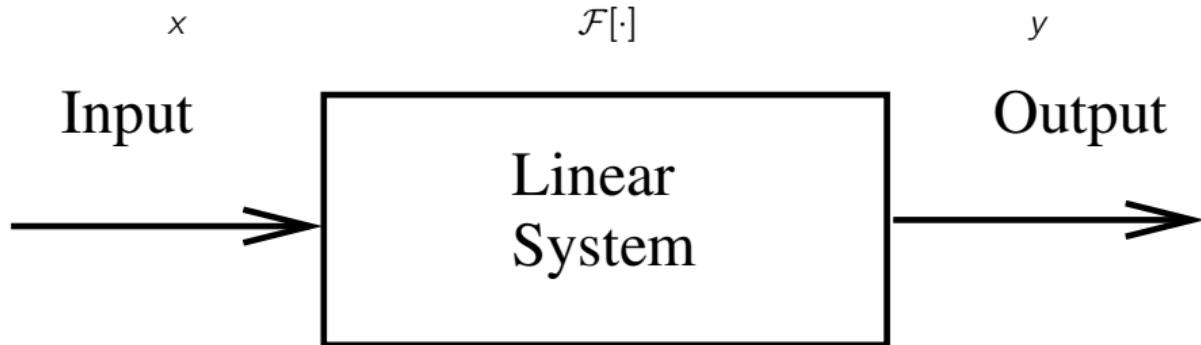


Dyn. response characteristics from data II

Estimated (black) and true (red) step response



Linear Dynamic Systems – notation



$x(t)$

x_t

$X(\omega)$

$X(z)$

$(X(s))$

Differential eq., $h(u)$

Difference eq., h_k , $h(B)$

$\mathcal{H}(\omega)$

$H(z)$

$H(s)$

$y(t)$

y_t

$Y(\omega)$

$Y(z)$

$Y(s))$

Dynamic Systems – Some characteristics

Def. Linear system:

$$\mathcal{F}[\lambda_1 x_1(t) + \lambda_2 x_2(t)] = \lambda_1 \mathcal{F}[x_1(t)] + \lambda_2 \mathcal{F}[x_2(t)]$$

Def. Time invariant system:

$$y(t) = \mathcal{F}[x(t)] \Rightarrow y(t - \tau) = \mathcal{F}[x(t - \tau)]$$

Def. Stable system: A system is said to be *stable* if any constrained input implies a constrained output.

Def. Causal system: A system is said to be *physically feasible* or *causal*, if the output at time t does not depend on future values of the input.

Description in the time domain (Convolution)

For *linear time invariant systems*:

- ▶ Continuous time:

$$y(t) = \int_{-\infty}^{\infty} h(u)x(t-u) du \quad (1)$$

- ▶ Discrete time:

$$y_t = \sum_{k=-\infty}^{\infty} h_k x_{t-k} \quad (2)$$

- ▶ $h(u)$ or h_k is called the *impulse response*
- ▶ $S_k = \sum_{j=-\infty}^k h_j$ is called the *step response* (similar def. in continuous time)
- ▶ The impulse response can be determined by "sending a 1 through the system" - $x_t = 1$ for $t = 0$ and zero otherwise

Example

- ▶ System: $y_t - ay_{t-1} = bx_t$
- ▶ Can be written: $y_t = bx_t + ay_{t-1} = bx_t + a(bx_{t-1} + ay_{t-2})$ or

$$y_t = b(x_t + ax_{t-1} + a^2x_{t-2} + a^3x_{t-3} + \dots) = b \sum_{k=0}^{\infty} a^k x_{t-k}$$

- ▶ Is the system *linear* and *time invariant*? – Yes and yes.
- ▶ The *impulse response* is $h_k = ba^k$, $k \geq 0$ (0 otherwise)
- ▶ Is the system causal? – Yes

$$\sum_{k=-\infty}^{\infty} |h_k| = \sum_{k=0}^{\infty} |b||a|^k = \begin{cases} |b|/(1-|a|) & ; \quad |a| < 1 \\ \infty & ; \quad |a| \geq 1 \end{cases}$$

- ▶ Is the system *stable*? – Yes, for $|a| < 1$ (stability does not depend on b)

Stability based on the impulse response function

If the impulse response function is absolutely convergent, the system is stable (Theorem 4.3).

- ▶ Continuous time:

$$\int_{-\infty}^{\infty} |h(u)| du < \infty$$

- ▶ Discrete time:

$$\sum_{k=-\infty}^{\infty} |h_k| < \infty$$

Example: Calculating the impulse response function.

The impulse response can be determined by 'sending a 1 through the system'. Consider the linear, time invariant system

$$y_t - 0.8y_{t-1} = 2x_t - x_{t-1}$$

By putting $x = \delta_k = \delta_{0k}$ (Kronecker delta) we see that $y_k = h_k = 0$ for $k < 0$. For $k = 0$ we get

$$\begin{aligned} y_0 &= 0.8y_{-1} + 2\delta_0 - \delta_{-1} \\ &= 0.8 \times 0 + 2 \times 1 - 0 = 2 \end{aligned}$$

i.e. $h_0 = 2$.

Example continued

$$y_t - 0.8y_{t-1} = 2x_t - x_{t-1}$$

Going on we get

$$y_1 = 0.8y_0 + 2\delta_1 - \delta_0 = 0.8 \times 2 + 2 \times 0 - 1 = 0.6$$

$$y_2 = 0.8y_1 = 0.48$$

$$y_k = 0.8^{k-1}0.6 \quad (k > 0)$$

Hence, the impulse response function is

$$h_k = \begin{cases} 0 & \text{for } k < 0 \\ 2 & \text{for } k = 0 \\ 0.8^{k-1}0.6 & \text{for } k > 0 \end{cases}$$

which clearly represents a causal system. Furthermore, the system is stable since $\sum_0^{\infty} |h_k| = 2 + 0.6(1 + 0.8 + 0.8^2 + \dots) = 5 < \infty$

The z -transform

- ▶ A way to describe dynamical systems in discrete time

$$Z(\{x_t\}) = X(z) = \sum_{t=-\infty}^{\infty} x_t z^{-t} \quad (z \in \mathbb{C})$$

- ▶ The z -transform of a time delay: $Z(\{x_{t-\tau}\}) = z^{-\tau} X(z)$
- ▶ The **transfer function** of the system is called $H(z) = \sum_{t=-\infty}^{\infty} h_t z^{-t}$

$$y_t = \sum_{k=-\infty}^{\infty} h_k x_{t-k} \Leftrightarrow Y(z) = H(z)X(z)$$

Linear Difference Equation

$$y_t + a_1 y_{t-1} + \cdots + a_p y_{t-p} = b_0 x_{t-\tau} + b_1 x_{t-\tau-1} + \cdots + b_q x_{t-\tau-q}$$
$$(1 + a_1 z^{-1} + \cdots + a_p z^{-p}) Y(z) = z^{-\tau} (b_0 + b_1 z^{-1} + \cdots + b_q z^{-q}) X(z)$$

Transfer function:

$$H(z) = \frac{z^{-\tau} (b_0 + b_1 z^{-1} + \cdots + b_q z^{-q})}{(1 + a_1 z^{-1} + \cdots + a_p z^{-p})}$$
$$= \frac{z^{-\tau} (1 - n_1 z^{-1})(1 - n_2 z^{-1}) \cdots (1 - n_q z^{-1}) b_0}{(1 - \lambda_1 z^{-1})(1 - \lambda_2 z^{-1}) \cdots (1 - \lambda_p z^{-1})}$$

Where the roots n_1, n_2, \dots, n_q are called the *zeros of the system* and $\lambda_1, \lambda_2, \dots, \lambda_p$ are called the *poles of the system*

The system is stable if all poles lie within the unit circle

Relation to the backshift operator

$$\begin{aligned}y_t + a_1 y_{t-1} + \cdots + a_p y_{t-p} &= b_0 x_{t-\tau} + b_1 x_{t-\tau-1} + \cdots + b_q x_{t-\tau-q} \\(1 + a_1 z^{-1} + \cdots + a_p z^{-p})Y(z) &= z^{-\tau}(b_0 + b_1 z^{-1} + \cdots + b_q z^{-q})X(z) \\(1 + a_1 B^1 + \cdots + a_p B^p)y_t &= B^\tau(b_0 + b_1 B^1 + \cdots + b_q B^q)x_t \\\varphi(B)y_t &= \omega(B)B^\tau x_t\end{aligned}$$

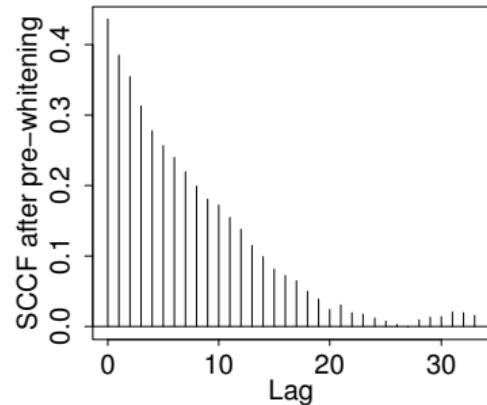
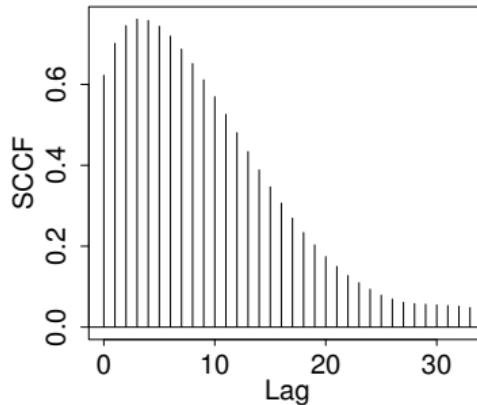
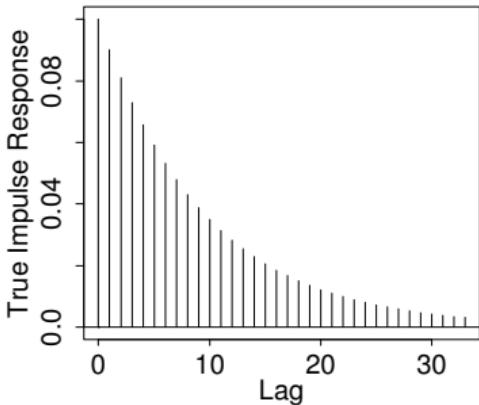
The output can be written:

$$y_t = \varphi^{-1}(B)\omega(B)B^\tau x_t = h(B)x_t = \left[\sum_{i=0}^{\infty} h_i B^i \right] x_t = \sum_{i=0}^{\infty} h_i x_{t-i}$$

$h(B)$ is **also** called the *transfer function*. Using $h(B)$ the system is assumed to be causal; compare with $H(z) = \sum_{t=-\infty}^{\infty} h_t z^{-t}$

Estimating the impulse response

- ▶ The poles and zeros characterize the impulse response (Appendix A and Chapter 8)
- ▶ If we can estimate the impulse response from recordings of input and output we can get information that allows us to *suggest a structure for the transfer function*



Cross covariance and cross correlation functions

Estimate of the cross covariance function:

$$C_{XY}(k) = \frac{1}{N} \sum_{t=1}^{N-k} (X_t - \bar{X})(Y_{t+k} - \bar{Y})$$

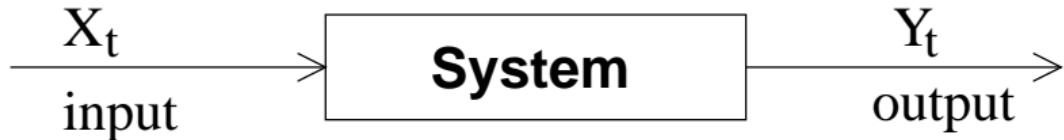
$$C_{XY}(-k) = \frac{1}{N} \sum_{t=1}^{N-k} (X_{t+k} - \bar{X})(Y_t - \bar{Y})$$

Estimate of the cross correlation function:

$$\hat{\rho}_{XY}(k) = C_{XY}(k) / \sqrt{C_{XX}(0)C_{YY}(0)}$$

If at least one of the processes is white noise and if the processes are uncorrelated then $\hat{\rho}_{XY}(k)$ is approximately normally distributed with mean 0 and variance $1/N$

Systems without measurement noise



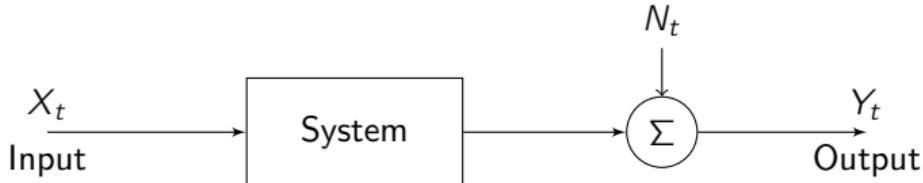
$$Y_t = \sum_{i=-\infty}^{\infty} h_i X_{t-i}$$

Given γ_{XX} and the system description we obtain

$$\gamma_{YY}(k) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h_i h_j \gamma_{XX}(k-j+i)$$

$$\gamma_{XY}(k) = \sum_{i=-\infty}^{\infty} h_i \gamma_{XX}(k-i).$$

Systems with measurement noise



$$Y_t = \sum_{i=-\infty}^{\infty} h_i X_{t-i} + N_t.$$

Given γ_{XX} and γ_{NN} we obtain

$$\gamma_{YY}(k) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h_i h_j \gamma_{XX}(k-j+i) + \gamma_{NN}(k)$$

$$\gamma_{XY}(k) = \sum_{i=-\infty}^{\infty} h_i \gamma_{XX}(k-i).$$

IMPORTANT ASSUMPTION: No feedback in the system.

Estimating the impulse response

- ▶ On a previous slide we saw that we got a good picture of the true impulse response when *pre-whitening* the data
- ▶ The reason is

$$\gamma_{XY}(k) = \sum_{i=-\infty}^{\infty} h_i \gamma_{XX}(k-i)$$

- ▶ and only if $\{X_t\}$ is white noise then we get $\boxed{\gamma_{XY}(k) = h_k \sigma_X^2}$
- ▶ Therefore if $\{X_t\}$ is white noise the SCCF $\hat{\rho}_{XY}(k)$ is proportional to \hat{h}_k
- ▶ Normally $\{X_t\}$ is not white noise – we fix this using pre-whitening

Pre-whitening

- a) A suitable ARMA-model is applied to the input series:

$$\eta(B)X_t = \nu(B)\alpha_t.$$

- b) We perform a *prewhitening* of the input series

$$\alpha_t = \nu(B)^{-1}\eta(B)X_t$$

- c) The output-series $\{Y_t\}$ is filtered with the same model, i.e.

$$\beta_t = \nu(B)^{-1}\eta(B)Y_t.$$

- d) Now the *impulse response function is estimated* by

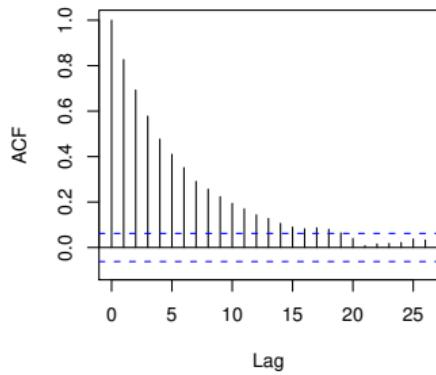
$$\hat{h}_k = C_{\alpha\beta}(k)/C_{\alpha\alpha}(0) = C_{\alpha\beta}(k)/S_\alpha^2.$$

Example using R

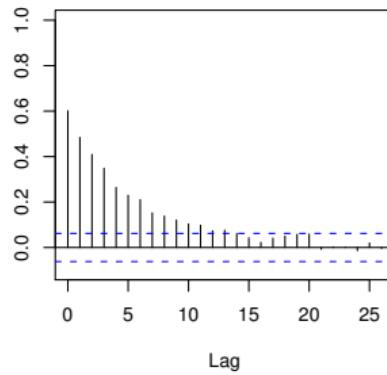
```
## ARMA structure for x; AR(1)
x.struct<-c(1,0,0)
## Estimate the model (check for convergence):
x.fit <- arima(x, order=x.struct, include.mean=F)
## Filter x and y:
x.filt <- x - x.fit$coef[1] * c(0,x[1:(length(x)-1)])
y.filt <- y - x.fit$coef[1] * c(0,y[1:(length(y)-1)])
##Estimate SCCF for the filtered series:
acf(cbind(y.filt, x.filt), type="correlation")
```

Graphical output

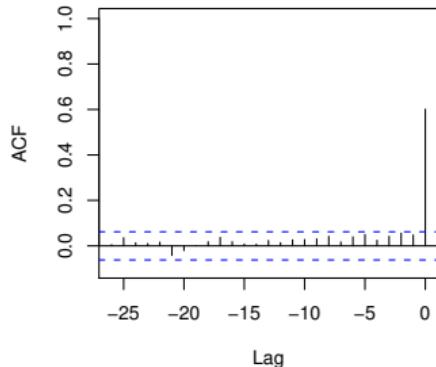
y.filter



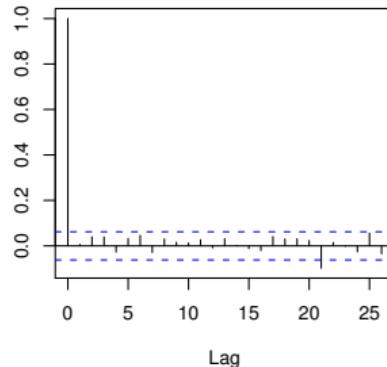
y.filter & x.filter



x.filter & y.filter

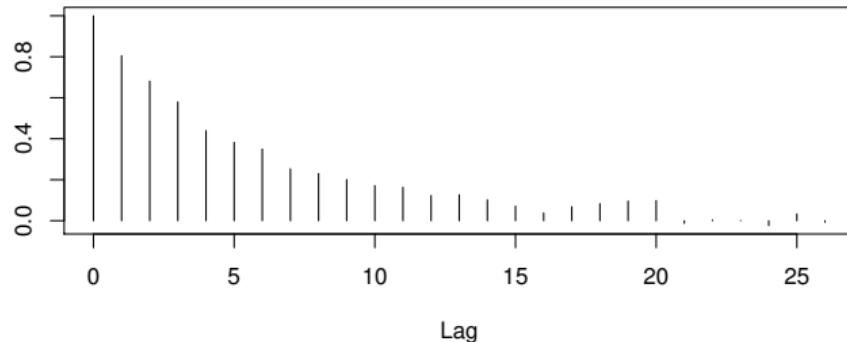


x.filter

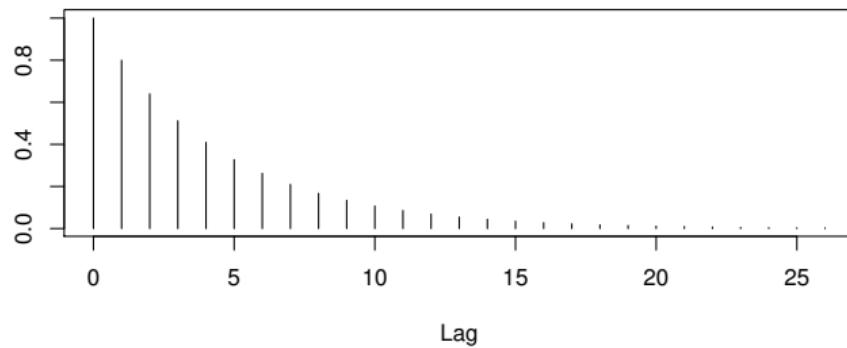


Impulse response functions

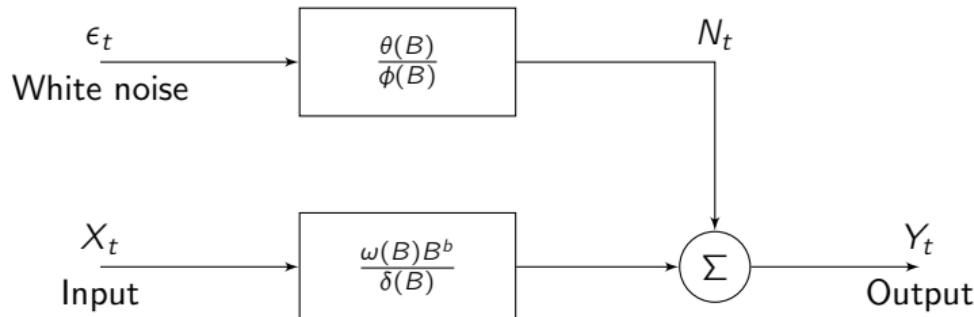
Estimated impulse response function



True Impulse response function $h_k=0.8^k$



Transfer function models



$$Y_t = \frac{\omega(B)}{\delta(B)} B^b X_t + \frac{\theta(B)}{\varphi(B)} \epsilon_t$$

- ▶ Also called Box-Jenkins models
- ▶ Can be extended to include more inputs – see the book.

Some names

The following are all sub-models of transfer function models.

- ▶ FIR: Finite Impulse Response (impulse response function(s) of finite length)
- ▶ ARX: Auto Regressive with eXogenous input
- ▶ ARMAX/CARMA: Auto Regressive Moving Average with eXogenous input / Controlled ARMA (Common poles in transfer functions)
- ▶ OE: Output Error (No model for the observation noise)

Regression models with ARMA noise (the `xreg` option to `arima` in R)

Identification of transfer function models

$$h(B) = \frac{\omega(B)B^b}{\delta(B)} = h_0 + h_1B + h_2B^2 + h_3B^3 + h_4B^4 + \dots$$

- ▶ Using pre-whitening we estimate the impulse response and “guess” an appropriate structure of $h(B)$ based on this.
- ▶ It is a good idea to experiment with some structures. With Matlab’s “ident” toolbox (use q^{-1} instead of B):

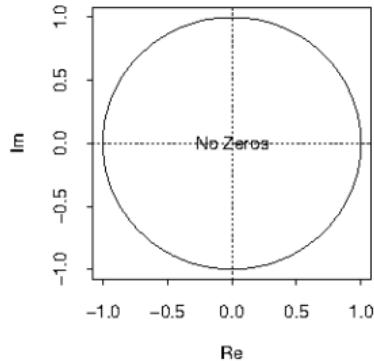
```
A = 1; B = 1; C = 1; D = 1; F = [1 -2.55 2.41 -0.85];  
mod = idpoly(A, B, C, D, F, 1, 1)  
impulse(mod)
```

- ▶ PEZdemo (complex poles/zeros should be in pairs):
<http://users.ece.gatech.edu/mcclella/matlabGUIs/>

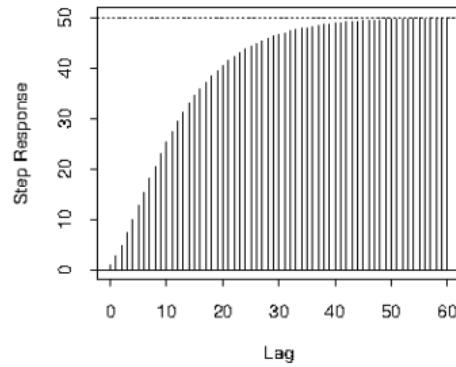
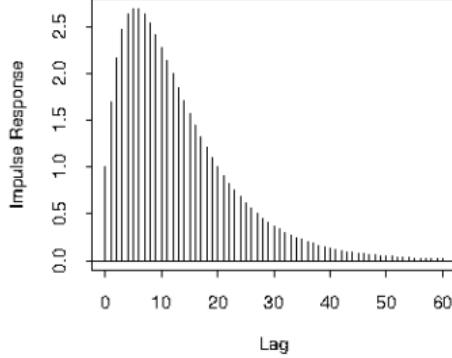
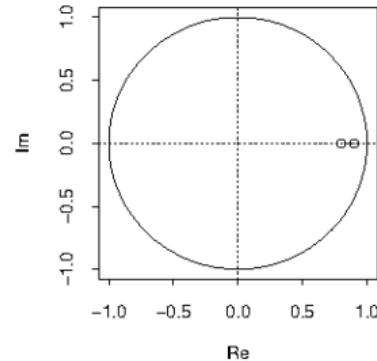
2 real poles

$$h(B) = \frac{1}{1 - 1.7B + 0.72B^2}$$

Zeros



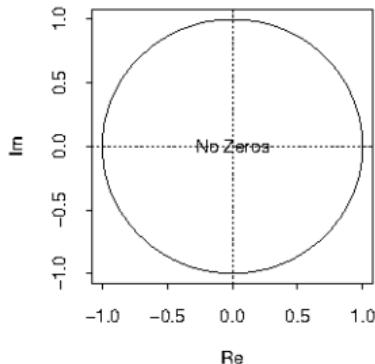
Poles



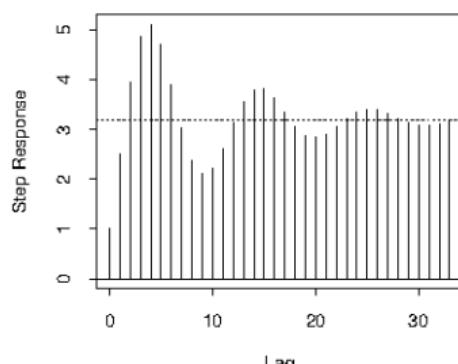
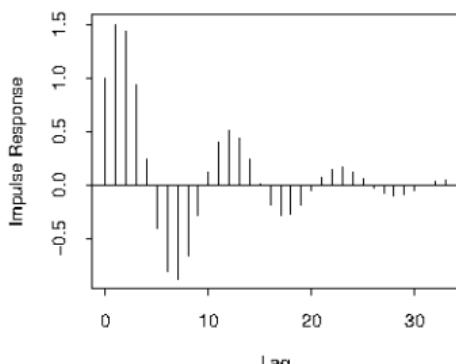
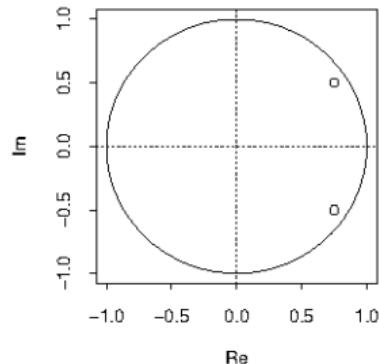
2 complex

$$h(B) = \frac{1}{1 - 1.5B + 0.81B^2}$$

Zeros



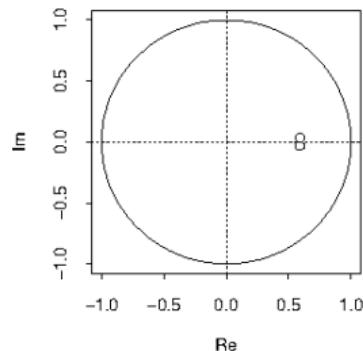
Poles



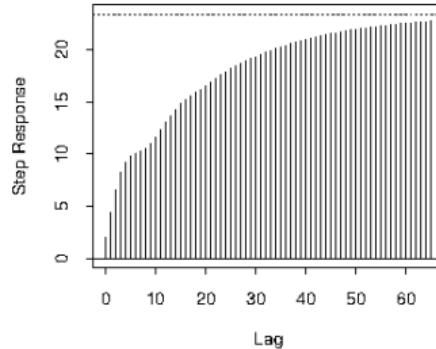
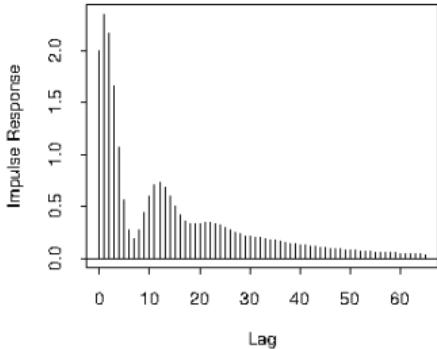
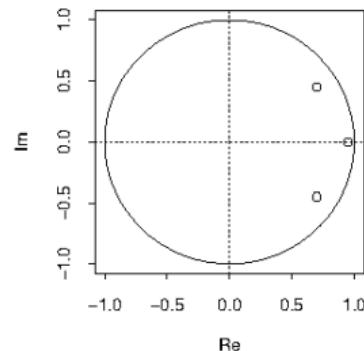
1 real, 2 comp

$$h(B) = \frac{2 - 2.35B + 0.69B^2}{1 - 2.35B + 2.02B^2 - 0.66B^3}$$

Zeros



Poles



Identification of the transfer function for the noise

- ▶ After selection of the structure of the transfer function of the input we estimate the parameters of the model

$$Y_t = \frac{\omega(B)}{\delta(B)} B^b X_t + N_t$$

- ▶ We extract the residuals $\{N_t\}$ and identifies a structure for an ARMA model of this series

$$N_t = \frac{\theta(B)}{\varphi(B)} \varepsilon_t \quad \Leftrightarrow \quad \varphi(B) N_t = \theta(B) \varepsilon_t$$

- ▶ We then have the full structure of the model and we estimate all parameters simultaneously

Estimation

- ▶ Form 1-step predictions, treating the input $\{X_t\}$ as known in the future (if $\{X_t\}$ is really stochastic we *condition* on the observed values)
- ▶ Select the parameters so that the sum of squares of these errors is as small as possible
- ▶ If $\{\varepsilon_t\}$ is normal then the ML estimates are obtained
- ▶ For FIR and ARX models we can write the model as $\mathbf{Y}_t = \mathbf{X}_t^T \boldsymbol{\theta} + \varepsilon_t$ and use LS-estimates
- ▶ Moment estimates: Based on the structure of the transfer function we find the theoretical impulse response and we make a match with the lowest lags in the estimated impulse response
- ▶ Output error estimates ...

Model validation

As for ARMA models with the additions:

- ▶ Test for cross correlation between the residuals and the input

$$\hat{\rho}_{\varepsilon X}(k) \sim Norm(0, 1/N)$$

which is (approximately) correct when $\{\varepsilon_t\}$ is white noise and when there is no correlation between the input and the residuals

- ▶ A *Portmanteau test(Ljung-Box)* can also be performed

Prediction $\hat{Y}_{t+k|t}$

We must consider two situations

- ▶ The input is controllable, i.e. we can decide it and we can predict under different input-scenarios. In this case the prediction error variance is originating from the ARMA-part only (N_t).
- ▶ The input is only known until the present time point t and to predict the output we must predict the input. In this case the prediction error variance depend also on the autocovariance of the input process. In the book the case where the input can be modelled as an ARMA-process is considered.

Prediction II

$$\hat{Y}_{t+k|t} = \sum_{i=0}^{k-1} h_i \hat{X}_{t+k-i|t} + \sum_{i=k}^{\infty} h_i X_{t+k-i} + \hat{N}_{t+k|t}.$$

$$Y_{t+k} - \hat{Y}_{t+k|t} = \sum_{i=0}^{k-1} h_i (X_{t+k-i} - \hat{X}_{t+k-i|t}) + N_{t+k} - \hat{N}_{t+k|t}$$

- ▶ If the input is controllable then $\hat{X}_{t+k-i|t} = X_{t+k-i}$
- ▶ The book also considers the case where output is known until time t and input until time $t+j$

Prediction III

- We have

$$N_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$$

- And if we model the input as an ARMA-process we have

$$X_t = \sum_{i=0}^{\infty} \bar{\psi}_i \eta_{t-i}$$

- And thereby we get:

$$V[Y_{t+k} - \hat{Y}_{t+k|t}] = \sigma_{\eta}^2 \sum_{\ell=0}^{k-1} \left(\sum_{i_1+i_2=\ell} h_{i_1} \bar{\psi}_{i_2} \right)^2 + \sigma_{\varepsilon}^2 \sum_{i=0}^{k-1} \psi_i^2$$

Example: Prediction of a transfer function model

- ▶ Consider the system

$$\begin{aligned}Y_t &= h(B)X_t + N_t \\&= h(B)X_t + \Psi(B)\epsilon_t \\&= \frac{0.4}{1 - 0.6B}X_t + \frac{1}{1 - 0.4B}\epsilon_t, \quad \sigma_\epsilon^2 = 0.036\end{aligned}$$

- ▶ The following data is available:

t	1	2	3	4	5	6	7	8	9	10
Y	2.040	3.050	2.340	2.490	3.300	3.530	2.720	2.460		
X	1.661	4.199	1.991	2.371	3.521	3.269	0.741	2.238	2.544	3.201

- ▶ In order to perform a prediction, $\hat{Y}_{9|8}$, we must filter X with $h(B)$ and forecast $N_{9|8} = \Psi(B)\epsilon_{9|8}$
- ▶ We can then evaluate

$$\hat{Y}_{9|8} = h(B)X_9 + \hat{N}_{9|8}$$

Example continued

Filtering

For $t \in \{1, \dots, 10\}$ we evaluate

$$h(B)X_t = 0.4 \cdot \sum_{i=0}^{\infty} (0.6B)^i X_{t-i} \approx 0.4 \cdot \sum_{i=0}^{t-1} (0.6B)^i X_{t-i}$$

And for $t \in \{1, \dots, 8\}$:

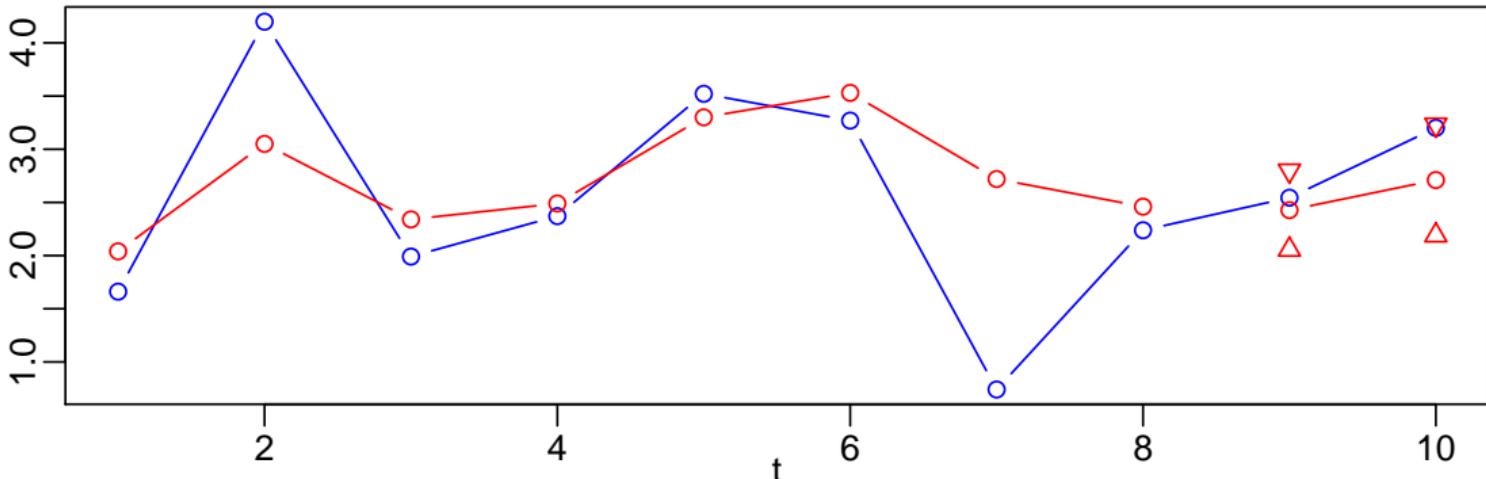
$$N_t = Y_t - h(B)X_t$$

Forecasting

For $k \in \{1, 2\}$:

$$\begin{aligned}\hat{N}_{8+k|8} &= N_8 \cdot 0.4^k \\ \hat{Y}_{8+k|8} &= h(B)X_{8+k} + \hat{N}_{8+k|8} \\ V(Y_{8+k} - \hat{Y}_{8+k|8}) &= V(\hat{Y}_{8+k|8}) = V\left(\sum_{i=0}^{k-1} \Psi^i \epsilon_{8+k-i}\right) \\ &= \begin{cases} V(\epsilon_9) = \sigma_\epsilon^2 \\ V(\epsilon_{10}) + 0.4^2 \cdot V(\epsilon_{10}) = (1 + 0.4^2)\sigma_\epsilon^2 \end{cases}\end{aligned}$$

t	historic								future	
	1	2	3	4	5	6	7	8	9	10
data										
Y	2.04	3.05	2.34	2.49	3.30	3.53	2.72	2.46		
X	1.66	4.20	1.99	2.37	3.52	3.27	0.74	2.24	2.54	3.20
filtered										
$h(B)X_t$	0.66	2.08	2.04	2.17	2.71	2.94	2.06	2.13	2.30	2.66
N	1.38	0.97	0.30	0.32	0.59	0.59	0.66	0.33		
forecasted										
$N_{t 8}$									0.13	0.05
$Y_{t 8}$									2.43	2.71



blue: X , red: Y

Intervention models

$$I_t = \begin{cases} 1 & t = t_0 \\ 0 & t \neq t_0 \end{cases}$$
$$Y_t = \frac{\omega(B)}{\delta(B)} I_t + \frac{\theta(B)}{\phi(B)} \varepsilon_t$$

See a real life example in the book.

Highlights

- ▶ Def. linear system

$$\mathcal{F}[\lambda_1 x_1(t) + \lambda_2 x_2(t)] = \lambda_1 \mathcal{F}[x_1(t)] + \lambda_2 \mathcal{F}[x_2(t)]$$

- ▶ Estimating Cross Covariance function

$$C_{XY}(k) = \frac{1}{N} \sum_{t=1}^{N-k} (X_t - \bar{X})(Y_{t+k} - \bar{Y})$$

$$C_{XY}(-k) = \frac{1}{N} \sum_{t=1}^{N-k} (X_{t+k} - \bar{X})(Y_t - \bar{Y})$$

- ▶ Transfer function model

$$Y_t = \frac{\omega(B)}{\delta(B)} B^b X_t + \frac{\theta(B)}{\varphi(B)} \varepsilon_t$$

- ▶ Use pre-whitening of input
- ▶ Fit ARMAX models using the `xreg` option to `arima`

Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

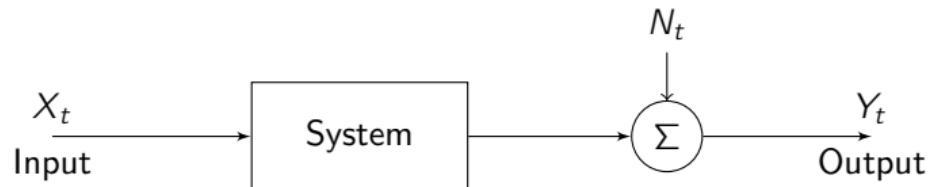
November 10, 2017

Outline of the lecture

- ▶ Chapter 9 – Multivariate time series

Multiple output models

Re-consider the univariate transfer function model:



$$Y_t = h(B)X_t + N_t$$

- ▶ What if there is a feedback from Y to X ?

Closed Loop Models

$$\begin{aligned}Y_t &= h_1(B)X_t + N_{1,t} \\X_t &= h_2(B)Y_t + N_{2,t}\end{aligned}$$

Or:

$$\begin{pmatrix} 1 & -h_1(B) \\ -h_2(B) & 1 \end{pmatrix} \begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} N_{1,t} \\ N_{2,t} \end{pmatrix}$$

- ▶ Two inputs (N_1, N_2);
- ▶ Two outputs (Y, X);
- ▶ Four transfer functions from input to output.

We will look at them individually first.

Transfer from N_1, N_2 to Y :

$$Y = h_1(B)(N_2 + h_2(B)Y) + N_1$$

Z-domain:

$$Y(z) = H_1(z)(N_2(z) + H_2(z)Y(z)) + N_1(z)$$

solving for $Y(z)$:

$$Y(z) = \frac{1}{1 - H_1(z)H_2(z)}N_1(z) + \frac{H_1(z)}{1 - H_1(z)H_2(z)}N_2(z)$$

Transfer functions from N_1, N_2 to Y :

$$\frac{1}{1 - H_1(z)H_2(z)}$$

and

$$\frac{H_1(z)}{1 - H_1(z)H_2(z)}$$

Transfer from N_1 , N_2 to X :

$$X = h_2(B)(N_1 + h_1(B)X) + N_2$$

Z-domain:

$$X(z) = H_2(z)(N_1(z) + H_1(z)X(z)) + N_2(z)$$

solving for $X(z)$:

$$X(z) = \frac{1}{1 - H_1(z)H_2(z)}N_2(z) + \frac{H_2(z)}{1 - H_1(z)H_2(z)}N_1(z)$$

Transfer functions from N_1 , N_2 to X :

$$\frac{H_2(z)}{1 - H_1(z)H_2(z)}$$

and

$$\frac{1}{1 - H_1(z)H_2(z)}$$

Multivariate transfer function

Model equation:

$$\begin{pmatrix} 1 & -h_1(B) \\ -h_2(B) & 1 \end{pmatrix} \begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} N_{1,t} \\ N_{2,t} \end{pmatrix}$$

Model equation in Z-domain:

$$\begin{pmatrix} 1 & -H_1(z) \\ -H_2(z) & 1 \end{pmatrix} \begin{pmatrix} Y(z) \\ X(z) \end{pmatrix} = \begin{pmatrix} N_1(z) \\ N_2(z) \end{pmatrix}$$

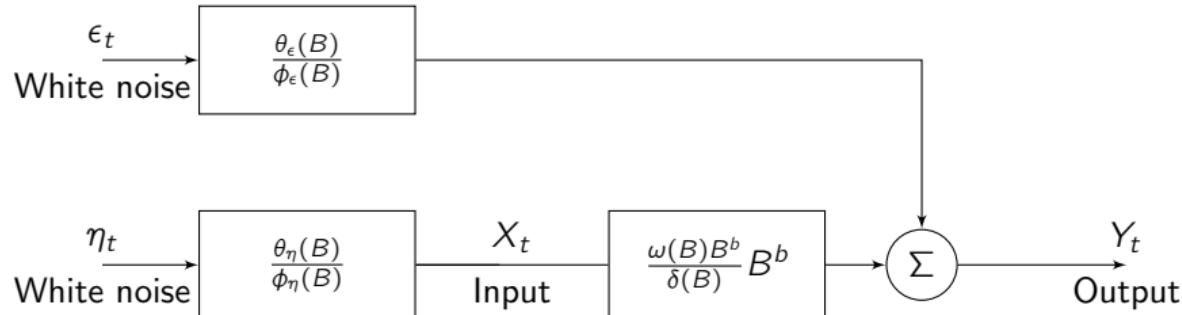
Thus,

$$\begin{pmatrix} Y(z) \\ X(z) \end{pmatrix} = \frac{1}{1 - H_1(z)H_2(z)} \begin{pmatrix} 1 & H_1(z) \\ H_2(z) & 1 \end{pmatrix} \begin{pmatrix} N_1(z) \\ N_2(z) \end{pmatrix}$$

Multivariate transfer function:

$$\begin{pmatrix} \frac{1}{1-H_1(z)H_2(z)} & \frac{H_1(z)}{1-H_1(z)H_2(z)} \\ \frac{H_2(z)}{1-H_1(z)H_2(z)} & \frac{1}{1-H_1(z)H_2(z)} \end{pmatrix}$$

Univ. Transfer function models with ARMA input



$$Y_t = \frac{\omega(B)}{\delta(B)} B^b X_t + \frac{\theta_\epsilon(B)}{\phi_\epsilon(B)} \epsilon_t$$

$$X_t = \frac{\theta_\eta(B)}{\phi_\eta(B)} \eta_t$$

we require $\{\epsilon_t\}$ and $\{\eta_t\}$ to be mutually uncorrelated.

Univ. Transfer function models with ARMA input continued...

$$Y_t = \frac{\omega(B)}{\delta(B)} B^b X_t + \frac{\theta_\varepsilon(B)}{\varphi_\varepsilon(B)} \varepsilon_t$$
$$X_t = \frac{\theta_\eta(B)}{\varphi_\eta(B)} \eta_t$$

which leads to:

$$\delta(B)\varphi_\varepsilon(B)Y_t = \varphi_\varepsilon(B)\omega(B)B^bX_t + \delta(B)\theta_\varepsilon(B)\varepsilon_t$$
$$\varphi_\eta(B)X_t = \theta_\eta(B)\eta_t$$

The term including X_t on the RHS is moved to the LHS:

$$\delta(B)\varphi_\varepsilon(B)Y_t - \varphi_\varepsilon(B)\omega(B)B^bX_t = \delta(B)\theta_\varepsilon(B)\varepsilon_t$$
$$\varphi_\eta(B)X_t = \theta_\eta(B)\eta_t$$

This can be written in matrix notation...

Univ. Transfer function models with ARMA input continued...

from previous slide:

$$\begin{aligned}\delta(B)\varphi_\varepsilon(B)Y_t - \varphi_\varepsilon(B)\omega(B)B^bX_t &= \delta(B)\theta_\varepsilon(B)\varepsilon_t \\ \varphi_\eta(B)X_t &= \theta_\eta(B)\eta_t\end{aligned}$$

Is equivalent to

$$\begin{bmatrix} \delta(B)\varphi_\varepsilon(B) & -\varphi_\varepsilon(B)\omega(B)B^b \\ 0 & \varphi_\eta(B) \end{bmatrix} \begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} \delta(B)\theta_\varepsilon(B) & 0 \\ 0 & \theta_\eta(B) \end{bmatrix} \begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix}$$

For multivariate ARMA-models in general:

- ▶ Replace the off diagonal zeroes by polynomials in B .
- ▶ This introduces feedback from Y to X or reverse
- ▶ Non-zero correlation between ε_t and η_t

Multivariate ARMA models

- ▶ The model can be written

$$\phi(B)(Y_t - c) = \theta(B)\epsilon_t$$

- ▶ The individual time series may have been transformed and differenced
- ▶ The variance-covariance matrix of the multivariate white noise process $\{\epsilon_t\}$ is denoted Σ .
- ▶ The matrices $\phi(B)$ and $\theta(B)$ have elements which are polynomials in the backshift operator
- ▶ The diagonal elements have leading terms of unity
- ▶ The off-diagonal elements have leading terms of zero (i.e. they normally start in B)

Air pollution in cities NO and NO_2

$$\begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \begin{bmatrix} 0.9 & -0.1 \\ 0.4 & 0.8 \end{bmatrix} \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix} + \begin{bmatrix} \xi_{1,t} \\ \xi_{2,t} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 30 & 21 \\ 21 & 23 \end{bmatrix}$$

Matrix formulation:

$$X_t - \begin{bmatrix} 0.9 & -0.1 \\ 0.4 & 0.8 \end{bmatrix} X_{t-1} = \xi_t \quad \text{or} \quad X_t - \phi_1 X_{t-1} = \xi_t$$

Matrix formulation using the backshift operator:

$$\begin{bmatrix} 1 - 0.9B & 0.1B \\ -0.4B & 1 - 0.8B \end{bmatrix} X_t = \xi_t \quad \text{or} \quad \phi(B) X_t = \xi_t$$

Stationarity and Invertability

The multivariate ARMA process

$$\phi(B)(Y_t - c) = \theta(B)\epsilon_t$$

is stationary if

$$\det(\phi(z^{-1})) = 0 \Rightarrow |z| < 1$$

is invertible if

$$\det(\theta(z^{-1})) = 0 \Rightarrow |z| < 1$$

Two formulations (centered data)

Either matrices with polynomials in B as elements:

$$\phi(B)Y_t = \theta(B)\epsilon_t$$

or without B , but with matrices as coefficients:

$$Y_t + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}$$

Auto Covariance Matrix Functions

$$\boldsymbol{\Gamma}_k = E[(\mathbf{Y}_{t-k} - \boldsymbol{\mu}_Y)(\mathbf{Y}_t - \boldsymbol{\mu}_Y)^T] = \boldsymbol{\Gamma}_{-k}^T$$

Example for bivariate case $\mathbf{Y}_t = (Y_{1,t} \ Y_{2,t})^T$:

$$\boldsymbol{\Gamma}_k = \begin{bmatrix} \gamma_{11}(k) & \gamma_{12}(k) \\ \gamma_{21}(k) & \gamma_{22}(k) \end{bmatrix} = \begin{bmatrix} \gamma_{11}(k) & \gamma_{12}(k) \\ \gamma_{12}(-k) & \gamma_{22}(k) \end{bmatrix}$$

We can describe these by plotting

- ▶ each autocovariance or autocorrelation function for $k = 0, 1, 2, \dots$ and
- ▶ each cross-covariance or cross-correlation function for $k = 0, \pm 1, \pm 2, \dots$

The Theoretical Autocovariance Matrix Functions

Using the matrix coefficients ϕ_1, \dots, ϕ_p and $\theta_1, \dots, \theta_q$, together with Σ , the theoretical Γ_k can be calculated:

Pure Autoregressive Models: Γ_k is found from a multivariate version of Theorem 5.10 in the book, which leads to the Yule-Walker equations

Pure Moving Average Models: Γ_k is found from a multivariate version of (5.65) in the book

Autoregressive Moving Average Models: Γ_k is found multivariate versions of (5.100) and (5.101) in the book

- ▶ Examples can be found in the book. (Page 255++)

Autocorrelation for VAR Models

VAR: Vector Auto Regressive - Multivariate AR

$$\phi(B)Y = \varepsilon$$

$$Y_t = -\phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} + \varepsilon_t$$

$$Y_t Y_t^T = -Y_t Y_{t-1}^T \phi_1^T - \cdots - Y_t Y_{t-p}^T \phi_p^T + Y_t \varepsilon_t^T$$

$$\Gamma(0) = -\Gamma(-1)\phi_1^T - \cdots - \Gamma(-p)\phi_p^T + \Sigma$$

$$= -\phi_1 \Gamma(-1)^T - \cdots - \phi_p \Gamma(-p)^T + \Sigma$$

$$Y_{t-k} Y_t^T = -Y_{t-k} Y_{t-1}^T \phi_1^T - \cdots - Y_{t-k} Y_{t-p}^T \phi_p^T + Y_{t-k} \varepsilon_t^T$$

$$\Gamma(k) = -\Gamma(k-1)\phi_1^T - \cdots - \Gamma(k-p)\phi_p^T$$

Multivariate Yule-Walker equations

Y is Vector $AR(k)$ ($VAR(k)$):

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_k Y_{t-k} + \varepsilon_t$$

$$\begin{pmatrix} \Gamma(0) & \Gamma(1)^T & \cdots & \Gamma(k-1)^T \\ \Gamma(1) & \Gamma(0) & \cdots & \Gamma(k-2)^T \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(k-1) & \Gamma(k-2) & \cdots & \Gamma(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_k \end{pmatrix} = \begin{pmatrix} \Gamma(1) \\ \Gamma(2) \\ \vdots \\ \Gamma(k) \end{pmatrix}$$

VAR(1) representation of VARMA processes

Just as in the univariate case, ARMA models may be written as VAR(1) models through stacking:

$$Y_t + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

may for $p \geq q + 1$ be written as

$$\begin{pmatrix} Z_{1,t} \\ Z_{2,t} \\ \vdots \\ Z_{p,t} \end{pmatrix} = \begin{pmatrix} -\phi_1 & I & 0 & \cdots & 0 \\ -\phi_2 & 0 & I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & I \\ -\phi_p & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} Z_{1,t-1} \\ Z_{2,t-1} \\ \vdots \\ Z_{p,t-1} \end{pmatrix} + \begin{pmatrix} I \\ \theta_1 \\ \vdots \\ \theta_{p-1} \end{pmatrix} \varepsilon_t$$

with $Z_{1,t} = Y_t$.

Identification using Autocovariance Matrix Functions

Sample Correlation Matrix Function; R_k near zero for pure moving average processes of order q when $k > q$

Sample Partial Correlation Matrix Function; S_k near zero for pure autoregressive processes of order p when $k > p$

Sample q -conditioned Partial Correlation Matrix Function; $S_k(q)$ near zero for autoregressive moving average processes of order (p, q) when $k > p$ – can be used for univariate processes also. Not so useful in practice.

Identification using (multivariate) prewhitening

- ▶ Fit univariate models to each individual series
- ▶ Investigate the residuals as a multivariate time series
- ▶ The cross correlations can then be compared with $\pm 2/\sqrt{N}$

This is **not** the same form of prewhitening as in Chapter 8

The multivariate model $\phi(B)Y_t = \theta(B)\epsilon_t$ is equivalent to

$$\text{diag}(\det(\phi(B)))Y_t = \text{adj}(\phi(B))\theta(B)\epsilon_t$$

Therefore the corresponding univariate models will have much higher order, so although this is often done in the literature: Don't take this approach!

Multivariate ARMA(p,q) processes (centered data)

- ▶ Matrices with polynomials in B as elements:

$$\phi(B)\mathbf{Y}_t = \theta(B)\boldsymbol{\epsilon}_t$$

So the coefficients are now matrices:

$$\mathbf{Y}_t + \phi_1 \mathbf{Y}_{t-1} + \dots + \phi_p \mathbf{Y}_{t-p} = \boldsymbol{\epsilon}_t + \theta_1 \boldsymbol{\epsilon}_{t-1} + \dots + \theta_q \boldsymbol{\epsilon}_{t-q}$$

- ▶ In general, no analytic solution exists.
- ▶ Therefore, estimation algorithms (or numerical optimization) is necessary.

Estimation procedures

For multivariate ARX(p)

- ▶ Least squares estimation is possible

For multivariate ARMAX(p,q)

- ▶ The Spliid method (Henrik Spliid, 1983)
- ▶ Maximum likelihood

See the book for details.

Highlights

- ▶ Closed loop model as multivariate transfer function

$$\begin{pmatrix} 1 & -h_1(B) \\ -h_2(B) & 1 \end{pmatrix} \begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} N_{1,t} \\ N_{2,t} \end{pmatrix}$$

- ▶ Multivariate ARMA models

$$\boldsymbol{\phi}(B)(\mathbf{Y}_t - \mathbf{c}) = \boldsymbol{\theta}(B)\boldsymbol{\epsilon}_t$$

is stationary if

$$\det(\boldsymbol{\phi}(z^{-1})) = 0 \Rightarrow |z| < 1$$

is invertible if

$$\det(\boldsymbol{\theta}(z^{-1})) = 0 \Rightarrow |z| < 1$$

- ▶ Auto covariance matrix functions

$$\boldsymbol{\Gamma}_k = E[(\mathbf{Y}_{t-k} - \boldsymbol{\mu}_Y)(\mathbf{Y}_t - \boldsymbol{\mu}_Y)^T] = \boldsymbol{\Gamma}_{-k}^T$$

- ▶ All VARMA models can be written as VAR(1)

Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

November 17, 2017

Outline of the lecture

- ▶ Introduction to marima
- ▶ Examples

Reading material

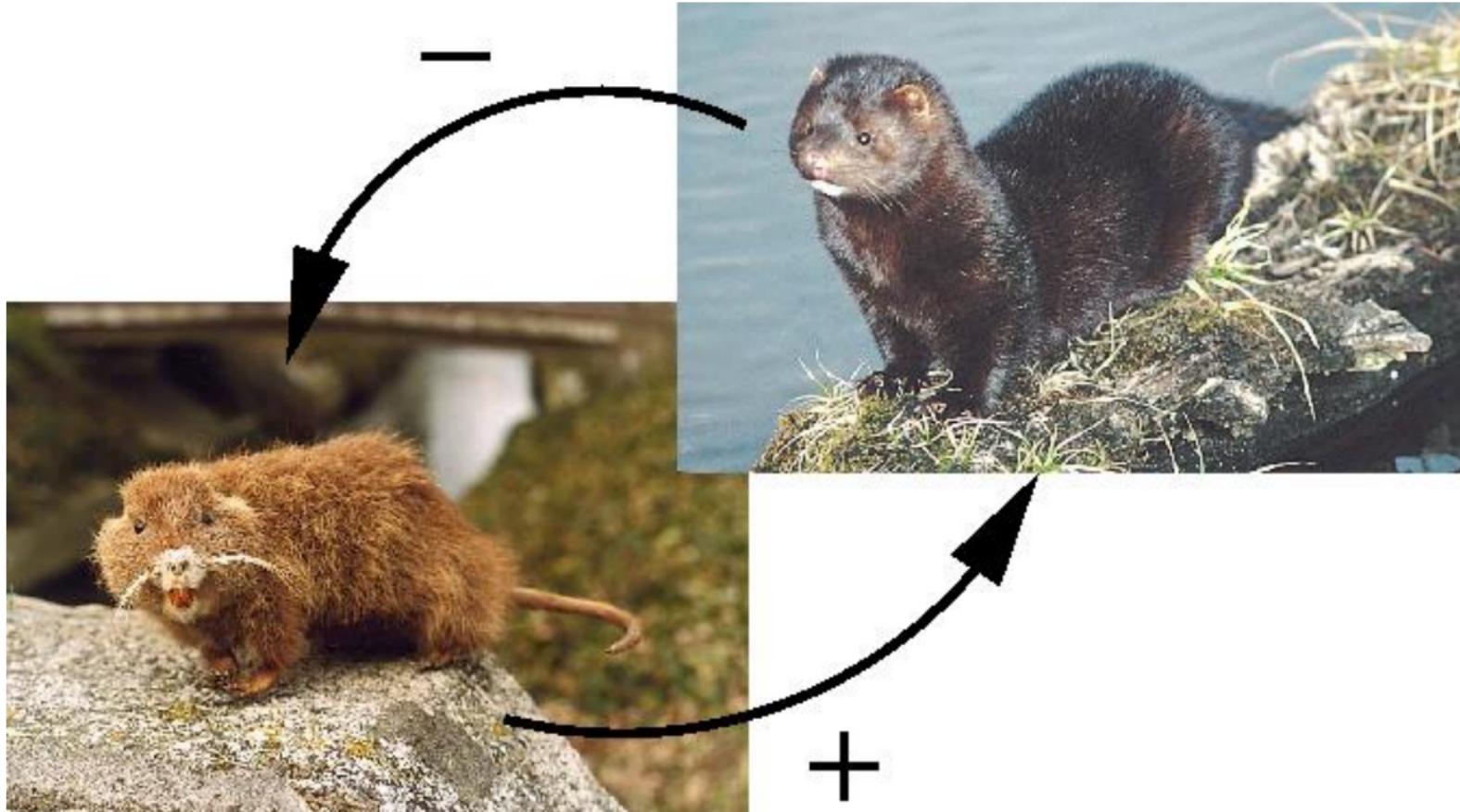
- ▶ A Fast Estimation Method for the Vector Autoregressive Moving Average Model With Exogenous Variables, Henrik Spliid
Journal of the American Statistical Association, Vol. 78, No. 384 (Dec., 1983), pp. 843-849
<http://www.jstor.org/stable/2288194>

Data Analysis

- ▶ The first case relates to the interaction of mink and muskrats
- ▶ Data source:
Jones, J.W. (1914) "Fur-farming in Canada", Commission of Conservation Canada, pp.209–214
<http://www.jstor.org/stable/2346944?origin=JSTOR-pdf>
- ▶ Fur sales, 1850-1911

URL: <http://robjhyndman.com/tsdldata/ecology1/>

Example – Muskrat and Mink skins traded



Model Validation

- ▶ For the individual residual series; all the methods from Chapter 6.
 - ▶ ACF, PACF
 - ▶ Sign tests
 - ▶ Marginal distribution of residuals
 - ▶ Possibly more...
- ▶ with the extension for the cross correlation as mentioned in Chapter 8.

Additional exercise

Use the marima package to fit a good model to the Mink-Muskrat data.

Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

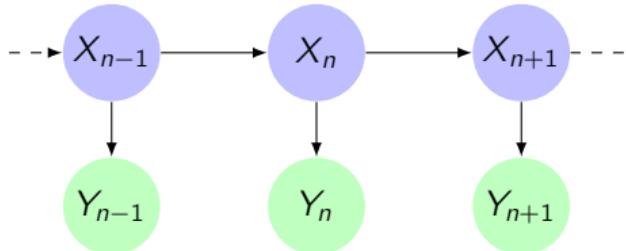
November 24, 2017

Outline of the lecture

State space models, 1st part:

- ▶ Model: Sec. 10.1
- ▶ The Kalman filter: Sec. 10.3
- ▶ An example on application of the Kalman filter.

State space models



- ▶ System model; A full description of the dynamical system (i.e. including the parameters):

$$X_t = f(X_{t-1}) + g(u_{t-1}) + e_{1,t}$$

- ▶ Observations; Noisy measurements of some parts (states) of the system:

$$Y_t = h(X_t) + e_{2,t}$$

- ▶ Goal; reconstruct and predict the state of the system

State space models; examples

- ▶ Estimate the temperature inside a solid block of material when we measure the temperature on the surface (with noise)
- ▶ Noisy measurements of the position of a ship; give a better estimate of the current position
- ▶ PK/PD-modeling: State: Amount of drug in blood, liver, muscles, ... Observations: Amount in blood (with noise), Input: Drug.

Determining the model structure

- ▶ The system model is often based on physical considerations; this often leads to dynamical models consisting of differential equations.
- ▶ An m 'th order differential equation can be formulated as m 1st order differential equations.
- ▶ Sampling such a system leads to a discrete-time state space model.
- ▶ Note that the parameters may change in the discretization.
- ▶ We shall only consider linear state space models.

The linear stochastic state space model

$$\text{System equation: } \mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{B}\mathbf{u}_{t-1} + \mathbf{e}_{1,t}$$

$$\text{Observation equation: } \mathbf{Y}_t = \mathbf{C}\mathbf{X}_t + \mathbf{e}_{2,t}$$

- ▶ \mathbf{X} : State vector
- ▶ \mathbf{Y} : Observation vector
- ▶ \mathbf{u} : Input vector
- ▶ \mathbf{e}_1 : System noise
- ▶ \mathbf{e}_2 : Observation noise
- ▶ $\dim(\mathbf{X}_t) = m$ is called the order of the system
- ▶ $\{\mathbf{e}_{1,t}\}$ and $\{\mathbf{e}_{2,t}\}$ mutually independent white noise
- ▶ $V[\mathbf{e}_1] = \Sigma_1$, $V[\mathbf{e}_2] = \Sigma_2$
- ▶ \mathbf{A} , \mathbf{B} , \mathbf{C} , Σ_1 , and Σ_2 are **known** matrices
- ▶ The state vector contains all information available for future evaluation; the process is a *Markov process*.
- ▶ It is possible to handle time-varying systems as well.

Example Air pollution in cities, NO and NO_2

$$\begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \begin{bmatrix} 0.9 & -0.1 \\ 0.4 & 0.8 \end{bmatrix} \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix} + \begin{bmatrix} \xi_{1,t} \\ \xi_{2,t} \end{bmatrix}, \quad \Sigma_\xi = \begin{bmatrix} 30 & 21 \\ 21 & 23 \end{bmatrix}$$

Suppose that the NO component is missing. How could we formulate this as a state-space model?

States:

$$\mathbf{x}_t = \begin{pmatrix} X_t^{NO} - \mu_{NO} \\ X_t^{NO_2} - \mu_{NO_2} \end{pmatrix}$$

Parameters:

$$A = \begin{pmatrix} 0.9 & -0.1 \\ 0.4 & 0.8 \end{pmatrix}, B = 0, V(\mathbf{e}_1) = \begin{pmatrix} 30 & 21 \\ 21 & 23 \end{pmatrix}$$
$$C = (0 \quad 1), V(\mathbf{e}_2) = 0$$

Example – a falling body I

- ▶ Height above ground: $z(t)$
- ▶ Initial conditions: Position $z(t_0)$ and velocity $z'(t_0)$
- ▶ Physical considerations: $\frac{d^2z}{dt^2} = -g$
- ▶ States: Position $x_1(t) = z(t)$ and velocity $x_2(t) = z'(t)$
- ▶ Only the position is measured $y(t) = x_1(t)$
- ▶ Continuous time description $\mathbf{x}(t) = [x_1(t) \ x_2(t)]^T$:

$$\begin{aligned}\mathbf{x}'(t) &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ -1 \end{bmatrix} g \\ \mathbf{y}(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}(t)\end{aligned}$$

Example – a falling body II

- ▶ Solving the equations:

$$\begin{aligned}x_1(t) &= -\frac{g}{2}(t - t_0)^2 + (t - t_0)x_2(t_0) + x_1(t_0) \\x_2(t) &= -g(t - t_0) + x_2(t_0)\end{aligned}$$

- ▶ Sampling: $t = kT$, $t_0 = (k - 1)T$, and $T = 1$

$$\begin{aligned}\mathbf{x}_k &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} -1/2 \\ -1 \end{bmatrix} g \\ \mathbf{y}_k &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}_k\end{aligned}$$

- ▶ Adding disturbances and measurement noise:

$$\begin{aligned}\mathbf{x}_k &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} -1/2 \\ -1 \end{bmatrix} g + \mathbf{e}_{1,k} \\ \mathbf{y}_k &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}_k + e_{2,k}\end{aligned}$$

Example – a falling body III

Given measurements of the position at time points $1, 2, \dots, k$ we could:

- ▶ **Predict** the future position and velocity $x_{k+n|k}$ ($n > 0$)
- ▶ **Reconstruct** the current position and velocity from noisy measurements $x_{k|k}$
- ▶ **Interpolate or smoothen** to find the best estimate of the position and velocity at a previous time point $x_{k+n|k}$ ($n < 0$) (estimate the path in the state space)

We will focus on reconstruction and prediction

Requirement – observability

In order to predict, reconstruct or interpolate the m -dimensional state in the system

$$\begin{aligned} \mathbf{X}_t &= \mathbf{A}\mathbf{X}_{t-1} + \mathbf{B}\mathbf{u}_{t-1} + \mathbf{e}_{1,t} \\ \mathbf{Y}_t &= \mathbf{C}\mathbf{X}_t + \mathbf{e}_{2,t} \end{aligned}$$

the system must be observable, i.e.

$$\text{rank} \left[\mathbf{C}^T : (\mathbf{CA})^T : \dots : (\mathbf{CA}^{m-1})^T \right] = m.$$

For the falling body (from the discrete-time description of the system):

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

$$\left[\mathbf{C}^T : (\mathbf{CA})^T \right] = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix} : \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \right)^T \right] = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

```
> qr( cbind(t(C), t(C %*% A)) )$rank  
[1] 2
```

The Kalman filter

Initialization:

$$\hat{\boldsymbol{X}}_{1|0} = E[\boldsymbol{X}_1] = \boldsymbol{\mu}_0$$

$$\Sigma_{1|0}^{xx} = V[\boldsymbol{X}_1] = \boldsymbol{V}_0$$

$$\Sigma_{1|0}^{yy} = \boldsymbol{C}\Sigma_{1|0}^{xx}\boldsymbol{C}^T + \boldsymbol{\Sigma}_2$$

For: $t = 1, 2, 3, \dots$

Reconstruction:

$$\boldsymbol{K}_t = \Sigma_{t|t-1}^{xx} \boldsymbol{C}^T \left(\Sigma_{t|t-1}^{yy} \right)^{-1}$$

$$\hat{\boldsymbol{X}}_{t|t} = \hat{\boldsymbol{X}}_{t|t-1} + \boldsymbol{K}_t \left(\boldsymbol{Y}_t - \boldsymbol{C}\hat{\boldsymbol{X}}_{t|t-1} \right)$$

$$\Sigma_{t|t}^{xx} = \Sigma_{t|t-1}^{xx} - \boldsymbol{K}_t \Sigma_{t|t-1}^{yy} \boldsymbol{K}_t^T$$

Prediction:

$$\hat{\boldsymbol{X}}_{t+1|t} = \boldsymbol{A}\hat{\boldsymbol{X}}_{t|t} + \boldsymbol{B}\boldsymbol{u}_t$$

$$\Sigma_{t+1|t}^{xx} = \boldsymbol{A}\Sigma_{t|t}^{xx}\boldsymbol{A}^T + \boldsymbol{\Sigma}_1$$

$$\Sigma_{t+1|t}^{yy} = \boldsymbol{C}\Sigma_{t+1|t}^{xx}\boldsymbol{C}^T + \boldsymbol{\Sigma}_2$$

Multi-step predictions

- ▶ Not part of the Kalman filter as stated above
- ▶ Can be calculated recursively for a given t starting with $k = 1$ for which $\hat{\mathbf{X}}_{t+k|t}$ and $\Sigma_{t+k|t}$ are calculated in the Kalman prediction step.

$$\begin{aligned}\hat{\mathbf{X}}_{t+k+1|t} &= \mathbf{A}\hat{\mathbf{X}}_{t+k|t} + \mathbf{B}\mathbf{u}_{t+k} \\ \Sigma_{t+k+1|t}^{xx} &= \mathbf{A}\Sigma_{t+k|t}^{xx}\mathbf{A}^T + \Sigma_1\end{aligned}$$

- ▶ The future input must be known/assumed.

Naming and history

- ▶ The filter is named after Rudolf E. Kalman, though Thorvald Nicolai Thiele and Peter Swerling actually developed a similar algorithm earlier.
- ▶ It was during a visit of Kalman to the NASA Ames Research Center that he saw the applicability of his ideas to the problem of trajectory estimation for the Apollo program, leading to its incorporation in the Apollo navigation computer.

From http://en.wikipedia.org/wiki/Kalman_filter

The Foundation of the Kalman filter

- ▶ Theorem 2.6 (Linear projection)
- ▶ The theorem is concerned with the random vectors \mathbf{X} and \mathbf{Y} for which the means, variances and covariances are used
- ▶ The state is called \mathbf{X}_t and the observation is called \mathbf{Y}_t and we could write down the theorem for these
- ▶ We have additional information; $\mathcal{Y}_{t-1}^T = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_{t-1}^T)$
- ▶ We include this information by considering the random vectors $\mathbf{X}_t|\mathcal{Y}_{t-1}$ and $\mathbf{Y}_t|\mathcal{Y}_{t-1}$ instead

$$\begin{aligned} E[(\mathbf{X}_t|\mathcal{Y}_{t-1}) | (\mathbf{Y}_t|\mathcal{Y}_{t-1})] &= E[\mathbf{X}_t|\mathbf{Y}_t, \mathcal{Y}_{t-1}] = \\ &E[\mathbf{X}_t|\mathcal{Y}_{t-1}] + \text{Cov}[\mathbf{X}_t, \mathbf{Y}_t|\mathcal{Y}_{t-1}] V^{-1}[\mathbf{Y}_t|\mathcal{Y}_{t-1}] (\mathbf{Y}_t - E[\mathbf{Y}_t|\mathcal{Y}_{t-1}]) \end{aligned}$$

$$\begin{aligned} V[(\mathbf{X}_t|\mathcal{Y}_{t-1}) | (\mathbf{Y}_t|\mathcal{Y}_{t-1})] &= V[\mathbf{X}_t|\mathbf{Y}_t, \mathcal{Y}_{t-1}] = \\ &V[\mathbf{X}_t|\mathcal{Y}_{t-1}] - \text{Cov}[\mathbf{X}_t, \mathbf{Y}_t|\mathcal{Y}_{t-1}] V^{-1}[\mathbf{Y}_t|\mathcal{Y}_{t-1}] \text{Cov}^T[\mathbf{X}_t, \mathbf{Y}_t|\mathcal{Y}_{t-1}] \end{aligned}$$

The Foundation of the Kalman filter II

$$E[\mathbf{X}_t | \mathcal{Y}_t, \mathcal{Y}_{t-1}] =$$

$$E[\mathbf{X}_t | \mathcal{Y}_{t-1}] + \text{Cov}[\mathbf{X}_t, \mathbf{Y}_t | \mathcal{Y}_{t-1}] V^{-1} [\mathbf{Y}_t | \mathcal{Y}_{t-1}] (\mathbf{Y}_t - E[\mathbf{Y}_t | \mathcal{Y}_{t-1}])$$

$$V[\mathbf{X}_t | \mathbf{Y}_t, \mathcal{Y}_{t-1}] =$$

$$V[\mathbf{X}_t | \mathcal{Y}_{t-1}] - \text{Cov}[\mathbf{X}_t, \mathbf{Y}_t | \mathcal{Y}_{t-1}] V^{-1} [\mathbf{Y}_t | \mathcal{Y}_{t-1}] \text{Cov}^T [\mathbf{X}_t, \mathbf{Y}_t | \mathcal{Y}_{t-1}]$$

Using definitions from previous slide the update equations are:

$$\hat{\mathbf{X}}_{t|t} = \hat{\mathbf{X}}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1}^{xy} \left(\boldsymbol{\Sigma}_{t|t-1}^{yy} \right)^{-1} \left(\mathbf{Y}_t - \hat{\mathbf{Y}}_{t|t-1} \right)$$

$$\boldsymbol{\Sigma}_{t|t}^{xx} = \boldsymbol{\Sigma}_{t|t-1}^{xx} - \boldsymbol{\Sigma}_{t|t-1}^{xy} \left(\boldsymbol{\Sigma}_{t|t-1}^{yy} \right)^{-1} \left(\boldsymbol{\Sigma}_{t|t-1}^{xy} \right)^T$$

$$\boldsymbol{K}_t = \boldsymbol{\Sigma}_{t|t-1}^{xy} \left(\boldsymbol{\Sigma}_{t|t-1}^{yy} \right)^{-1}$$

\boldsymbol{K}_t is called the *Kalman gain*, because it determines how much the 1-step prediction error influence the update of the state estimate

The Foundation of the Kalman filter III

The 1-step predictions are obtained directly from the state space model:

$$\begin{aligned}\hat{\mathbf{X}}_{t+1|t} &= \mathbf{A}\hat{\mathbf{X}}_{t|t} + \mathbf{B}\mathbf{u}_t \\ \hat{\mathbf{Y}}_{t+1|t} &= \mathbf{C}\hat{\mathbf{X}}_{t+1|t}\end{aligned}$$

Which results in the prediction errors:

$$\begin{aligned}\tilde{\mathbf{X}}_{t+1|t} &= \mathbf{X}_{t+1} - \hat{\mathbf{X}}_{t+1|t} = \mathbf{A}\tilde{\mathbf{X}}_{t|t} + \mathbf{e}_{1,t+1} \\ \tilde{\mathbf{Y}}_{t+1|t} &= \mathbf{Y}_{t+1} - \hat{\mathbf{Y}}_{t+1|t} = \mathbf{C}\tilde{\mathbf{X}}_{t+1|t} + \mathbf{e}_{2,t+1}\end{aligned}$$

And therefore:

$$\begin{aligned}\Sigma_{t+1|t}^{xx} &= \mathbf{A}\Sigma_{t|t}^{xx}\mathbf{A}^T + \Sigma_1 \\ \Sigma_{t+1|t}^{yy} &= \mathbf{C}\Sigma_{t+1|t}^{xx}\mathbf{C}^T + \Sigma_2 \\ \Sigma_{t+1|t}^{xy} &= \Sigma_{t+1|t}^{xx}\mathbf{C}^T\end{aligned}$$

Example: The falling body revised

Description of the system:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} -1/2 \\ -1 \end{bmatrix} \quad \mathbf{C} = [1 \ 0]$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = [10000]$$

Initialization: Released 10000 m above ground at 0 m/s

$$\hat{\mathbf{X}}_{1|0} = \begin{bmatrix} 10000 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma}_{1|0}^{xx} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \boldsymbol{\Sigma}_{1|0}^{yy} = [10000]$$

Simulation of a falling body – initialization

```
z0 <- 10000
A <- matrix(c(1,0,1,1),nrow=2)
B <- matrix(c(-.5,-1),nrow=2)
C <- matrix(c(1,0),nrow=1)
Sigma1 <- matrix(c(2,.8,.8,1),nrow=2)
Sigma2 <- matrix(10000)
g <- 9.82; N <- 300
X <- matrix(nrow=2,ncol=N) ## Allocating space
X[,1] <- c(z0,0)
Y <- numeric(N)
Y[1] <- C%*%X[,1]+sqrt(Sigma2) %*% rnorm(1)
```

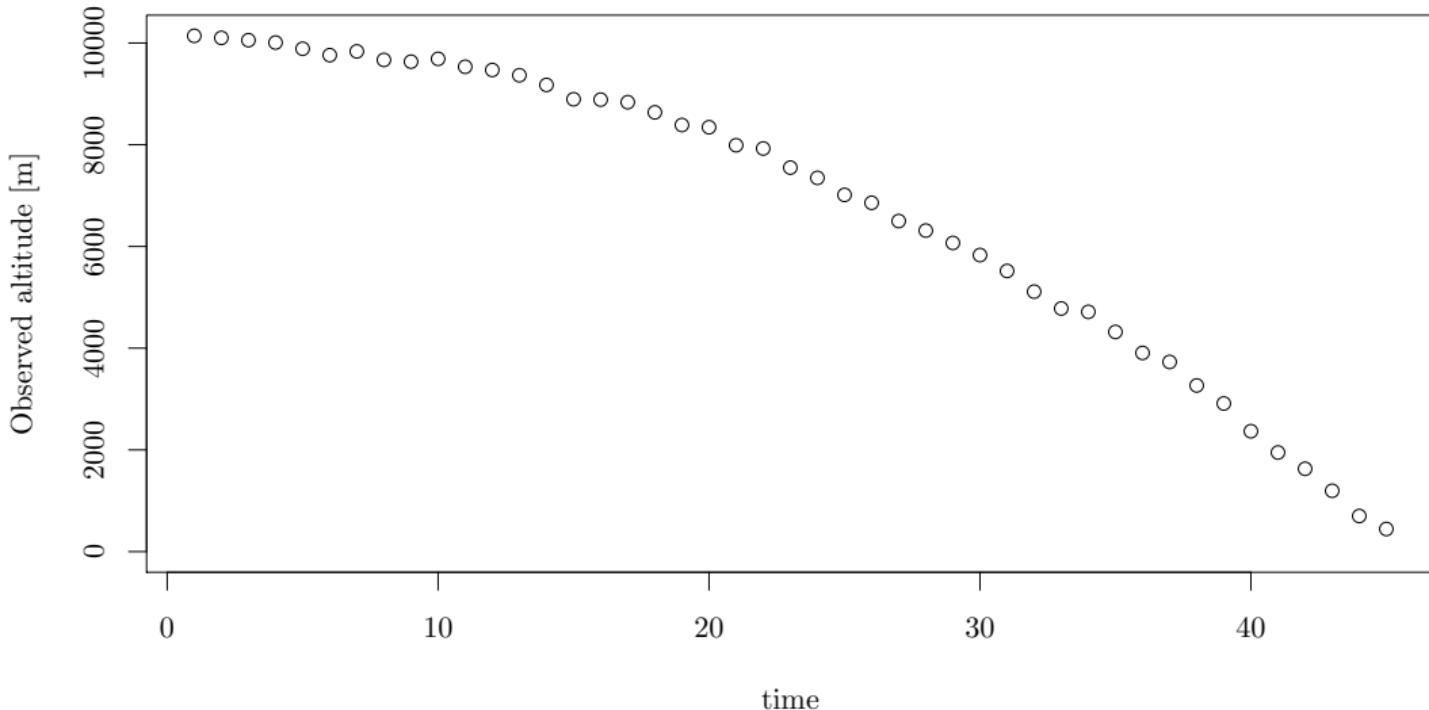
Simulation of a falling body - simulation

```
for (I in 2:N){  
  X[,I] <- A %*% X[,I-1,drop=FALSE] + B%*%g +  
    chol(Sigma1) %*% matrix(rnorm(2),ncol=1)  
  Y[I] <- C %*% X[,I] + sqrt(Sigma2) %*% rnorm(1)  
}  
Nhit <- min(which(X[1,>0))-1  
X <- X[,1:Nhit]  
Y <- Y[1:Nhit]
```

Why the Cholesky factorization?

- ▶ Remember that if $Z \sim N(0, I)$, then $Y = QZ \sim N(0, QQ^T)$.
- ▶ The Cholesky factorization is one way to solve $QQ^T = \Sigma$ for Q .

The falling body – observations



Kalman filter applied to a falling body II

1st observation ($t = 1$): $y_1 = 10171$

Reconstruction: $K_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$

$$\hat{\mathbf{x}}_{1|1} = \begin{bmatrix} 10000 \\ 0 \end{bmatrix} \quad \Sigma_{1|1}^{xx} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Prediction:

$$\hat{\mathbf{x}}_{2|1} = \begin{bmatrix} 9995.09 \\ -9.82 \end{bmatrix} \quad \Sigma_{2|1}^{xx} = \begin{bmatrix} 2 & 0.8 \\ 0.8 & 1 \end{bmatrix} \quad \Sigma_{2|1}^{yy} = [10002]$$

Kalman filter applied to a falling body III

2nd observation ($t = 2$): $y_2 = 10046$

Reconstruction: $K_2 = [\begin{array}{cc} 0.00020 & 0.00008 \end{array}]^T$

$$\hat{\mathbf{x}}_{2|2} = \begin{bmatrix} 9995.1 \\ -9.81 \end{bmatrix} \quad \Sigma_{2|2}^{xx} = \begin{bmatrix} 2 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Prediction:

$$\hat{\mathbf{x}}_{3|2} = \begin{bmatrix} 9980.38 \\ -19.63 \end{bmatrix} \quad \Sigma_{3|2}^{xx} = \begin{bmatrix} 6.6 & 2.6 \\ 2.6 & 2 \end{bmatrix} \quad \Sigma_{3|2}^{yy} = [\begin{array}{c} 10006.6 \end{array}]$$

Kalman filter applied to a falling body IV

3rd observation ($t = 3$): $y_3 = 10082$

Reconstruction: $K_3 = [\begin{array}{cc} 0.00066 & 0.00026 \end{array}]^T$

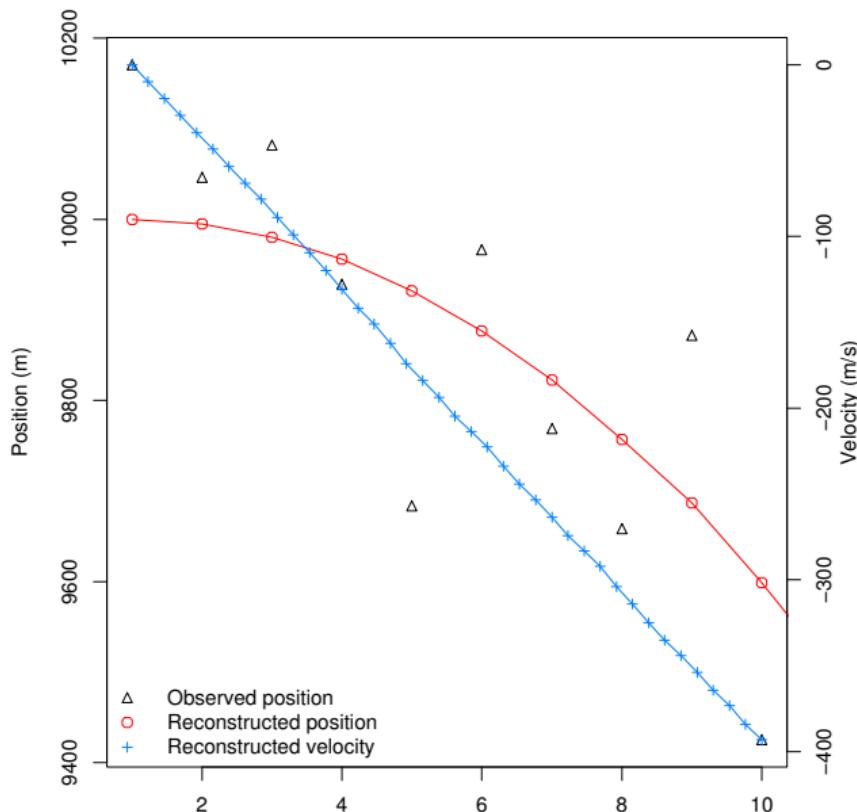
$$\hat{\mathbf{X}}_{3|3} = \begin{bmatrix} 9980.45 \\ -19.6 \end{bmatrix} \quad \Sigma_{3|3}^{xx} = \begin{bmatrix} 6.59 & 2.6 \\ 2.6 & 2 \end{bmatrix}$$

Prediction:

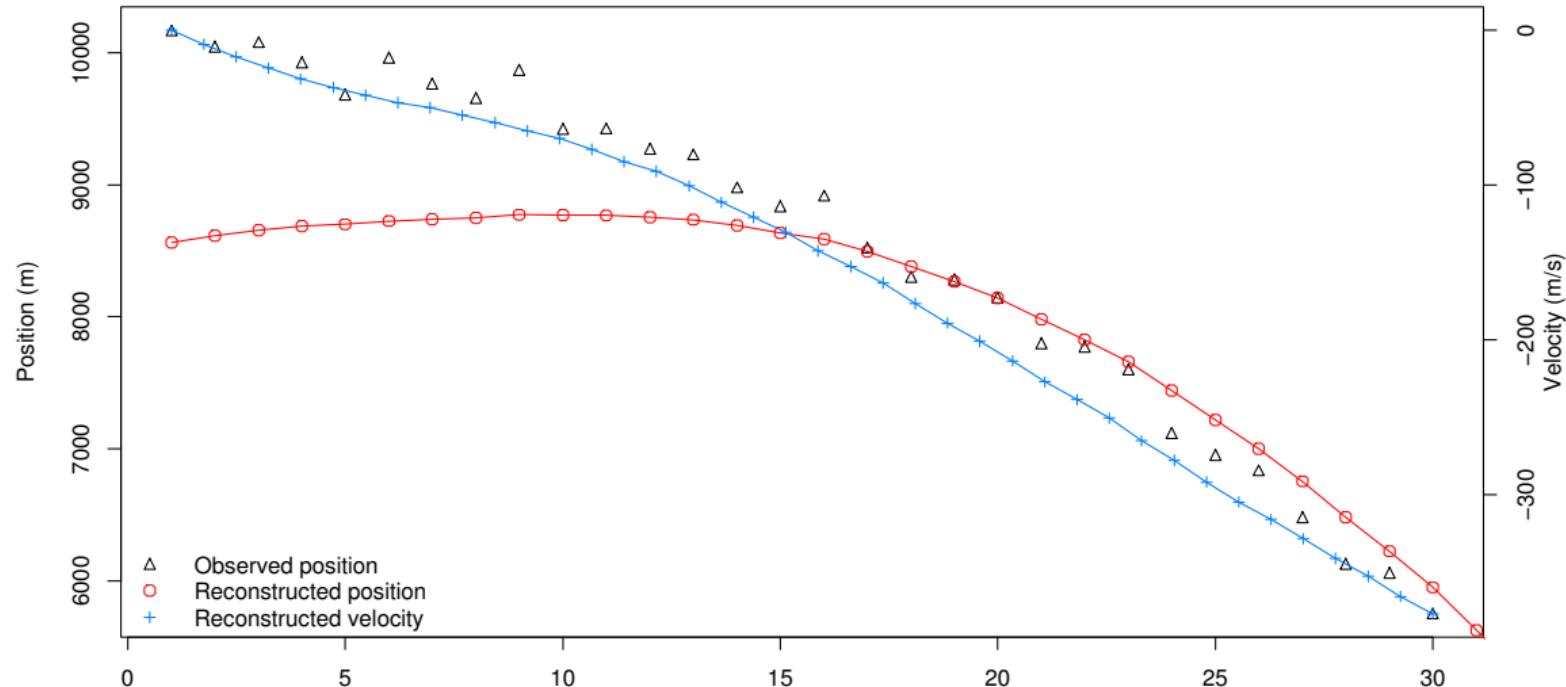
$$\hat{\mathbf{X}}_{4|3} = \begin{bmatrix} 9955.94 \\ -29.41 \end{bmatrix} \quad \Sigma_{4|3}^{xx} = \begin{bmatrix} 15.79 & 5.4 \\ 5.4 & 3 \end{bmatrix} \quad \Sigma_{4|3}^{yy} = [\begin{array}{c} 10015.79 \end{array}]$$

Falling body – the 10 first time points

Add uncertainty !!!



Falling body – wrong initial state



Highlights

- ▶ State space model:

$$\text{System equation: } \mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{B}\mathbf{u}_{t-1} + \mathbf{e}_{1,t}$$

$$\text{Observation equation: } \mathbf{Y}_t = \mathbf{C}\mathbf{X}_t + \mathbf{e}_{2,t}$$

- ▶ Sampling
- ▶ Observability

$$\text{rank} \left[\mathbf{C}^T : (\mathbf{CA})^T : \dots : (\mathbf{CA}^{m-1})^T \right] = m.$$

- ▶ Kalman filter
 - ▶ Reconstruction
 - ▶ Prediction

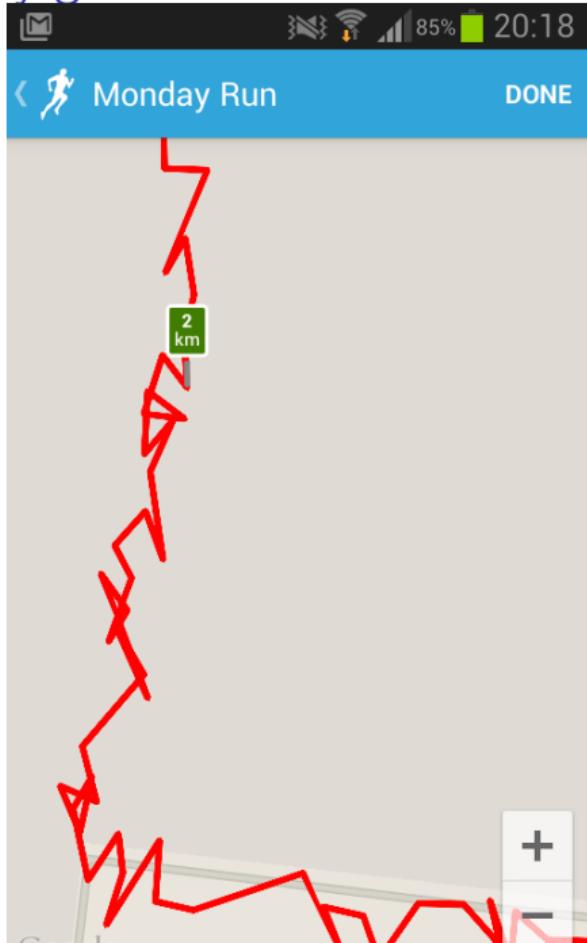
Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

December 1, 2017

Where did he actually go?



Outline of the lecture

State space models, 2nd part:

- ▶ ARMA-models on state space form, Sec. 10.4
- ▶ Example: Random walk with measurement noise
- ▶ The Kalman filter when some observations are missing
- ▶ ML-estimates in state space models, Sec. 10.6
- ▶ Time-varying systems
- ▶ Example AR(1) through measurement noise

Cursory material:

- ▶ Signal extraction, Sec. 10.4.1
- ▶ Time series with missing observations, Sec. 10.5

The linear stochastic state space model

$$\text{System equation: } \mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{B}\mathbf{u}_{t-1} + \mathbf{e}_{1,t}$$

$$\text{Observation equation: } \mathbf{Y}_t = \mathbf{C}\mathbf{X}_t + \mathbf{e}_{2,t}$$

- ▶ \mathbf{X} : State vector
- ▶ \mathbf{Y} : Observation vector
- ▶ \mathbf{u} : Input vector
- ▶ \mathbf{e}_1 : System noise
- ▶ \mathbf{e}_2 : Observation noise
- ▶ $\dim(\mathbf{X}_t) = m$ is called the order of the system
- ▶ $\{\mathbf{e}_{1,t}\}$ and $\{\mathbf{e}_{2,t}\}$ mutually independent white noise
- ▶ $V[\mathbf{e}_1] = \Sigma_1$, $V[\mathbf{e}_2] = \Sigma_2$
- ▶ \mathbf{A} , \mathbf{B} , \mathbf{C} , Σ_1 , and Σ_2 are **known** matrices
- ▶ The state vector contains all information available for future evaluation; the state vector is a *Markov process*.

The ARMA(p, q) model as a state space model

$$Y_t + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

State space form:

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{G}\boldsymbol{\varepsilon}_t$$

$$\mathbf{Y}_t = \mathbf{C}\mathbf{X}_t$$

Or:

$$\mathbf{X}_t = \begin{bmatrix} -\phi_1 & 1 & 0 & \cdots & 0 \\ -\phi_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\phi_{d-1} & 0 & 0 & 0 & 1 \\ -\phi_d & 0 & 0 & \cdots & 0 \end{bmatrix} \mathbf{X}_{t-1} + \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{d-1} \end{pmatrix} \boldsymbol{\varepsilon}_t$$

$$\mathbf{C} = [1 \ 0 \ \cdots \ 0]$$

where $d = \max(p, q + 1)$ and any extra parameter is fixed to zero.

For multivariate processes, just plug in matrices, and use $/$ in stead of 1.

Random walk with measurement noise

Consider the state space model

$$X_t = X_{t-1} + \eta_t$$

$$Y_t = X_t + \varepsilon_t$$

where $\{\eta_t\}$ and $\{\varepsilon_t\}$ are white noise processes with $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$ and $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

- ▶ $\{X_t\}$ is a random walk, that is not directly observed.
- ▶ The observations, $\{Y_t\}$, are influenced by measurement noise.
- ▶ What is the ARIMA structure of the Y process?
- ▶ Hint:

$$\nabla Y_t = \nabla X_t + \nabla \varepsilon_t = \eta_t + \varepsilon_t - \varepsilon_{t-1}$$

Random walk with measurement noise II

$$\nabla Y_t = \eta_t + \varepsilon_t - \varepsilon_{t-1}$$

ACF for ∇Y_t :

$$\rho(k) = \begin{cases} 1 & k = 0 \\ -\sigma_\varepsilon^2 / (\sigma_\eta^2 + 2\sigma_\varepsilon^2) & k = 1 \\ 0 & k > 1 \end{cases}$$

This is the ACF of an $MA(1)$ process; thus, Y is $IMA(1, 1)$

Alternative formulation:

$$\nabla Y_t = \xi_t + \theta_1 \xi_{t-1}, \quad \theta_1 < 0,$$

where ξ is white noise with variance σ_ξ^2 .

Random walk with measurement noise III

Parameter relations in the two formulations, found by equaling the ACF expressions:

$$\begin{aligned}(1 + \theta_1^2)\sigma_\xi^2 &= \sigma_\eta^2 + 2\sigma_\varepsilon^2 \\ \theta_1\sigma_\xi^2 &= -\sigma_\varepsilon^2\end{aligned}$$

- ▶ The ARMA process coefficients for the MA-parts are covariance parameters in the State Space formulation.
- ▶ The ARMA representation may be used to derive estimates for Σ_1 , Σ_2 .

The linear stochastic state space model

$$\text{System equation: } \mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{B}\mathbf{u}_{t-1} + \mathbf{e}_{1,t}$$

$$\text{Observation equation: } \mathbf{Y}_t = \mathbf{C}\mathbf{X}_t + \mathbf{e}_{2,t}$$

- ▶ \mathbf{X} : State vector
- ▶ \mathbf{Y} : Observation vector
- ▶ \mathbf{u} : Input vector
- ▶ \mathbf{e}_1 : System noise
- ▶ \mathbf{e}_2 : Observation noise
- ▶ $\dim(\mathbf{X}_t) = m$ is called the order of the system
- ▶ $\{\mathbf{e}_{1,t}\}$ and $\{\mathbf{e}_{2,t}\}$ mutually independent white noise
- ▶ $V[\mathbf{e}_1] = \Sigma_1$, $V[\mathbf{e}_2] = \Sigma_2$
- ▶ \mathbf{A} , \mathbf{B} , \mathbf{C} , Σ_1 , and Σ_2 are **known** matrices

The Kalman filter

Initialization

$$\hat{\mathbf{X}}_{1|0} = E[\mathbf{X}_1] = \boldsymbol{\mu}_0$$

$$\Sigma_{1|0}^{xx} = V[\mathbf{X}_1] = \mathbf{V}_0$$

$$\Sigma_{1|0}^{yy} = \mathbf{C}\Sigma_{1|0}^{xx}\mathbf{C}^T + \Sigma_2$$

For: $t = 1, 2, 3, \dots$

Reconstruction:

$$\boldsymbol{\kappa}_t = \Sigma_{t|t-1}^{xx} \mathbf{C}^T \left(\Sigma_{t|t-1}^{yy} \right)^{-1}$$

$$\hat{\mathbf{X}}_{t|t} = \hat{\mathbf{X}}_{t|t-1} + \boldsymbol{\kappa}_t \left(\mathbf{Y}_t - \mathbf{C}\hat{\mathbf{X}}_{t|t-1} \right)$$

$$\Sigma_{t|t}^{xx} = \Sigma_{t|t-1}^{xx} - \boldsymbol{\kappa}_t \Sigma_{t|t-1}^{yy} \boldsymbol{\kappa}_t^T$$

Prediction:

$$\hat{\mathbf{X}}_{t+1|t} = \mathbf{A}\hat{\mathbf{X}}_{t|t} + \mathbf{B}\mathbf{u}_t$$

$$\Sigma_{t+1|t}^{xx} = \mathbf{A}\Sigma_{t|t}^{xx}\mathbf{A}^T + \Sigma_1$$

$$\Sigma_{t+1|t}^{yy} = \mathbf{C}\Sigma_{t+1|t}^{xx}\mathbf{C}^T + \Sigma_2$$

- ▶ What happens if the observation \mathbf{Y}_t is missing for some t ?

Estimation in ARMA(p, q)-models using the KF

- ▶ Using the Kalman filter we can get the mean and variance of the one-step predictions of the observations:

$$\begin{aligned}\hat{Y}_{t+1|t} &= C\hat{X}_{t+1|t} \\ \Sigma_{t+1|t}^{yy} &= C\Sigma_{t+1|t}^{xx}C^T + \Sigma_2\end{aligned}$$

- ▶ The Kalman filter can handle missing observations
- ▶ An ARMA(p, q)-model can be written as a state space model
- ▶ This gives us a way of calculating ML-estimates in the ARMA(p, q)-model even when some observations are missing.

The ARMA(p, q) model as a state space model

$$Y_t + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

State space form:

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t$$

$$\mathbf{Y}_t = \mathbf{C}\mathbf{X}_t$$

For $p \geq q$:

$$\mathbf{X}_t = \begin{bmatrix} -\phi_1 & 1 & 0 & \cdots & 0 \\ -\phi_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\phi_{d-1} & 0 & 0 & 0 & 1 \\ -\phi_d & 0 & 0 & \cdots & 0 \end{bmatrix} \mathbf{X}_{t-1} + \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{d-1} \end{pmatrix} \boldsymbol{\varepsilon}_t$$

$$\mathbf{C} = [1 \ 0 \ \cdots \ 0]$$

where $d = \max(p, q + 1)$ and any extra parameter is fixed to zero.

ML-estimates in state space models

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{G}\mathbf{e}_{1,t}$$

$$\mathbf{Y}_t = \mathbf{C}\mathbf{X}_t + \mathbf{e}_{2,t}$$

- ▶ $\{\mathbf{e}_{1,t}\}$ and $\{\mathbf{e}_{2,t}\}$ are mutually uncorrelated normally distributed white noise
- ▶ $V(\mathbf{e}_{1,t}) = \Sigma_1$ and $V(\mathbf{e}_{2,t}) = \Sigma_2$
- ▶ For ARMA(p, q)-models we have \mathbf{A} , \mathbf{C} , and \mathbf{G} as stated on the previous slide. Furthermore, $\mathbf{e}_{1,t} = \varepsilon_t$, $\Sigma_1 = \sigma_\varepsilon^2$, and $\Sigma_2 = 0$

Maximum Likelihood Estimates

- ▶ Let \mathcal{Y}_{N^*} contain the available observations and let θ contain the parameters of the model
- ▶ The likelihood function is the density of the random vector corresponding to the observations and given the set of parameters:

$$L(\theta; \mathcal{Y}_{N^*}) = f(\mathcal{Y}_{N^*} | \theta)$$

- ▶ The ML-estimates is found by selecting θ so that the density function is as large as possible at the actual observations
- ▶ The random variables $\mathbf{Y}_{N^*} | \mathcal{Y}_{N^*-1}$ and \mathcal{Y}_{N^*-1} are independent:

$$\begin{aligned} L(\theta; \mathcal{Y}_{N^*}) &= f(\mathcal{Y}_{N^*} | \theta) = f(\mathbf{Y}_{N^*} | \mathcal{Y}_{N^*-1}, \theta) f(\mathcal{Y}_{N^*-1} | \theta) \\ &= f(\mathbf{Y}_{N^*} | \mathcal{Y}_{N^*-1}, \theta) f(\mathbf{Y}_{N^*-1} | \mathcal{Y}_{N^*-2}, \theta) \cdots f(\mathbf{Y}_1 | \theta) \end{aligned}$$

- ▶ The conditional densities can be found using the Kalman filter

MLE / KF – Prediction

- ▶ Assume that at time t we have:

$$\hat{\mathbf{X}}_{t|t} = E[\mathbf{X}_t | \mathcal{Y}_t] \quad \text{and} \quad \Sigma_{t|t}^{xx} = V[\mathbf{X}_t | \mathcal{Y}_t]$$

- ▶ Using the model we obtain predictions for time $t + 1$:

$$\begin{aligned}\hat{\mathbf{X}}_{t+1|t} &= A\hat{\mathbf{X}}_{t|t} \\ \Sigma_{t+1|t}^{xx} &= A\Sigma_{t|t}^{xx}A^T + G\Sigma_1G^T \\ \hat{\mathbf{Y}}_{t+1|t} &= C\hat{\mathbf{X}}_{t+1|t} \\ \Sigma_{t+1|t}^{yy} &= C\Sigma_{t+1|t}^{xx}C^T + \Sigma_2\end{aligned}$$

- ▶ Due to the normality of the white noise process $f(\mathbf{Y}_{t+1} | \mathcal{Y}_t, \theta)$ is then the (multivariate) normal density (see Chapter 2) with mean $\hat{\mathbf{Y}}_{t+1|t}$ and variance-covariance $\Sigma_{t+1|t}^{yy}$ ($= R_{t+1}$)

MLE / KF – Reconstruction

At time $t + 1$ there are two possibilities for the reconstruction part:

The observation \mathbf{Y}_{t+1} is available:

We update the state estimate using the reconstruction step of the Kalman Filter:

$$\mathbf{K}_{t+1} = \boldsymbol{\Sigma}_{t+1|t}^{xx} \mathbf{C}^T \left(\boldsymbol{\Sigma}_{t+1|t}^{yy} \right)^{-1}$$

$$\hat{\mathbf{X}}_{t+1|t+1} = \hat{\mathbf{X}}_{t+1|t} + \mathbf{K}_{t+1} \left(\mathbf{Y}_{t+1} - \hat{\mathbf{Y}}_{t+1|t} \right)$$

$$\boldsymbol{\Sigma}_{t+1|t+1}^{xx} = \boldsymbol{\Sigma}_{t+1|t}^{xx} - \mathbf{K}_{t+1} \boldsymbol{\Sigma}_{t+1|t}^{yy} \mathbf{K}_{t+1}^T$$

The observation \mathbf{Y}_{t+1} is missing:

We get no new information and we use:

$$\hat{\mathbf{X}}_{t+1|t+1} = \hat{\mathbf{X}}_{t+1|t}$$

$$\boldsymbol{\Sigma}_{t+1|t+1}^{xx} = \boldsymbol{\Sigma}_{t+1|t}^{xx}$$

MLE / KF – The likelihood function

- ▶ Using the prediction errors and variances

$$\begin{aligned}\tilde{\mathbf{Y}}_i &= \mathbf{Y}_i - \hat{\mathbf{Y}}_{i|i-1} \\ R_i &= \Sigma_{i|i-1}^{yy}\end{aligned}$$

- ▶ The likelihood function can be expressed as

$$L(\boldsymbol{\theta}; \mathcal{Y}_{N^*}) = \prod_{i=1}^{N^*} [(2\pi)^m \det \mathbf{R}_i]^{-\frac{1}{2}} \exp \left[-\frac{1}{2} \tilde{\mathbf{Y}}_i^T \mathbf{R}_i^{-1} \tilde{\mathbf{Y}}_i \right]$$

- ▶ In practice optimization is based on

$$\log L(\boldsymbol{\theta}; \mathcal{Y}_{N^*}) = -\frac{1}{2} \sum_{i=1}^N \left(\log \det \mathbf{R}_i + \tilde{\mathbf{Y}}_i^T \mathbf{R}_i^{-1} \tilde{\mathbf{Y}}_i \right) + c$$

- ▶ The variance of the estimates can be approximated by the 2nd order derivatives of the log-likelihood.

MLE / KF IV – Initialization

- ▶ The only outstanding issue is “prediction” of \hat{Y}_1 , i.e. calculation of $\hat{Y}_{1|0}$
- ▶ This can be done by setting $\hat{X}_{0|0} = \mathbf{0}$ and $\Sigma_{0|0}^{xx} = \alpha I$, where I is the identity matrix and α is a ‘large’ constant (we don’t know what it is)
- ▶ Alternatively, we can fix the initial state $\hat{X}_{0|0}$ and set $\Sigma_{0|0}^{xx} = \mathbf{0}$, whereby $\Sigma_{1|0}^{xx} = G\Sigma_1 G^T$
- ▶ Or combinations thereof - recommended
- ▶ The important part is that the (un-)certainty of $\hat{X}_{0|0}$ is reflected in $\Sigma_{0|0}^{xx}$.

Autocovariance functions with missing data

Define the observation indicator a as

$$a(t) = \begin{cases} 1 & \text{if } Y_t \text{ is observed;} \\ 0 & \text{if } Y_t \text{ is missing.} \end{cases}$$

Define similarly

$$C_a(k) = \frac{1}{N} \sum_{t=1}^{N-|k|} a(t)a(t+|k|).$$

- ▶ a indicates which data points are present
- ▶ For large N , $C_a(k)$ measures the fraction of pairs of data k time steps away from each other that are observed.

Autocovariance functions with missing data II

The indicator, $a(t)$ is used to define and estimate of the mean of $\{Y_t\}$

$$\bar{\mu}_y = \frac{\sum_{t=1}^N a(t) Y_t}{\sum_{t=1}^N a(t)}$$

- $\bar{\mu}_y$ is the mean of the observed Y_t 's.

And defining:

$$C_a^\square(k) = \frac{1}{N} \sum_{t=1}^{N-|k|} a(t)a(t+|k|)(Y_t - \bar{\mu}_y)(Y_{t+|k|} - \bar{\mu}_y).$$

- Note: $C_a^\square(k)$ leave pairs out if one of the observations is missing.
- The sample autocovariance is estimated by:

$$C_{YY}(k) = \frac{C_a^\square(k)}{C_a(k)}$$

- These estimates are available with the acf function in R, by using
 > `acf(...,na.action=na.omit)`

Time-varying systems

$$\text{System equation: } \mathbf{X}_t = \mathbf{A}_t \mathbf{X}_{t-1} + \mathbf{B}_t \mathbf{u}_{t-1} + \mathbf{e}_{1,t}$$

$$\text{Observation equation: } \mathbf{Y}_t = \mathbf{C}_t \mathbf{X}_t + \mathbf{e}_{2,t}$$

- ▶ \mathbf{X} : State vector
- ▶ \mathbf{Y} : Observation vector
- ▶ \mathbf{u} : Input vector
- ▶ \mathbf{e}_1 : System noise
- ▶ \mathbf{e}_2 : Observation noise
- ▶ $\dim(\mathbf{X}_t) = m$ is called the order of the system
- ▶ $\{\mathbf{e}_{1,t}\}$ and $\{\mathbf{e}_{2,t}\}$ mutually independent white noise
- ▶ $V[\mathbf{e}_{1,t}] = \Sigma_{1,t}$, $V[\mathbf{e}_{2,t}] = \Sigma_{2,t}$
- ▶ \mathbf{A}_t , \mathbf{B}_t , \mathbf{C}_t , $\Sigma_{1,t}$, and $\Sigma_{2,t}$ are **known** matrices at any point in time

The Kalman filter for time varying systems

Initialization

$$\begin{aligned}\hat{\mathbf{X}}_{1|0} &= E[\mathbf{X}_1] = \boldsymbol{\mu}_0 \\ \boldsymbol{\Sigma}_{1|0}^{xx} &= V[\mathbf{X}_1] = \mathbf{V}_0 \Rightarrow \\ \boldsymbol{\Sigma}_{1|0}^{yy} &= \mathbf{C}_1 \boldsymbol{\Sigma}_{1|0}^{xx} \mathbf{C}_1^T + \boldsymbol{\Sigma}_{2,1}\end{aligned}$$

For: $t = 1, 2, 3, \dots$

Reconstruction:

$$\begin{aligned}\boldsymbol{\mathcal{K}}_t &= \boldsymbol{\Sigma}_{t|t-1}^{xx} \mathbf{C}_t^T \left(\boldsymbol{\Sigma}_{t|t-1}^{yy} \right)^{-1} \\ \hat{\mathbf{X}}_{t|t} &= \hat{\mathbf{X}}_{t|t-1} + \boldsymbol{\mathcal{K}}_t \left(\mathbf{Y}_t - \mathbf{C}_t \hat{\mathbf{X}}_{t|t-1} \right) \\ \boldsymbol{\Sigma}_{t|t}^{xx} &= \boldsymbol{\Sigma}_{t|t-1}^{xx} - \boldsymbol{\mathcal{K}}_t \boldsymbol{\Sigma}_{t|t-1}^{yy} \boldsymbol{\mathcal{K}}_t^T\end{aligned}$$

Prediction:

$$\begin{aligned}\hat{\mathbf{X}}_{t+1|t} &= \mathbf{A}_{t+1} \hat{\mathbf{X}}_{t|t} + \mathbf{B}_{t+1} \mathbf{u}_t \\ \boldsymbol{\Sigma}_{t+1|t}^{xx} &= \mathbf{A}_{t+1} \boldsymbol{\Sigma}_{t|t}^{xx} \mathbf{A}_{t+1}^T + \boldsymbol{\Sigma}_{1,t+1} \\ \boldsymbol{\Sigma}_{t+1|t}^{yy} &= \mathbf{C}_{t+1} \boldsymbol{\Sigma}_{t+1|t}^{xx} \mathbf{C}_{t+1}^T + \boldsymbol{\Sigma}_{2,t+1}\end{aligned}$$

Example: An AR(1) and obs. noise

The heat transfer from a certain body to its surroundings is dominated by conduction. The temperature of the body is given by

$$\frac{dT}{dt} = \frac{1}{R}(T_{surr} - T)$$

Let the surrounding temperature be constantly 0. Then

$$\frac{dT}{dt} = -\frac{1}{R}(T) = aT$$

The solution to the differential equation is

$$T = T_0 e^{at}$$

A discretization of this yields

$$T_{t+1} = e^{a\Delta t} \cdot T_t$$

We know in reality, such a process is influenced by noise:

$$T_{t+1} = -\phi \cdot T_t + e_t, \quad e_t \sim \mathcal{N}(0, \sigma_e^2)$$

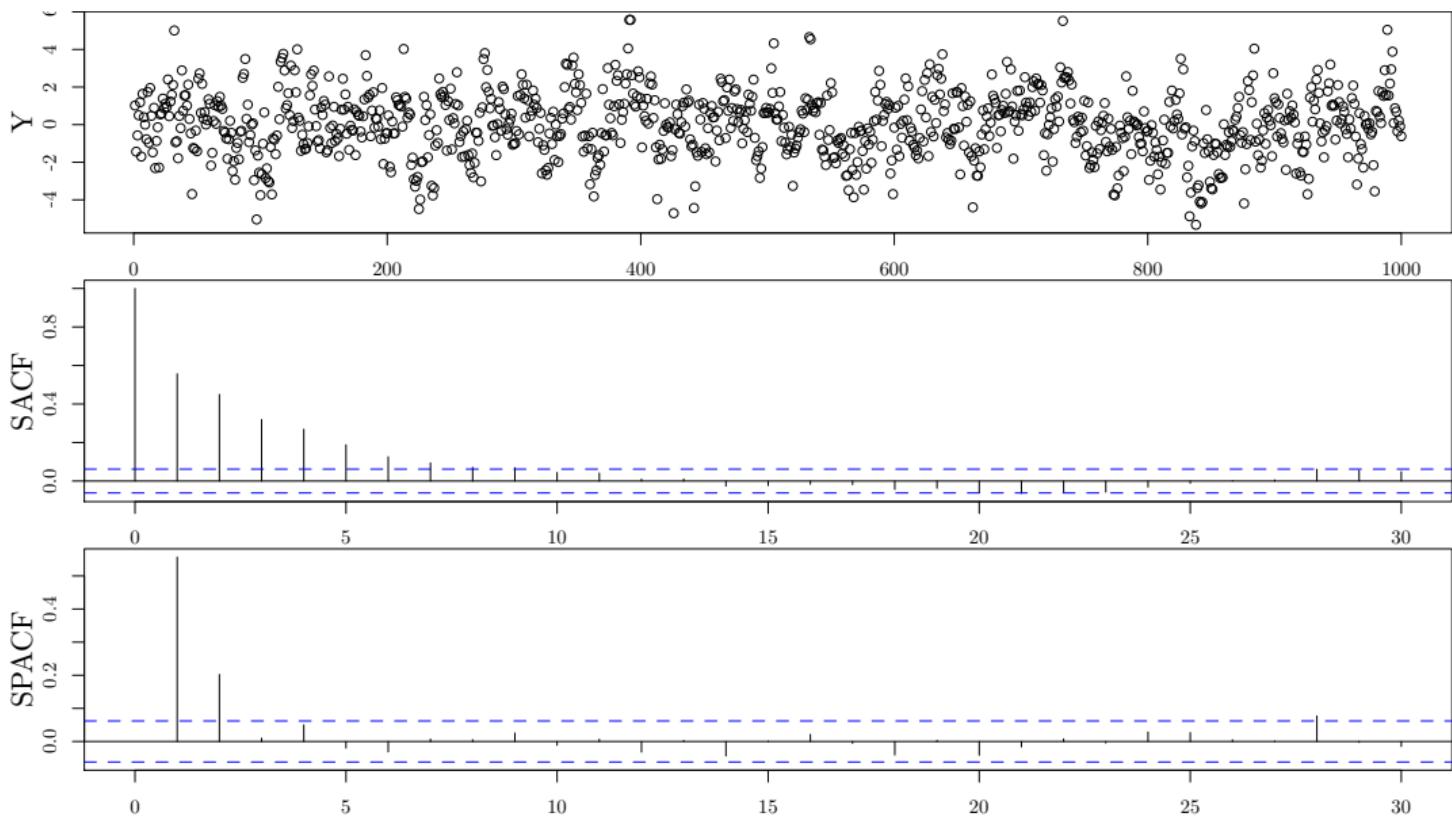
We measure the temperature at each time step:

$$Y_t = T_t + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$$

Example: An AR(1) and obs. noise – Simulation

```
a <- 0.8
N <- 1000
X <- numeric(N)
X[1] <- 1
for (i in 2:N){
  X[i] <- a * X[i-1] + rnorm(1, sd=0.8)
}
Y <- X + rnorm(N, sd=1.0)
```

Example: An AR(1) and obs. noise – What process is this?



Highlights

- ▶ ARMA models on State space form
- ▶ Kalman filter
 - ▶ Handling missing values
 - ▶ Prediction
 - ▶ Maximum likelihood estimation of parameters
 - ▶ Comparing models using likelihood

Time Series Analysis

Lasse Engbo Christiansen

Department of Applied Mathematics and Computer Science
Technical University of Denmark

December 8, 2017

Outline of the lecture

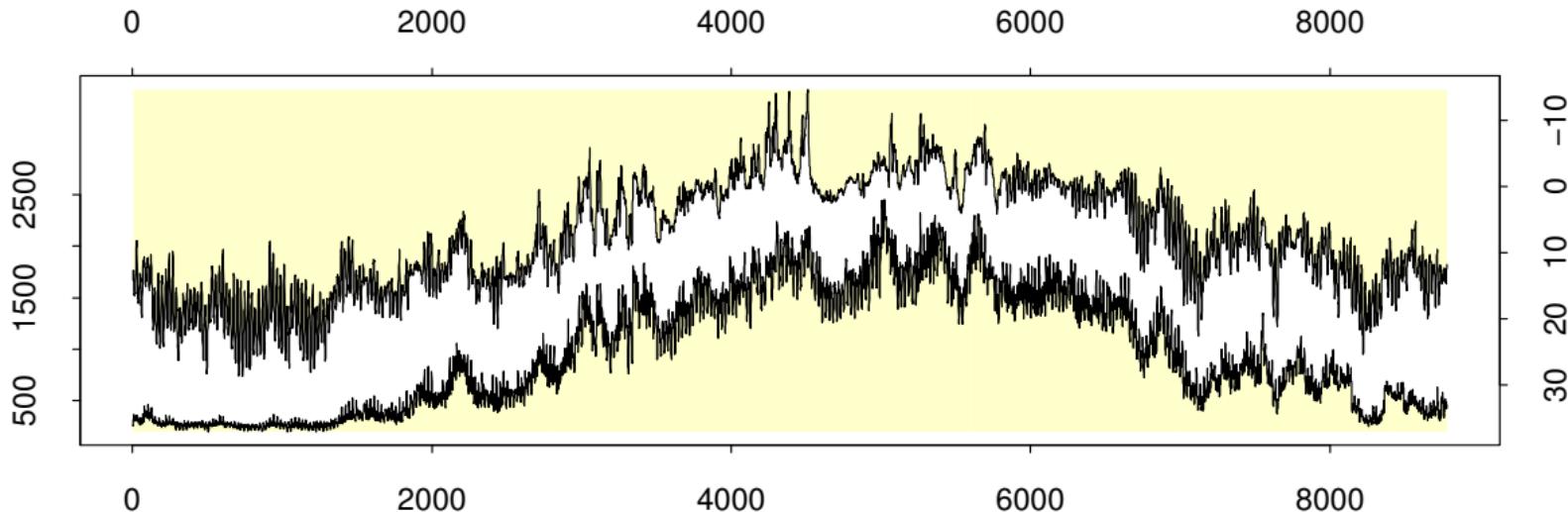
Recursive and adaptive estimation:

- ▶ Introduction to Chapter 11
- ▶ Recursive LS, Section 11.1
- ▶ Recursive pseudo-linear regression, Section 11.2
- ▶ Model-based adaptive estimation, Section 11.4

Further topics:

- ▶ Non-linear Time Series
- ▶ Stochastic differential equations
- ▶ Cointegration

Why recursive and adaptive estimation?



- ▶ As time passes we get more information
- ▶ New information should be included by “adjustment” rather than recalculating everything
- ▶ Models are approximations
- ▶ The best approximation may change over time
- ▶ Makes it possible to produce software which learns as new data becomes available

RLS – Types of models considered

REG:

$$Y_t = \mu + \beta_1 U_{1,t} + \beta_2 U_{2,t} + \dots + \beta_m U_{m,t} + \varepsilon_t$$

FIR:

$$Y_t = \mu + \omega(B)U_t + \varepsilon_t$$

$$= \mu + \omega_0 U_t + \omega_1 U_{t-1} + \dots + \omega_s U_{t-s} + \varepsilon_t$$

AR:

$$\phi(B)Y_t = \mu + \varepsilon_t \Leftrightarrow$$

$$Y_t = \mu - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \dots - \phi_p Y_{t-p} + \varepsilon_t$$

ARX:

$$\phi(B)Y_t = \mu + \omega(B)U_t + \varepsilon_t \Leftrightarrow$$

$$Y_t = \mu - \phi_1 Y_{t-1} - \dots - \phi_p Y_{t-p} + \omega_0 U_t + \dots + \omega_s U_{t-s} + \varepsilon_t$$

Generic form of the models considered

$$\begin{aligned}Y_t &= \mathbf{x}_t^T \boldsymbol{\theta} + \varepsilon_t \\&= \theta_1 x_{1,t} + \theta_2 x_{2,t} + \dots + \theta_\ell x_{\ell,t} + \varepsilon_t\end{aligned}$$

Example:

$$Y_t = \mu \cdot \underbrace{1}_{x_{1,t}} + \phi_2 \cdot \underbrace{(-Y_{t-2})}_{x_{2,t}} + \omega_1 \cdot \underbrace{U_{t-1}}_{x_{3,t}} + \varepsilon_t$$

LS-estimate at time t

Model:

$$Y_t = \mathbf{x}_t^T \boldsymbol{\theta} + \varepsilon_t$$

Data (\mathbf{x} may contain lagged values of the “real” input/output):

$$\begin{aligned} Y_1, Y_2, Y_3, Y_4, \dots, Y_{t-1}, Y_t \\ \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \dots, \mathbf{x}_{t-1}, \mathbf{x}_t \end{aligned}$$

LS-estimate based on t observations:

$$\begin{aligned} S_t(\boldsymbol{\theta}) &= \sum_{s=1}^t (Y_s - \mathbf{x}_s^T \boldsymbol{\theta})^2 \\ \hat{\boldsymbol{\theta}}_t &= \arg \min_{\boldsymbol{\theta}} S_t(\boldsymbol{\theta}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \end{aligned}$$

From one time step to the next (in an easy way)

The trick is to realize that:

$$R_t = \mathbf{X}^T \mathbf{X} = \mathbf{x}_1 \mathbf{x}_1^T + \mathbf{x}_2 \mathbf{x}_2^T + \dots + \mathbf{x}_t \mathbf{x}_t^T = \sum_{s=1}^t \mathbf{x}_s \mathbf{x}_s^T$$

$$\mathbf{h}_t = \mathbf{X}^T \mathbf{Y} = \mathbf{x}_1 Y_1 + \mathbf{x}_2 Y_2 + \dots + \mathbf{x}_t Y_t = \sum_{s=1}^t \mathbf{x}_s Y_s$$

Where:

$$\mathbf{x}_t \mathbf{x}_t^T = \begin{bmatrix} x_{1,t} x_{1,t} & x_{1,t} x_{2,t} & \cdots & x_{1,t} x_{\ell,t} \\ x_{2,t} x_{1,t} & x_{2,t} x_{2,t} & \cdots & x_{2,t} x_{\ell,t} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell,t} x_{1,t} & x_{\ell,t} x_{2,t} & \cdots & x_{\ell,t} x_{\ell,t} \end{bmatrix} \quad \mathbf{x}_t Y_t = \begin{bmatrix} x_{1,t} Y_t \\ x_{2,t} Y_t \\ \vdots \\ x_{\ell,t} Y_t \end{bmatrix}$$

The RLS algorithm

$$\hat{\theta}_t = \mathbf{R}_t^{-1} \mathbf{h}_t$$

$$\mathbf{R}_t = \sum_{s=1}^t \mathbf{x}_s \mathbf{x}_s^T = \mathbf{x}_t \mathbf{x}_t^T + \sum_{s=1}^{t-1} \mathbf{x}_s \mathbf{x}_s^T = \underline{\mathbf{x}_t \mathbf{x}_t^T + \mathbf{R}_{t-1}}$$

$$\mathbf{h}_t = \sum_{s=1}^t \mathbf{x}_s Y_s = \mathbf{x}_t Y_t + \sum_{s=1}^{t-1} \mathbf{x}_s Y_s = \underline{\mathbf{x}_t Y_t + \mathbf{h}_{t-1}}$$

Initialization:

- ▶ $\mathbf{R}_0 = \mathbf{0}$ (matrix of zeros)
- ▶ $\mathbf{h}_0 = \mathbf{0}$ (vector of zeros)
- ▶ Wait until $\hat{\theta}_t$ until \mathbf{R}_t is invertible

The RLS algorithm – 2 equivalent formulations

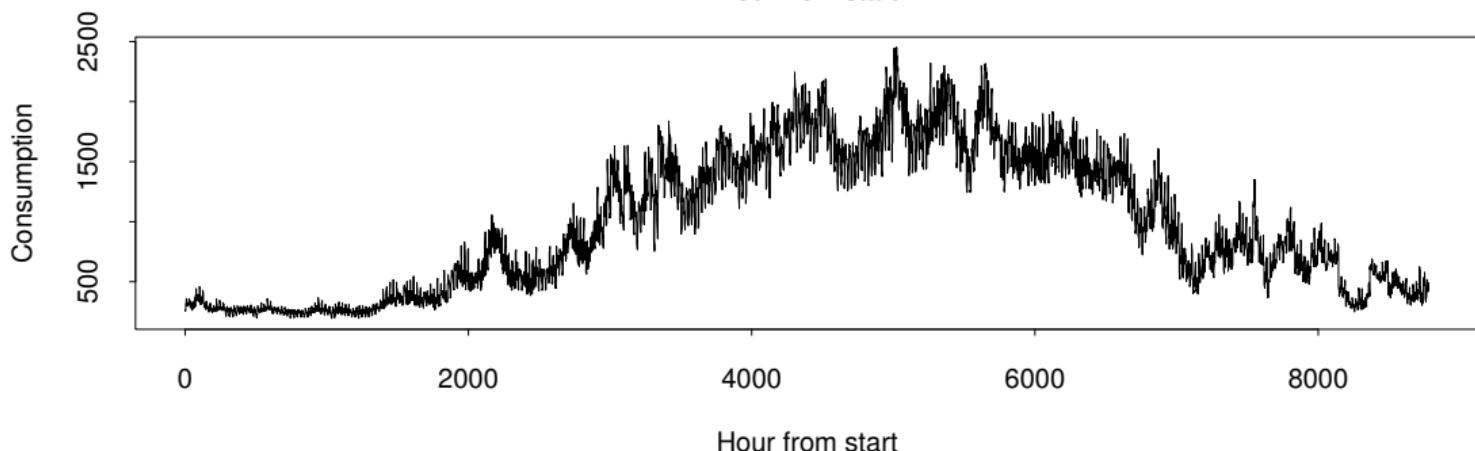
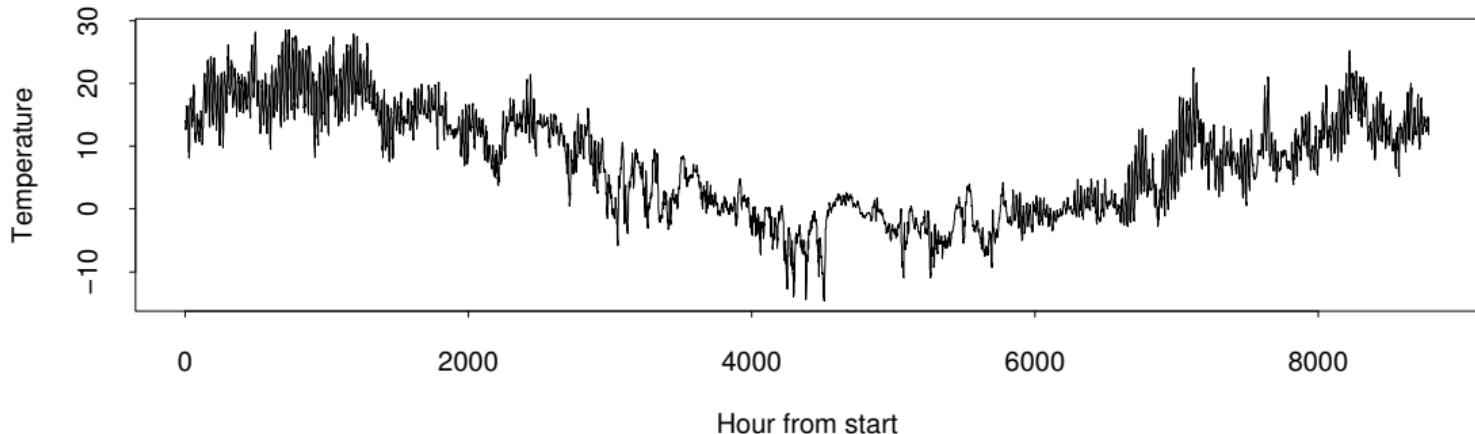
1. Eliminating h_t :

$$\begin{aligned}R_t &= \mathbf{x}_t \mathbf{x}_t^T + R_{t-1} \\ \hat{\boldsymbol{\theta}}_t &= \hat{\boldsymbol{\theta}}_{t-1} + R_t^{-1} \mathbf{x}_t (Y_t - \mathbf{x}_t^T \hat{\boldsymbol{\theta}}_{t-1})\end{aligned}$$

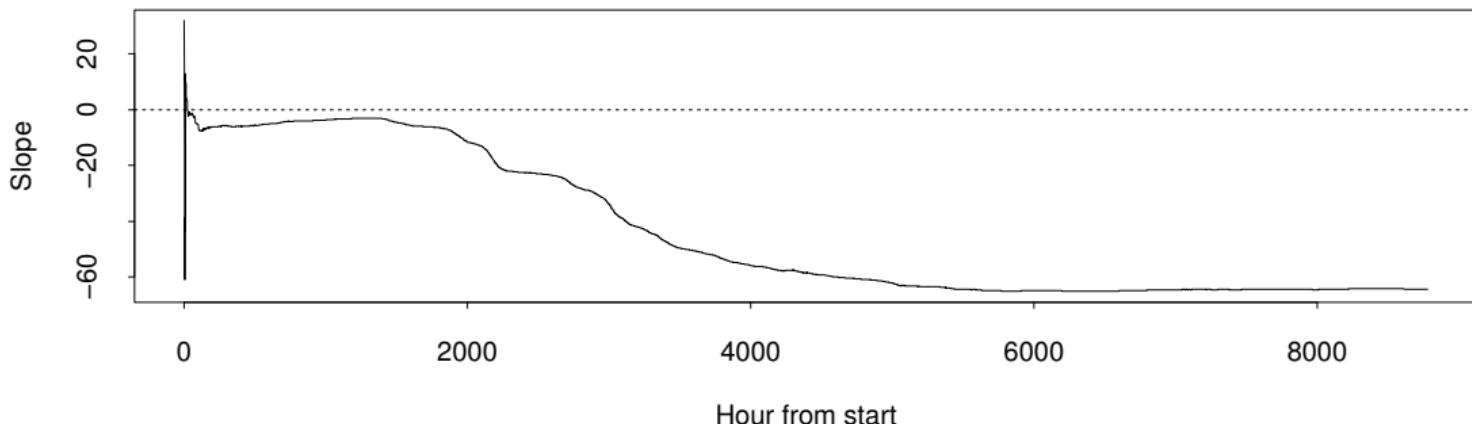
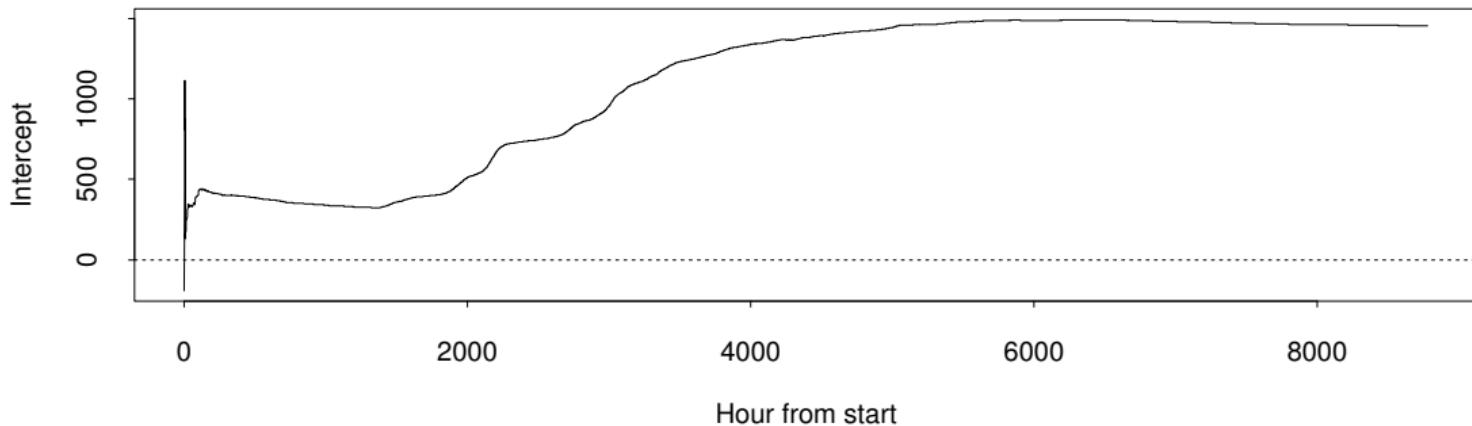
2. Eliminating h_t and avoiding matrix-inversion $P = R^{-1}$:

$$\begin{aligned}K_t &= \frac{P_{t-1} \mathbf{x}_t}{1 + \mathbf{x}_t^T P_{t-1} \mathbf{x}_t} \\ \hat{\boldsymbol{\theta}}_t &= \hat{\boldsymbol{\theta}}_{t-1} + K_t (Y_t - \mathbf{x}_t^T \hat{\boldsymbol{\theta}}_{t-1}) \\ P_t &= P_{t-1} - \frac{P_{t-1} \mathbf{x}_t \mathbf{x}_t^T P_{t-1}}{1 + \mathbf{x}_t^T P_{t-1} \mathbf{x}_t}\end{aligned}$$

Example: $HC_t = \mu + \theta_1 T_t + \varepsilon_t$



$$\text{Example: } HC_t = \mu + \theta_1 T_t + \varepsilon_t$$



Forgetting old observations

- ▶ So far we have a way of updating the estimates as the data set grows
- ▶ If we want a method which forgets old observations we apply weights which start at 1 and goes to 0 when observations gets old:

$$\hat{\boldsymbol{\theta}}_t = \arg \min_{\boldsymbol{\theta}} S_t(\boldsymbol{\theta}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$$

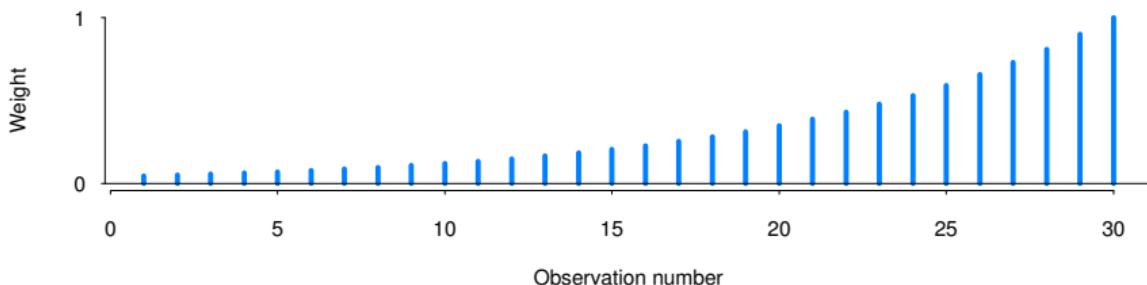
$$S_t(\boldsymbol{\theta}) = \sum_{s=1}^t \beta(t, s) (Y_s - \mathbf{x}_s^T \boldsymbol{\theta})^2$$

where $\mathbf{W} = \text{diag}(\beta(t, 1), \beta(t, 2), \dots, \beta(t, t-1), 1)$

- ▶ $\beta(t, s)$ express how we assign weights to old observations

Exponential decay of weights

- Let's first consider $\beta(t, s) = \lambda^{t-s}$ ($0 < \lambda \leq 1$)
 - $\lambda = 1$: What we did with the previous algorithms
 - $0 < \lambda < 1$: We "forget" in an exponential manner



- In the general case it turns out that if the sequence of weights can be written
 - $\beta(t, s) = \lambda(t)\beta(t-1, s)$ $1 \leq s \leq t-1$
 - $\beta(t, t) = 1$

Then the estimates can be updated recursively

The Adaptive Recursive LS algorithm

$$R_t = x_t x_t^T + \lambda(t) R_{t-1}$$

$$h_t = x_t Y_t + \lambda(t) h_{t-1}$$

$$\hat{\theta}_t = R_t^{-1} h_t$$

The Adaptive RLS algorithm – 2 equivalent formulations

1. Eliminating h_t :

$$\begin{aligned}R_t &= \mathbf{x}_t \mathbf{x}_t^T + \lambda(t) R_{t-1} \\ \hat{\boldsymbol{\theta}}_t &= \hat{\boldsymbol{\theta}}_{t-1} + R_t^{-1} \mathbf{x}_t (Y_t - \mathbf{x}_t^T \hat{\boldsymbol{\theta}}_{t-1})\end{aligned}$$

2. Eliminating h_t and avoiding matrix-inversion:

$$\begin{aligned}K_t &= \frac{P_{t-1} \mathbf{x}_t}{\lambda(t) + \mathbf{x}_t^T P_{t-1} \mathbf{x}_t} \\ \hat{\boldsymbol{\theta}}_t &= \hat{\boldsymbol{\theta}}_{t-1} + K_t (Y_t - \mathbf{x}_t^T \hat{\boldsymbol{\theta}}_{t-1}) \\ P_t &= \frac{1}{\lambda(t)} \left(P_{t-1} - \frac{P_{t-1} \mathbf{x}_t \mathbf{x}_t^T P_{t-1}}{\lambda(t) + \mathbf{x}_t^T P_{t-1} \mathbf{x}_t} \right)\end{aligned}$$

Constant forgetting

- ▶ If $\lambda(t) = \lambda$ we call λ the *forgetting factor* and define the *memory* as

$$T_0 = \sum_{i=0}^{\infty} \lambda^i = 1 + \lambda + \lambda^2 + \lambda^3 + \lambda^4 + \dots = \frac{1}{1 - \lambda}$$

- ▶ Given a data set an optimal value of λ can be found by “trial and error”
- ▶ It is often a good idea to select the values of λ to be investigated so that the corresponding values of T_0 are approximately equidistant
- ▶ The criteria to evaluate may depend on the application, but the sum of squared one-step prediction errors is often appropriate
- ▶ An initialization period should be excluded from the evaluation

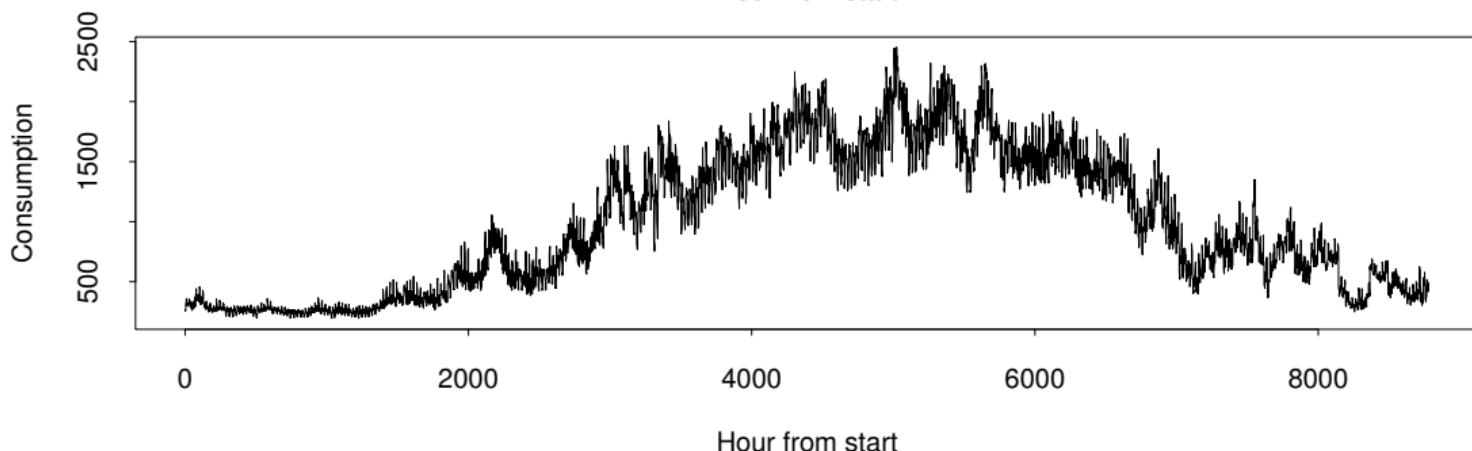
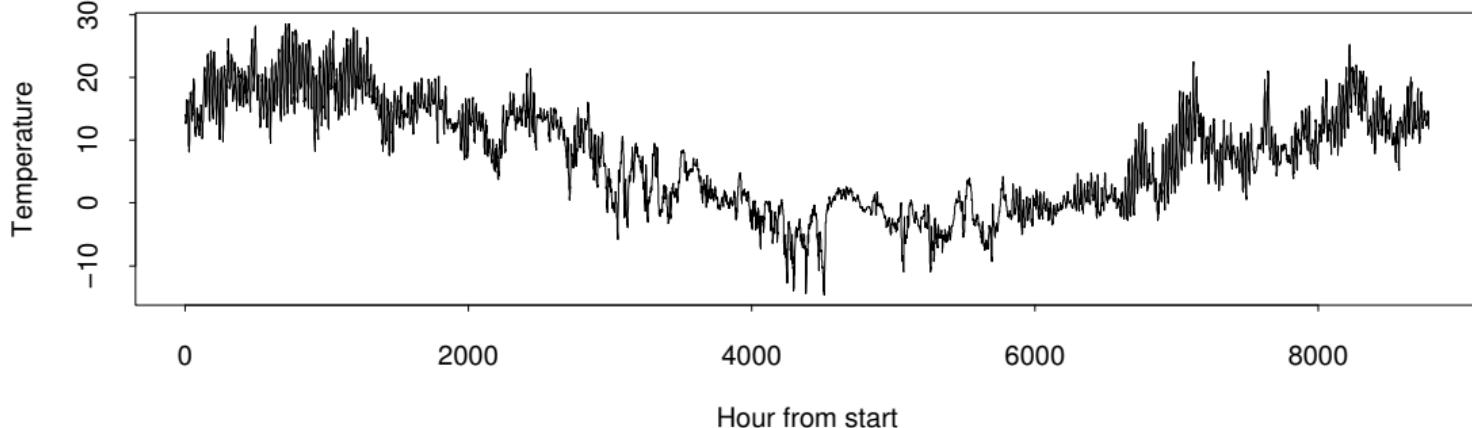
Variable forgetting

- ▶ Many methods exists
- ▶ One is based on the aim of keeping the criteria functions defining the estimates constant at S_0
- ▶ Leads to:

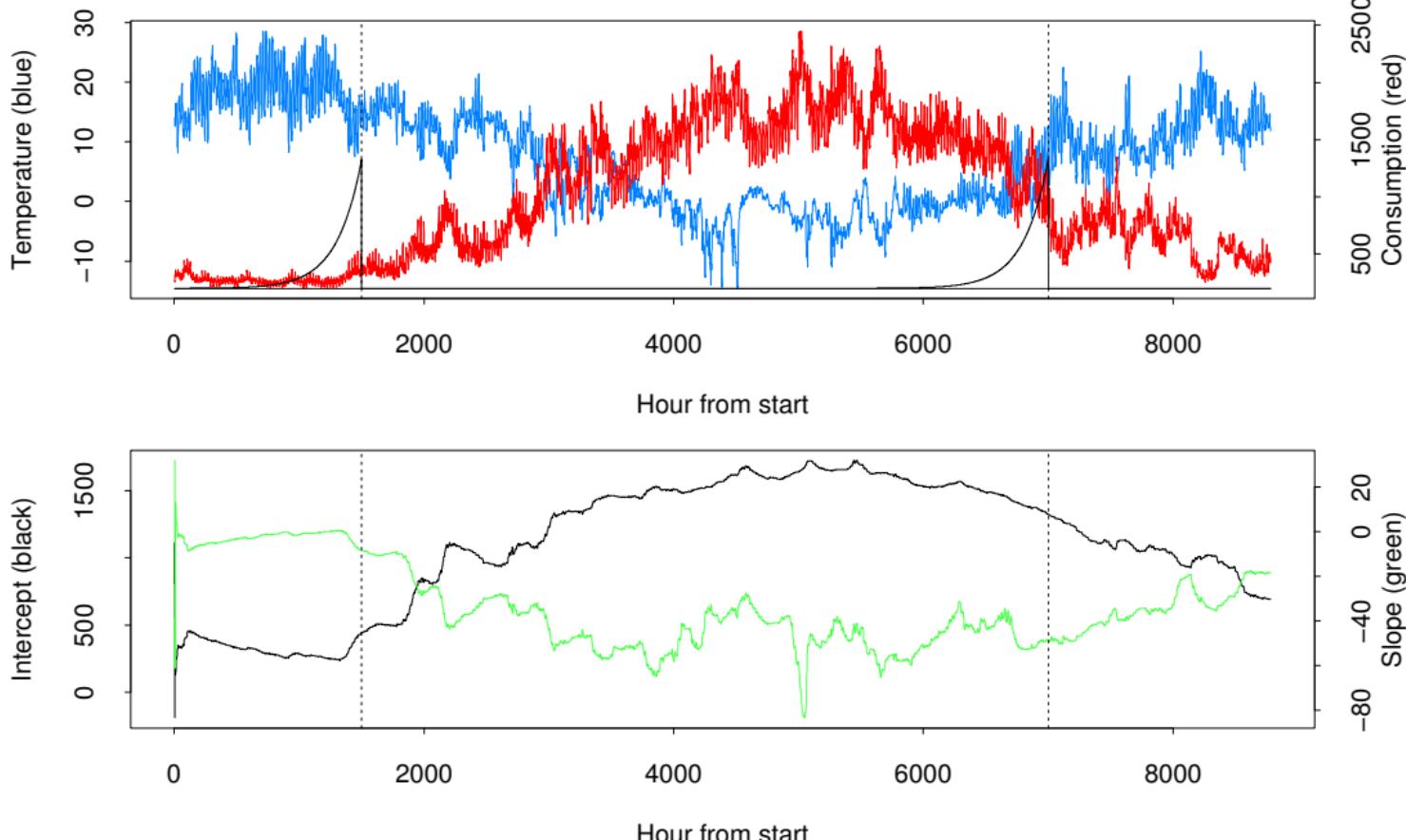
$$\lambda(t) \simeq 1 - \frac{\varepsilon_t^2}{S_0 [1 + \mathbf{x}_t^T \mathbf{P}_{t-1} \mathbf{x}_t]}$$

- ▶ A lower bound λ_{\min} on $\lambda(t)$ should be applied
- ▶ For optimal tuning of this method S_0 could be varied

Example: $HC_t = \mu + \theta_1 T_t + \varepsilon_t$



Example: $HC_t = \mu + \theta_1 T_t + \varepsilon_t$, $\lambda = 0.995$



Recursive pseudo-linear regression

- ▶ Problem: The ARMA structure cannot be estimated using regression directly.
- ▶ However, given the parameters, θ , the one-step prediction residuals can be calculated and used for regression.
- ▶ The model becomes

$$\hat{Y}_{t|t-1}(\theta) = \mathbf{X}_t^T(\theta)\boldsymbol{\theta}$$

- ▶ We minimize

$$S_t(\boldsymbol{\theta}) = \lambda(t)S_{t-1}(\boldsymbol{\theta}) + (Y_t - \mathbf{X}_t^T(\boldsymbol{\theta})\boldsymbol{\theta})^2$$

with respect to $\boldsymbol{\theta}$.

- ▶ Then, the RPLR algorithm is:

$$\mathbf{R}_t = \mathbf{x}_t \mathbf{x}_t^T + \lambda(t) \mathbf{R}_{t-1}$$

$$\mathbf{h}_t = \mathbf{x}_t Y_t + \lambda(t) \mathbf{h}_{t-1}$$

$$\hat{\boldsymbol{\theta}}_t = \mathbf{R}_t^{-1} \mathbf{h}_t$$

- ▶ I.e. as before except that \mathbf{x}_t must be calculated at each step.

Model-based adaptive estimation

- ▶ What is this?

$$\begin{aligned}\boldsymbol{X}_{t+1} &= \boldsymbol{X}_t + \boldsymbol{e}_{1,t}, & V(\boldsymbol{e}_{1,t}) &= \boldsymbol{\Sigma}_1 \\ Y_t &= \boldsymbol{C}_t \boldsymbol{X}_t + \boldsymbol{e}_{2,t}, & V(\boldsymbol{e}_{2,t}) &= \boldsymbol{\Sigma}_2\end{aligned}$$

- ▶ How do we predict and reconstruct such a system?
- ▶ Now the parameters are the latent state

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \boldsymbol{e}_{1,t}, & V(\boldsymbol{e}_{1,t}) &= \boldsymbol{\Sigma}_1 \\ Y_t &= \boldsymbol{X}_t^T \boldsymbol{\theta}_t + \boldsymbol{e}_{2,t}, & V(\boldsymbol{e}_{2,t}) &= \boldsymbol{\Sigma}_2\end{aligned}$$

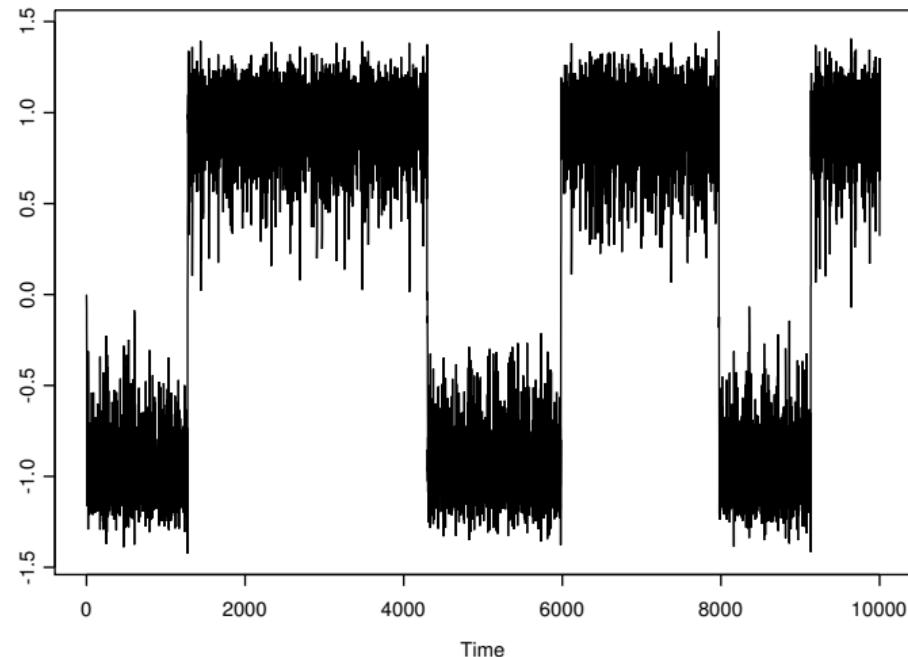
- ▶ That means we can use the Kalman filter for tracking the parameters.
- ▶ See the formulation of the Kalman filter in the book.

Models with time-varying parameters

- ▶ ARMA processes where the parameters are time-varying.
- ▶ The parameters can follow deterministic functions of time or be stochastic processes.
- ▶ When the parameters are stochastic processes, the models are called *double stochastic*.
 - ▶ This could be an ARMA structure where the parameters are other ARMA processes.
- ▶ The Kalman filter is the central tool in estimation on such processes.

Non-linear time series

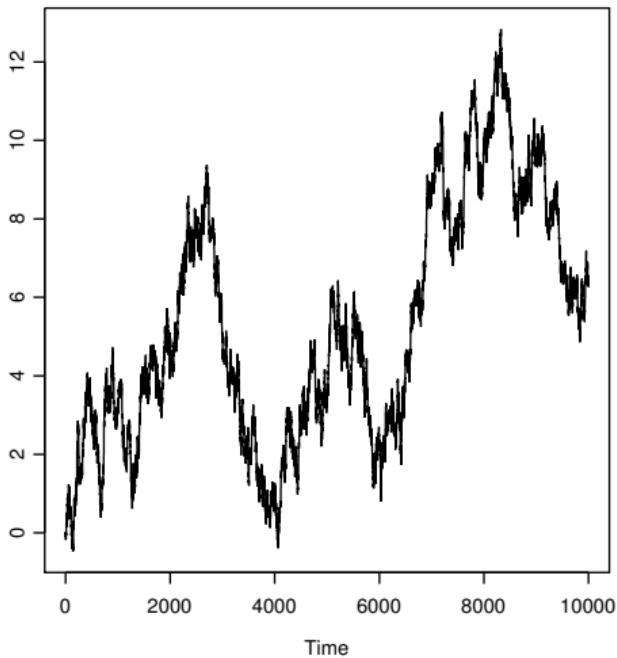
There are of course many, many possible formulations.



$$dX_t = -X_t(X_t + 1)(X_t - 1)dt + \sigma dW_t,$$

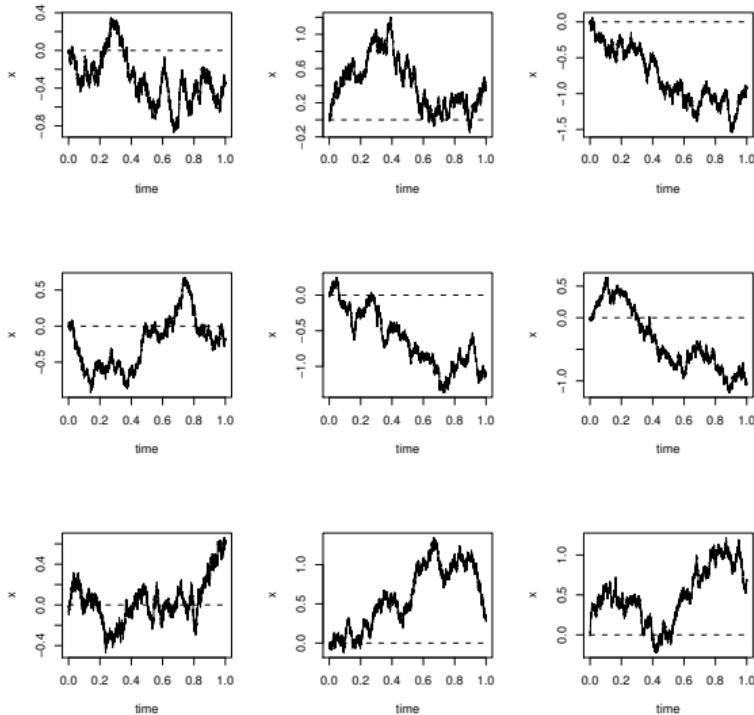
where $\sigma = 0.11$, and W is a Brownian Motion.

Brownian Motion



The continuous-time counterpart to a random walk

9 Brownian Motions



Properties of a Brownian Motion W

- ▶ The stochastic process W is continuous with probability 1;
- ▶ The stochastic process W is nowhere differentiable with probability 1;
- ▶ If you wish to straighten out the graph over any interval, the length will be infinite due to infinite variation.

Stochastic Differential Equations

The equation

$$dX_t = -X_t(X_t + 1)(X_t - 1)dt + \sigma dW_t$$

is a Stochastic Differential Equation (SDE).

More general SDE's:

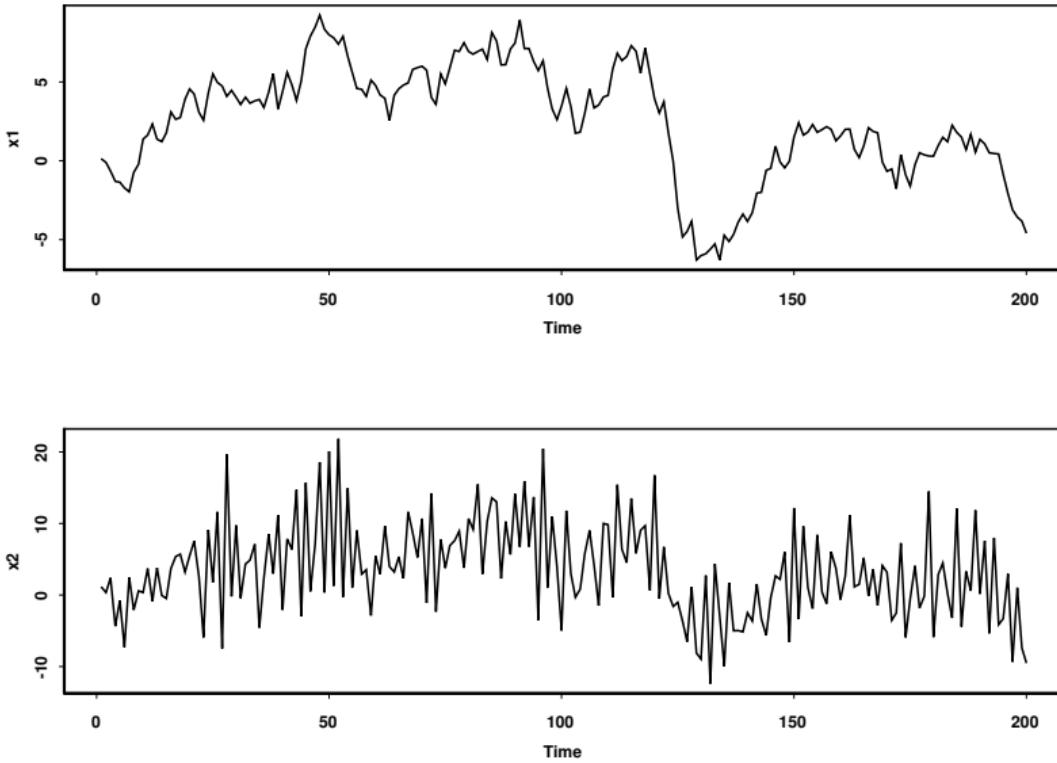
$$dX_t = a(X, t)dt + b(X, t)dW_t$$

The equation is really an integral equation:

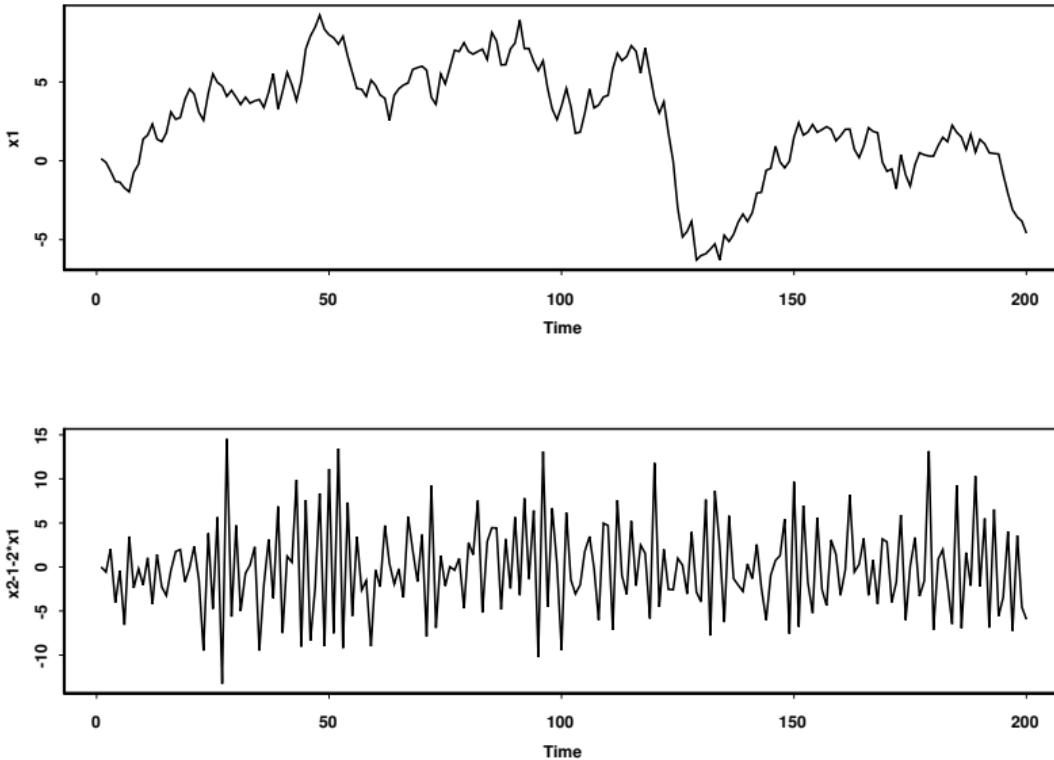
$$X(t) = \int_0^t a(X, s) ds + \int_0^t b(X, s) dW(s)$$

The last term is a so-called *stochastic integral* (advanced topic).

Cointegration 1



Cointegration 2



Same Time Series in different coordinates

Cointegration links

- ▶ Cointegration in economics:
<http://isc.temple.edu/economics/notes/cointegration/cointegration.HTM>
- ▶ Clive Grangers lecture:
http://www.nobelprize.org/nobel_prizes/economics/laureates/2003/granger-lecture.pdf
- ▶ Justification from the Swedish Academy of Sciences (much like link 1):
http://www.nobelprize.org/nobel_prizes/economics/laureates/2003/advanced-economicsciences2003.pdf

More links:

- ▶ Spatial time series (Chlorophyll dynamics):
http://spg.ucsd.edu/Satellite_Projects/Chlorophyll_dynamics_Santa_Barbara_Basin/Chlorophyll_dynamics_Santa_Barbara_Basin.htm
- ▶ Sunspot research at NASA:
http://science.nasa.gov/science-news/science-at-nasa/2008/11jul_solarcycleupdate/
- ▶ Time Series Data Library by Rob J Hyndman:
<http://datamarket.com/data/list/?q=provider:tsdl>

Related courses

- ▶ 02427 Advanced time series analysis
- ▶ 02425 Diffusions and stochastic differential equations
- ▶ 02407 Stochastic processes
- ▶ Many more ;-)

Highlights

- ▶ Recursive LS for many types of models.
- ▶ Adaptive Recursive LS
- ▶ Kalman filtering to trace parameters
- ▶ Where next:
 - ▶ Continuous time
 - ▶ Non-linearities
 - ▶ Stochastic Differential Equations

Merry Christmas!