Advanced Data Analysis Methods

Henrik Madsen
www.smart-cities-centre.org

Lawrence Berkeley National Lab

hmad.dtu@gmail.com

Outline

Introduction

- Various methods of advanced modelling are needed for an increasing number of complex physical, chemical and biological systems.
- For a model to describe the future evolution of the system, it must
  1. capture the inherently non-linear behavior of the system.
  2. provide means to accommodate for noise due to approximation, input, and measurement errors.
- Calls for methods that are capable of **bridging the gap between physical and statistical modelling**.

Which type of model to use?

- Simple / Complex
- Lumped / Distributed
- Linear / Non-linear
- Time-invariant / Time-varying
- Discrete / Continuous time
- Deterministic / Stochastic
- Black box / Grey box / White box
- Parametric / Non-parametric

Base the decision on

- Purpose
- Prior knowledge
- Available data
- Tools

Nonlinear versus linear modelling

- The aim of the modelling effort may be generally expressed as follows: Find a **nonlinear** function h such that $\{\epsilon_t\}$ defined by

$$h(X_t, X_{t-1}, \dots) = \epsilon_t \tag{1}$$

  is a sequence of independent random variables.
- Suppose also that the model is *causally invertible*, i.e. the equation above may be 'solved' such that we may write

$$X_t = h'(\epsilon_t, \epsilon_{t-1}, \dots). \tag{2}$$

Nonlinear vs. linear model buiding (cont.)

- Suppose that h' is sufficiently well-behaved to be expanded in a Taylor series

$$\begin{aligned} X_t &= \mu + \sum_{k=0}^{\infty} g_k \epsilon_{t-k} + \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} g_{kl}\epsilon_{t-k}\epsilon_{t-l} \\ &\quad + \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty} g_{klm}\epsilon_{t-k}\epsilon_{t-l}\epsilon_{t-m} + \dots \end{aligned} \tag{3}$$

- The functions

$$\mu = h'(0),\ g_k = \left(\frac{\partial h'}{\partial \epsilon_{t-k}}\right),\ g_{kl} = \left(\frac{\partial^2 h'}{\partial \epsilon_{t-k}\partial \epsilon_{t-l}}\right),\ \text{etc.} \tag{4}$$

  are called the *Volterra series* for the process $\{X\}$. The sequences $g_k, g_{kl}, \dots$ are called the *kernels* of the Volterra series.

Non-linear vs. linear model building (cont.)

- For **linear systems**

$$g_{kl} = g_{klm} = g_{klmn} = \ldots = 0 \tag{5}$$

- Hence the system is completely characterized by either

$$\{g_k\} \quad : \quad \text{Impulse response function}$$

or

$$\mathcal{H}(\rightarrow) \quad : \quad \text{Frequency response function}$$

Non-linear vs. linear model building (cont.)

- In general there is *no such thing as a transfer function* for non-linear systems.
- However, an *infinite sequence of generalized transfer functions* may be defined as

$$
\begin{aligned}
H_1(\omega_1) &= \sum_{k=0}^{\infty} g_k e^{-i\omega_1 k} \\
H_2(\omega_1, \omega_2) &= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} g_{kl} e^{-i(\omega_1 k+\omega_2 l)} \\
H_3(\omega_1, \omega_2, \omega_3) &= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty} g_{klm} e^{-i(\omega_1 k+\omega_2 l+\omega_3 m)} \\
&\vdots
\end{aligned}
$$

Non-linear vs. linear model building (cont.)

Let U_t and X_t denote the input and the output of a given system.

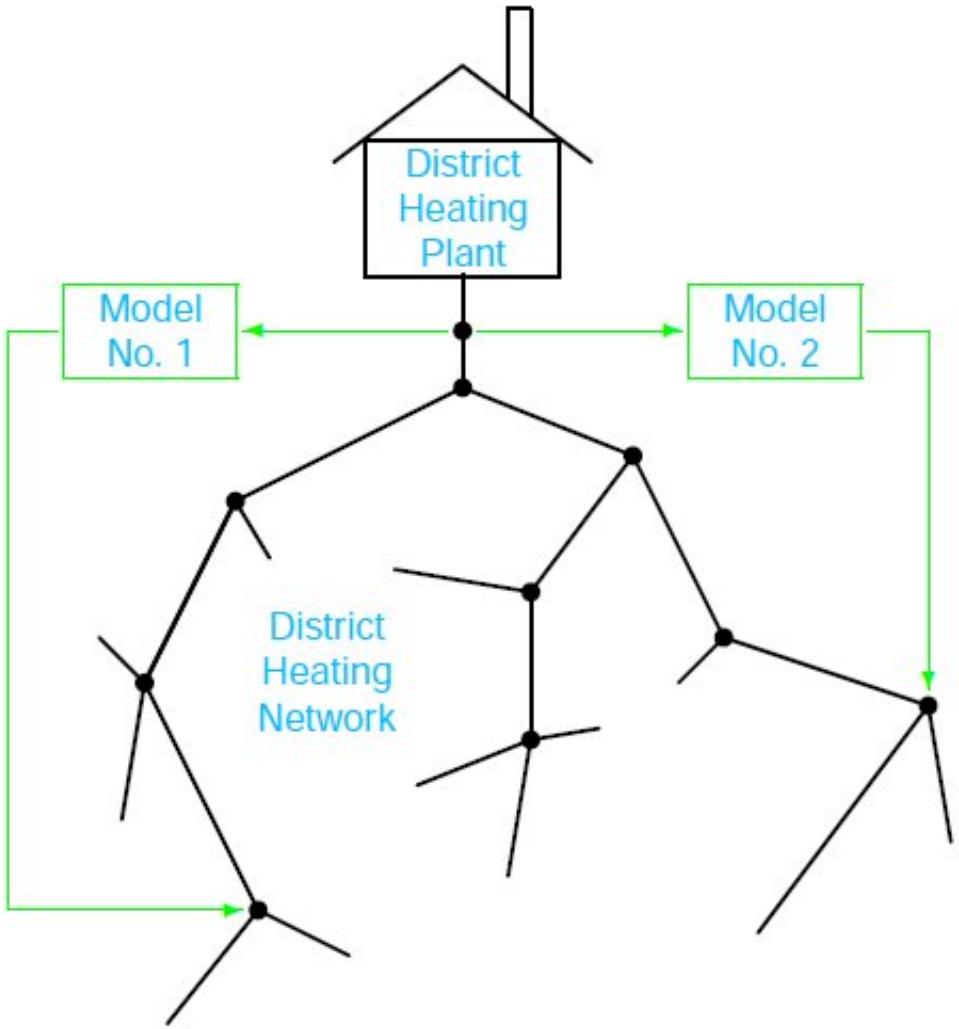
- For **linear** systems it is well known that
  - L1 If the input is a single harmonic $U_t = A_0 e^{i\omega_0 t}$ then the output is a single harmonic of *the same frequency*, but with the amplitude scaled by $|H(\omega_0)|$ and the phase shifted by $\arg H(\omega_0)$.
  - L2 Due to the linearity, the *principle of superposition* is valid, and the total output is the sum of the outputs corresponding to the individual frequency components of the input. (Hence the system is completely described by knowing the response to all frequencies – that is what the transfer function supplies).
- For **non-linear** systems, however, neither of the properties (L1) or (L2) hold.
  - NL1 For an input with frequency ω_0, the output will, in general, contain also components at the frequencies $2\omega_0, 3\omega_0, \ldots$ (*frequency multiplication*).
  - NL2 For two inputs with frequencies ω_0 and ω_1, the output will contain components at frequencies $\omega_0, \omega_1, (\omega_0 + \omega_1)$ and all harmonics of the frequencies (*intermodulation distortion*).

Conditional parametric ARX-model

$$y_t = \sum_{i \in L_y} a_i(x_{t-m}) y_{t-i} + \sum_{i \in L_u} b_i(x_{t-m}) u_{t-i} + e_t, \tag{6}$$

- The **functions** $a_i(x_{t-m})$ and $b_i(x_{t-m})$ must be estimated
- The model may be written as $y_t = \boldsymbol{z}_t^T \theta(\boldsymbol{x}_t) + e_t$

DH network principle



Application

- One consumer consisting of 84 households
- Measurements:
  - Flow at plant
  - Temperature at plant
  - Temperature at consumer

Application - Models

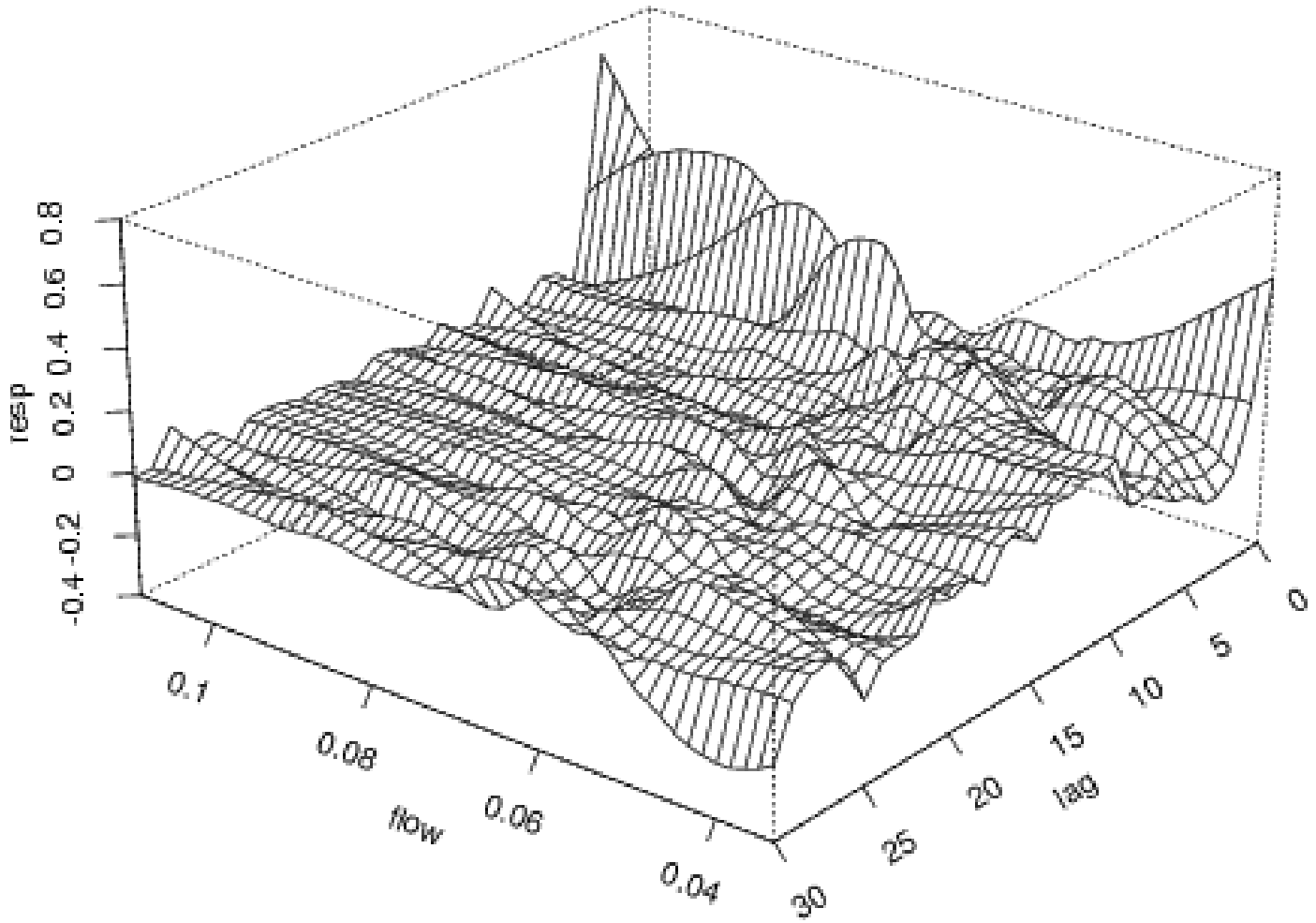
- FIR-model

$$y_t = \sum_{i=0}^{30} b_i(x_t) u_{t-i} + e_t$$

- ARX-model

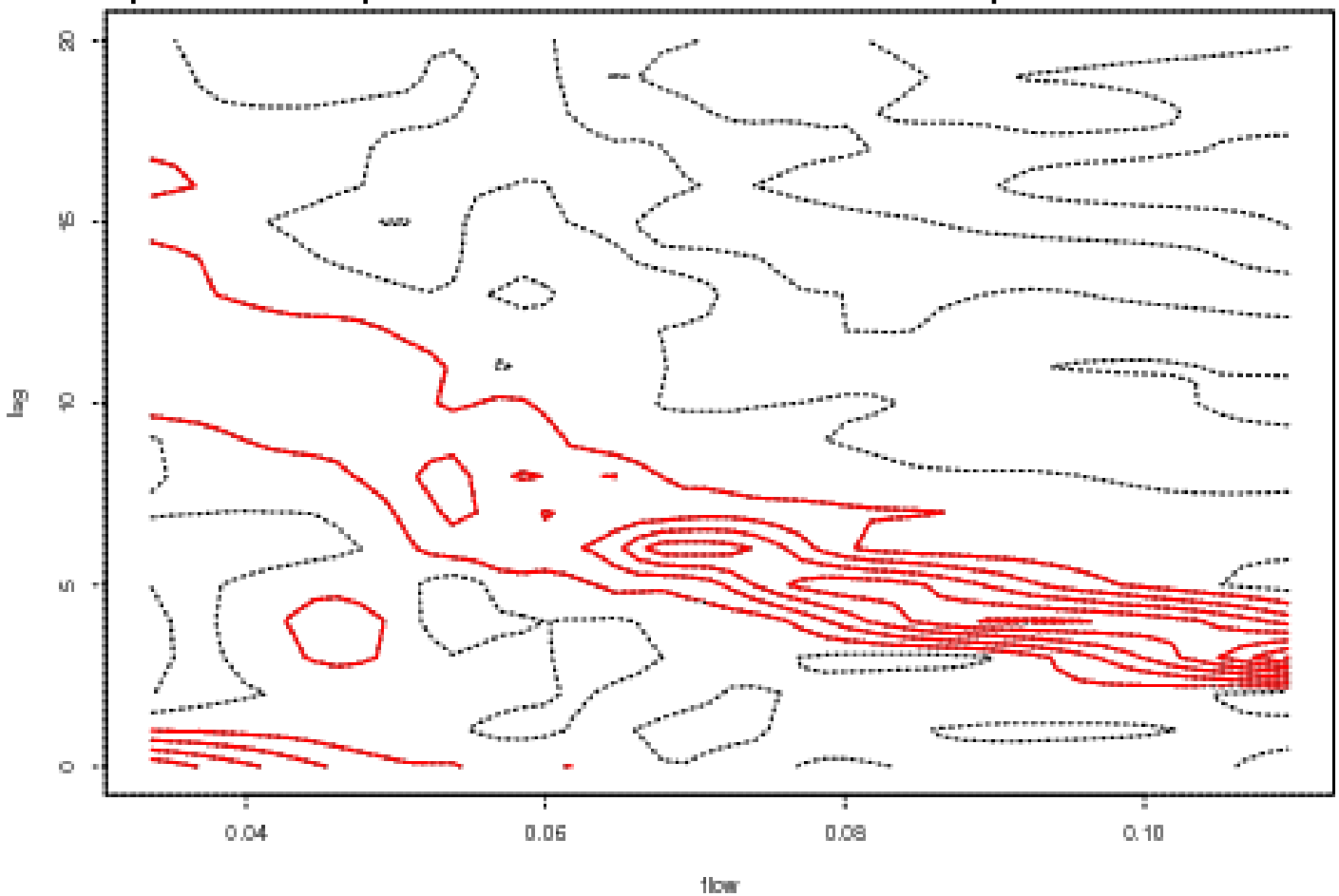
$$y_t = a(x_t) y_{t-1} + \sum_{i=3}^{15} b_i(x_t) u_{t-i} + e_t$$

Impulse Response of FIR-model (40%)



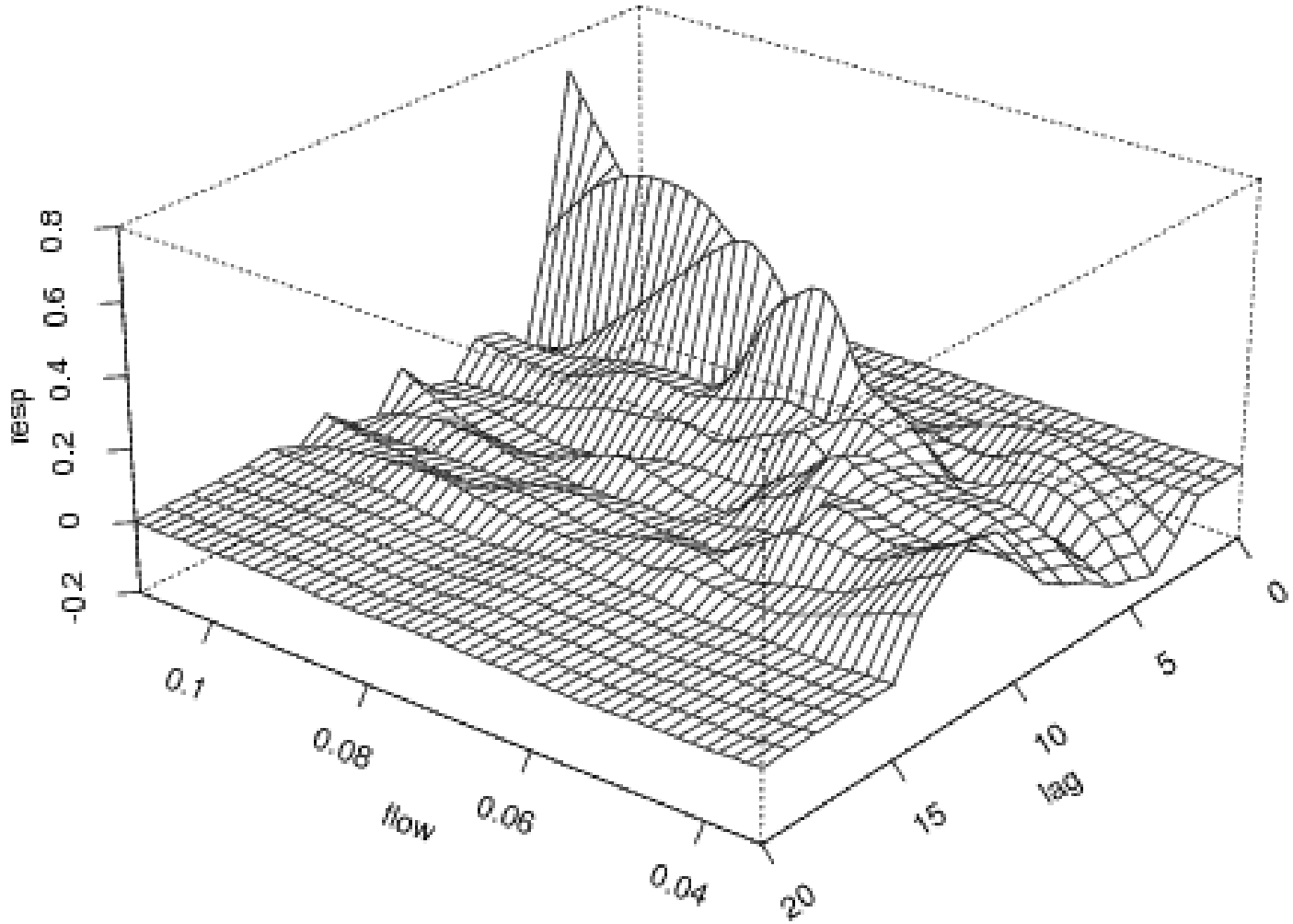
Impulse Response - FIR model

Impulse Response: -0.1 to 0.7 ^oC in steps of 0.1 ^oC



— Positive

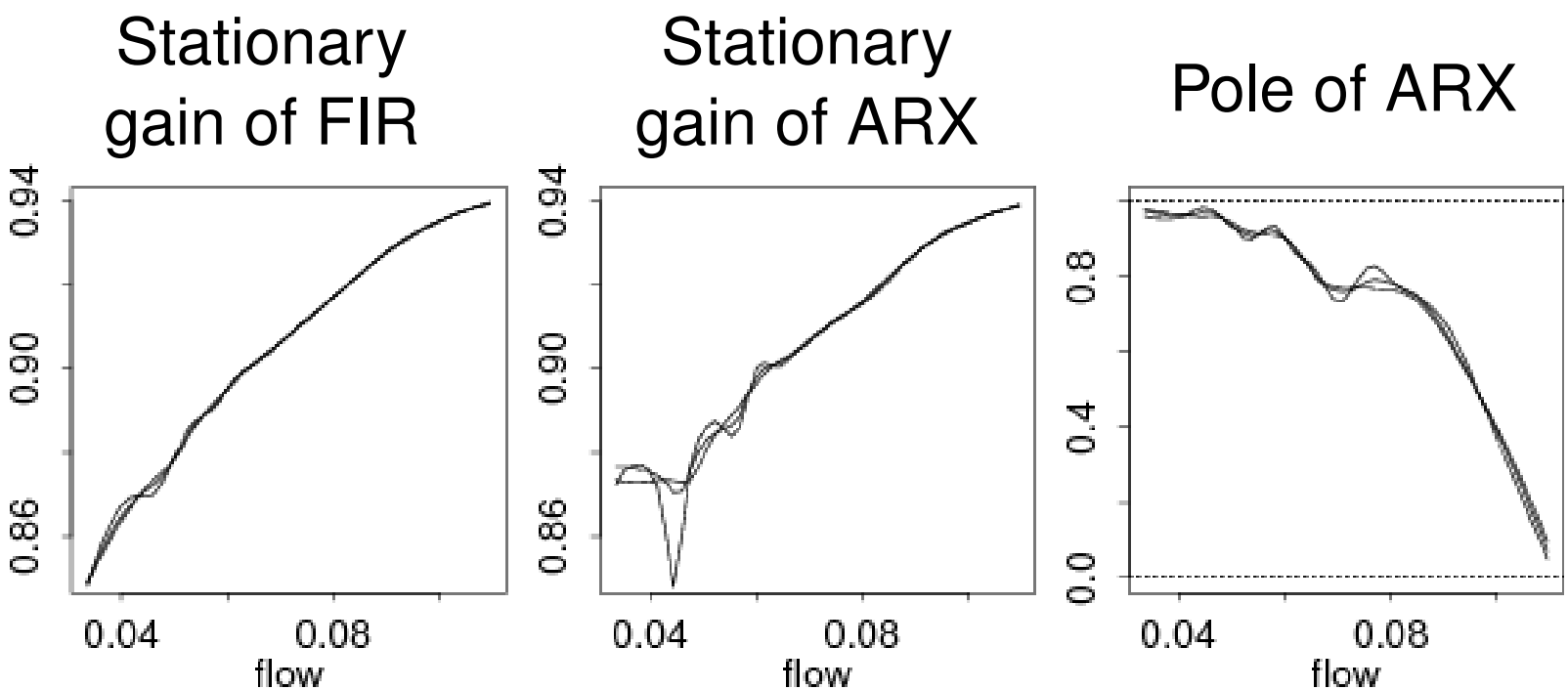
Impulse Response of ARX-model (40%)



Characteristics

Characteristics

30%, 40%, 50%



Conclusions - Transfer function

- Time delay decreasing with increasing flow
- 6-15% temperature loss depending on flow
- Possible inaccuracy of the model at low flows (input design?)

Model Predictive Controller for DH-systems

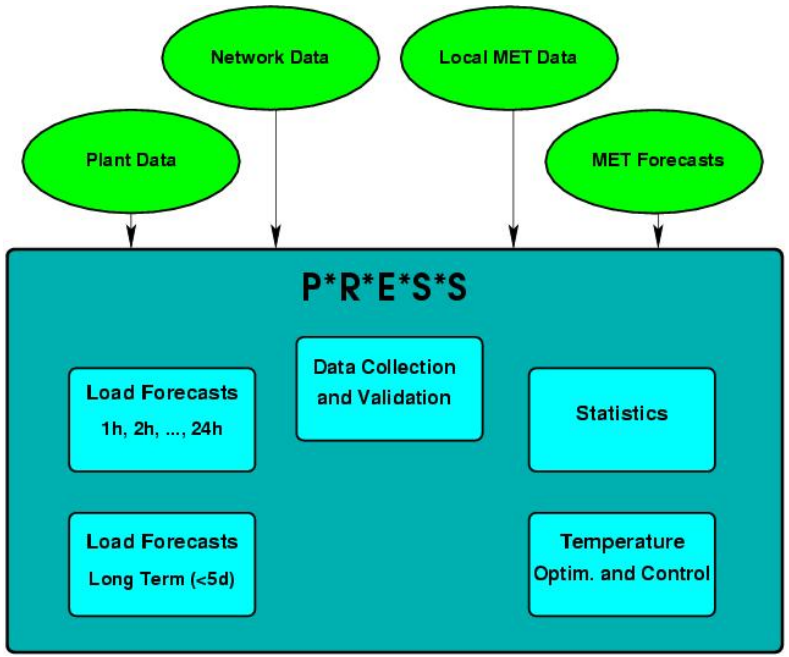
The netpoint temperature control is implemented using the XGPC controller:

$$\begin{aligned} \min_{\boldsymbol{u}_t} J(\boldsymbol{\Gamma}_t, \boldsymbol{\Lambda}_t, \boldsymbol{\Omega}_t; t, \boldsymbol{u}_t) &= E[(\boldsymbol{y}_t - \boldsymbol{y}_t^0)^T \boldsymbol{\Gamma}_t (\boldsymbol{y}_t - \boldsymbol{y}_t^0) + \boldsymbol{u}_t^T \boldsymbol{\Lambda}_t \boldsymbol{u}_t + 2\boldsymbol{\omega}_t^T \boldsymbol{u}_t] \\ \boldsymbol{y}_t &= \boldsymbol{H}_t \boldsymbol{u}_t + \boldsymbol{v}_t + \boldsymbol{e}_t \end{aligned}$$

The flow rate control is implemented using the relation $p_t = c_w q_t (T_t^s - T_t^r)$. The supply temperature is found as

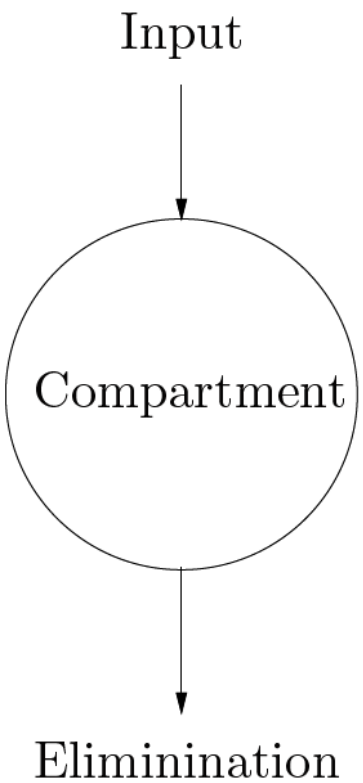
$$T_{t+1}^s = \sum_{i=1}^{N_u} w_i \left[\hat{T}_{t+i|t}^r + \frac{\hat{p}_{t+i|t}}{c_w \; q^0} \right] . \tag{7}$$

Implementation of Models and Controllers



Installed at about 15 DH plants in Denmark. Observed savings are 10 - 20 pct of the heat loss, and the pay-back time is from 5 month to 2 years.

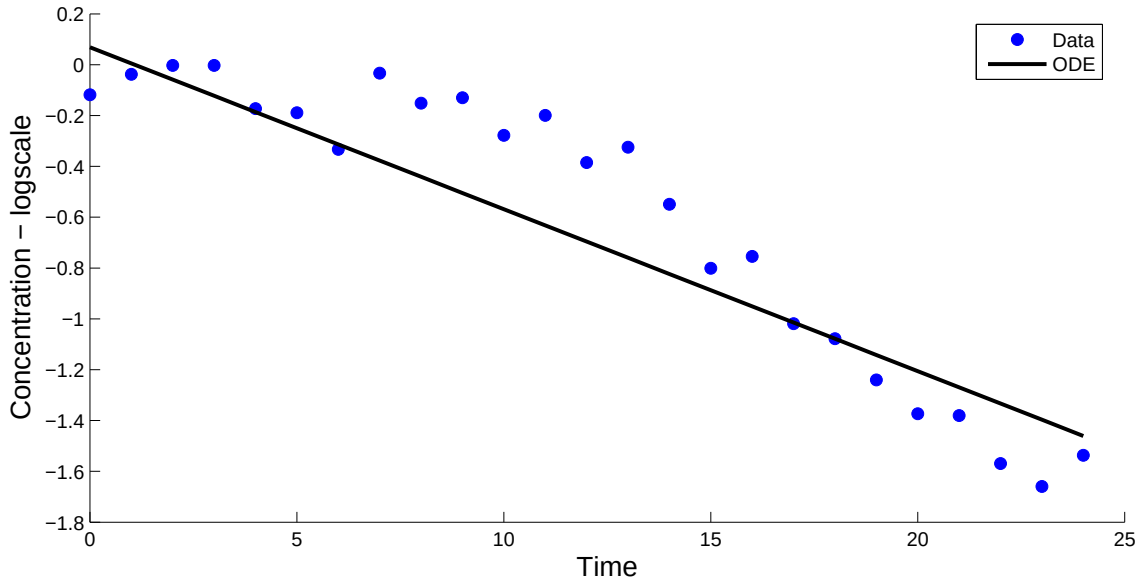
Why Stochastic Differential Equations?
Problem Scenario



Ordinary differential equation

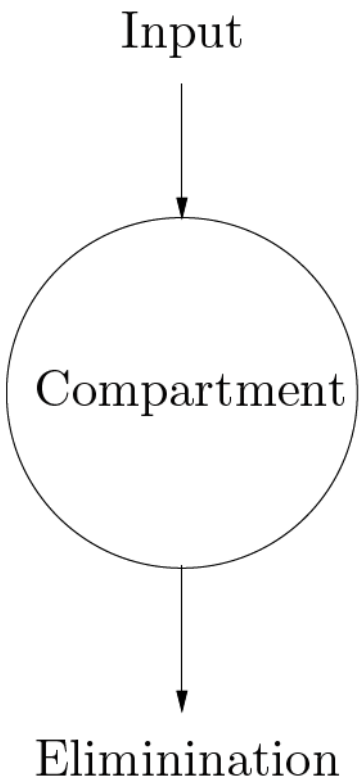
$$dA = -KA\,dt$$
$$Y = A + \epsilon$$

ODE



- Autocorrelated residuals!!

Problem Scenario

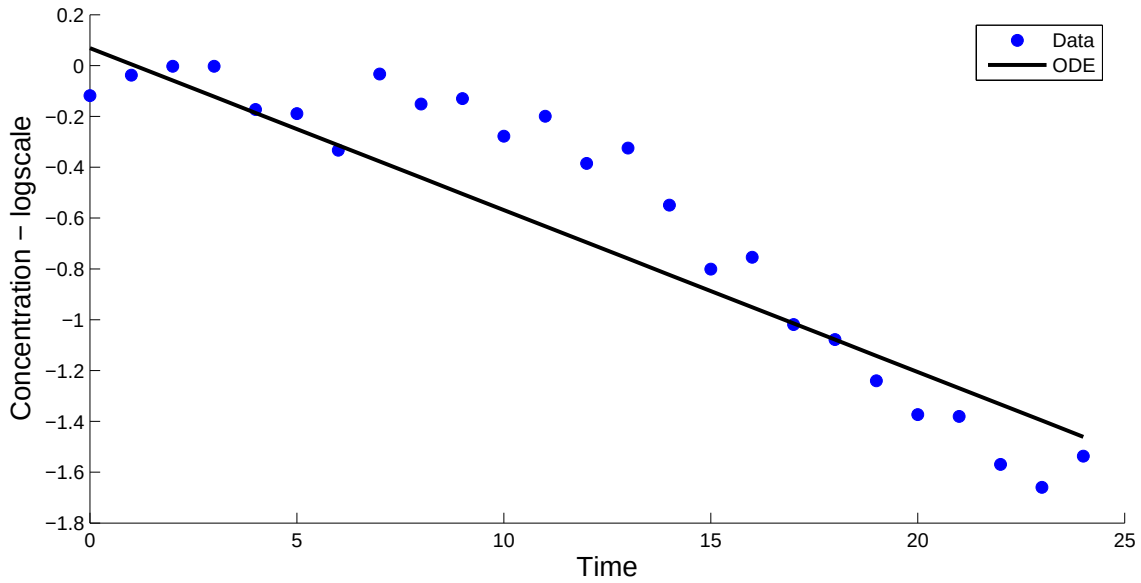


Stochastic differential equation

$$dA = -KA\,dt + dw$$

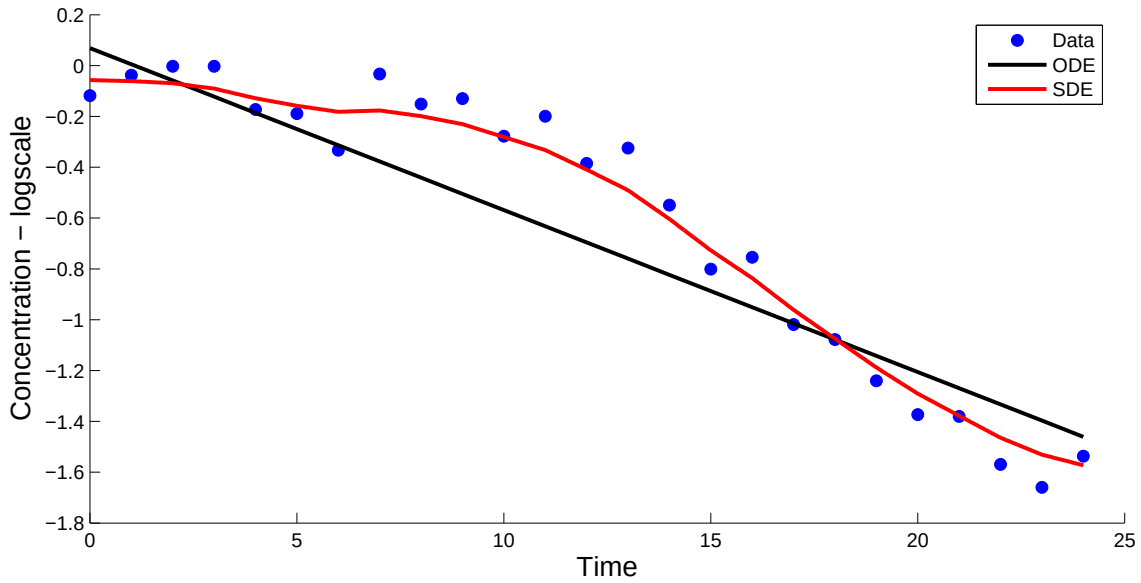
$$Y = A + e$$

ODE vs SDE



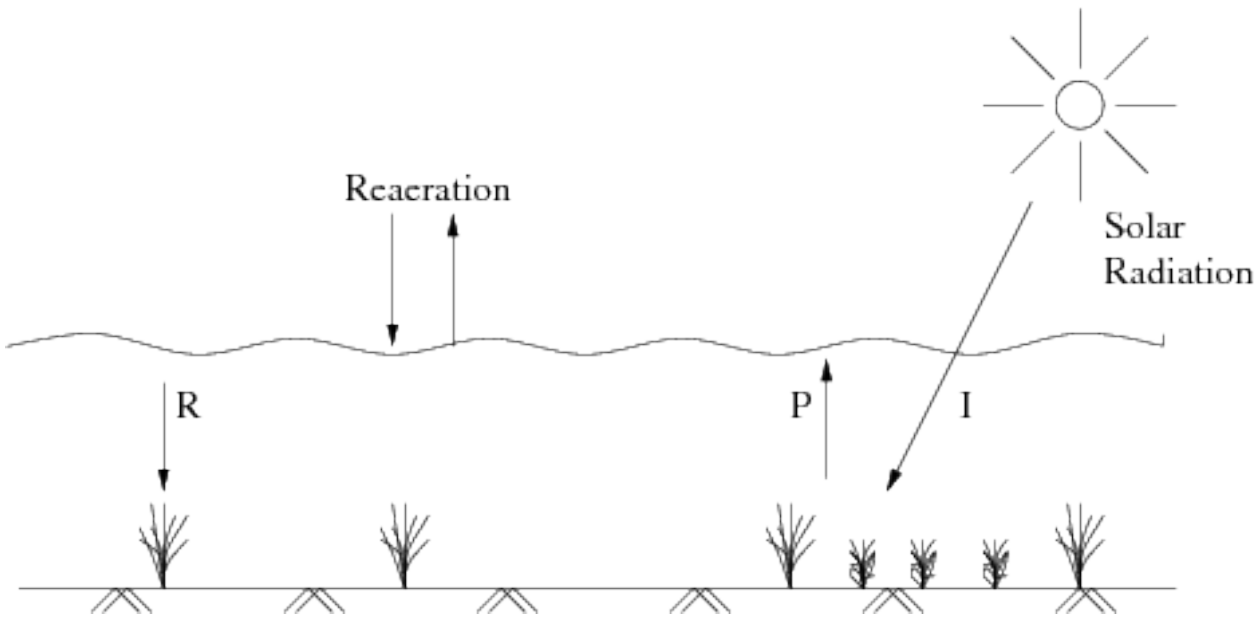
- Uncorrelated residuals
- System noise
- Measurement noise

ODE vs SDE



- Uncorrelated residuals
- System noise
- Measurement noise

Grey box modelling of oxygen concentration - A sketch of the physical system



Grey box modelling of oxygen concentration - A white box model

Model found in the literature:

$$\frac{dC}{dt} = \frac{K}{h\sqrt{h}}(C_m(T) - C) + P(I) - R(T)$$

$$P(I) = P_m E_0 \frac{I}{P_m + E_0 I} (= \beta I)$$

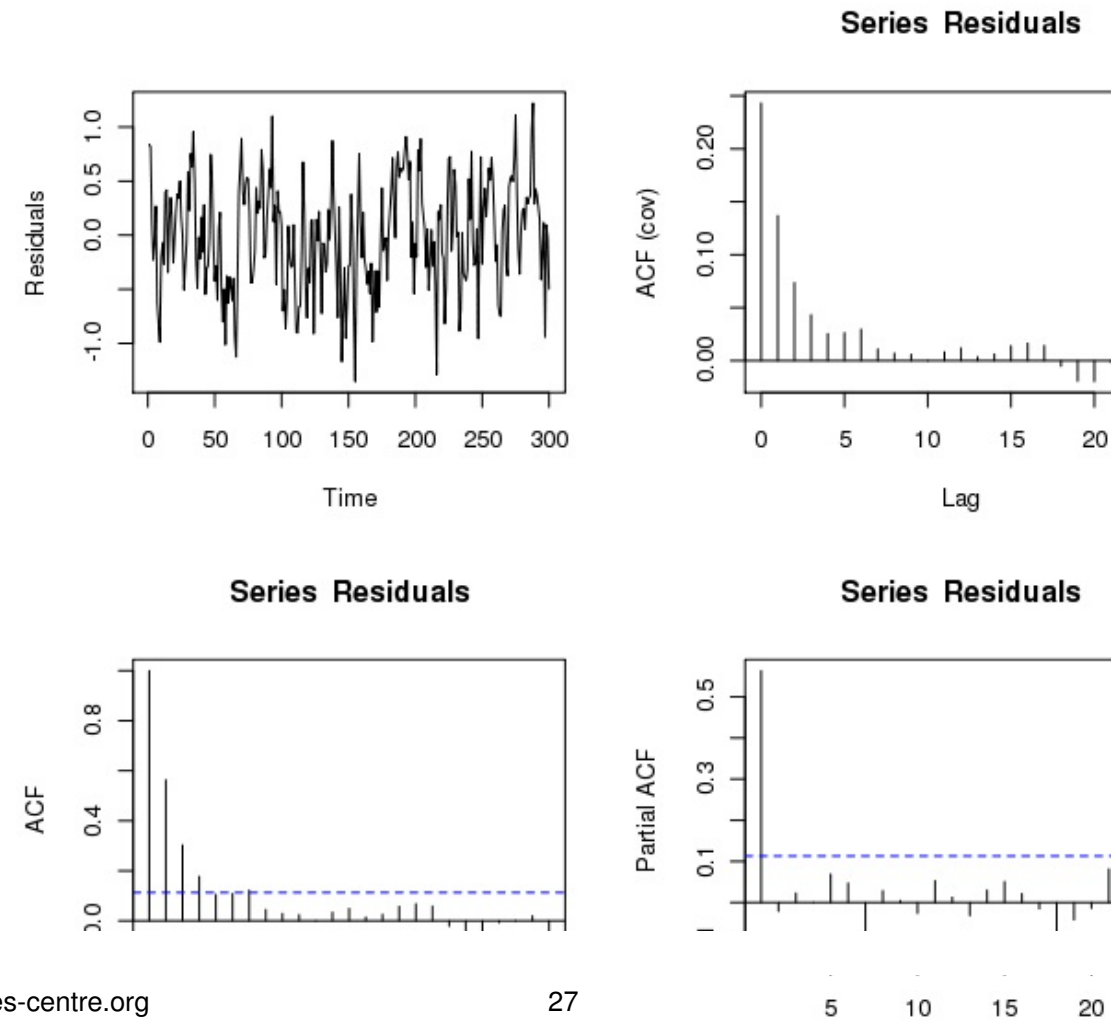
$$R(T) = R_{15}\theta^{T-15} \quad [mg/l]$$

$$C_m(T) = 14.54 - 0.39T + 0.01T^2 \quad [mg/l]$$

- Simple - however, a non-linear model.
- Uncertainty of prediction does not depend on horizon.

Model validation

The **autocorrelation** and **partial autocorrelation** function for the residuals from the first order model



Grey box model of oxygen concentration

The following nonlinear state space (Hidden Markov) model has been found:

***The system equation**:

$$\begin{bmatrix} dC \\ dL \end{bmatrix} = \begin{bmatrix} \frac{K}{h\sqrt{h}} & -K_c \\ K_3 & -K_l \end{bmatrix} \begin{bmatrix} C \\ L \end{bmatrix} dt + \begin{bmatrix} \beta & \frac{\sqrt{C}K_b}{h} \\ 0 & \gamma \end{bmatrix} \begin{bmatrix} I \\ P_r \end{bmatrix} dt + \begin{bmatrix} \frac{K}{h\sqrt{h}} C_m(T) - R(T) \\ 0 \end{bmatrix} dt + \begin{bmatrix} dW_1(t) \\ dW_2(t) \end{bmatrix}$$

***The observation equation**:

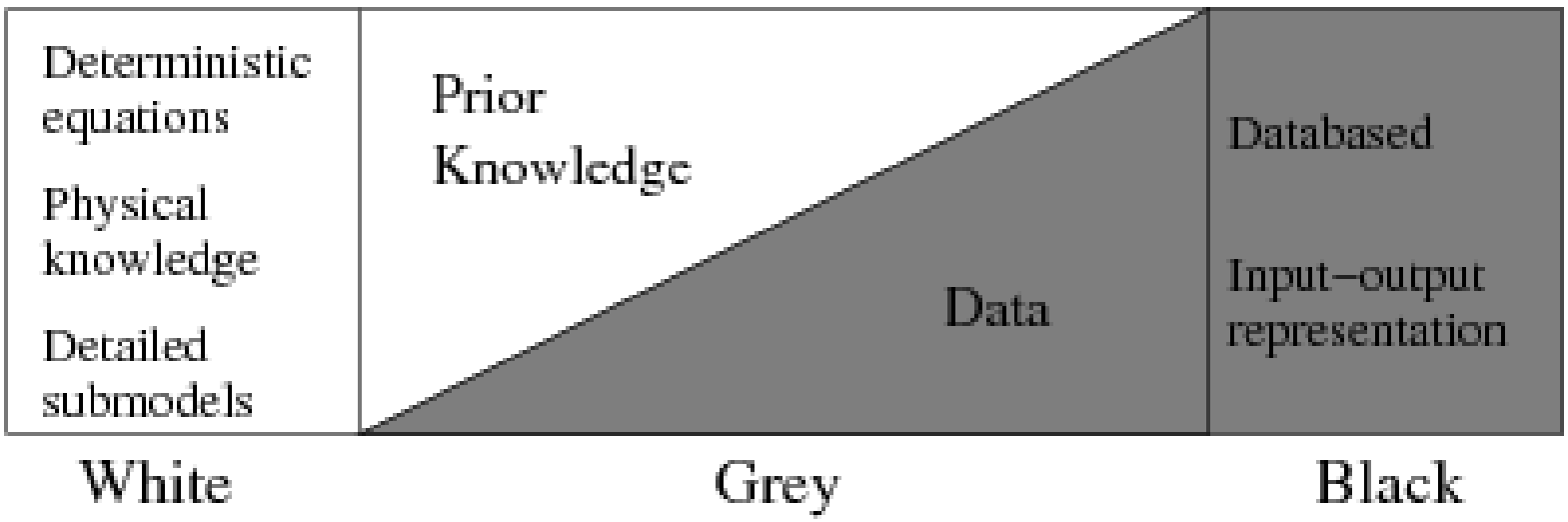
$$C_r = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} C \\ L \end{bmatrix} + e$$

Model types

- **White box models**
  - the model structure is known and deterministic.
  - uncertainty is discarded and the model tends to be overspecified.
- **Black box models**
  - data based models - input/output models.
  - the model and its parameters have little physical significance.
- **Grey box models**
  - between white and black box models

The grey box modelling concept

- Combines prior physical knowledge with information in data.
- The model is not completely described by physical equations, but equations and the parameters are physically interpretable.



Why use grey box modelling?

- Prior physical knowledge can be used.
- Non-linear and non-stationary models are easily formulated.
- Missing data are easily accommodated.
- It is possible to estimate environmental variables that are not measured.
- Available physical knowledge and statistical modelling tools is combined to estimate the parameters of a rather complex dynamic system.
- The parameters contain information from the data that can be directly interpreted by the scientists.
- Fewer parameters $\rightarrow$ more power in the statistical tests.
- The physical expert and the statistician can collaborate in the model formulation.

Stochastic Differential Equations (SDE's)

- Ordinary Differential Equations (ODE's) provide deterministic description of a system:

$$d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{U}_t, t)dt \quad t \geq 0.$$

  where $\boldsymbol{f}$ is a known function of the time t and the state $\boldsymbol{X}$ and input $\boldsymbol{U}$.
- To describe the deviation between the ODE and the true variation of the state an additive noise term is introduced.
- Physical arguments for including the noise part:
  1. Modelling approximations.
  2. Unrecognized inputs.
  3. Measurements of the input are noise corrupted.

The continuous-discrete time non-linear stochastic state space model

The system equation (set of Itô stochastic differential eqs.)

$$\mathrm{d}\boldsymbol{X}_t = f(\boldsymbol{X}_t, \boldsymbol{U}_t, \boldsymbol{\theta})\,\mathrm{d}t + G(\boldsymbol{X}_t, \boldsymbol{U}_t, \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{W}_t, \quad \boldsymbol{X}_{t_0} = \boldsymbol{X}_0$$

Notation

$\boldsymbol{X}_t \in \mathbb{R}^n$	State vector
$\boldsymbol{U}_t \in \mathbb{R}^r$	Known input vector
f	Drift term
G	diffusion term
$\boldsymbol{W}_t$	Wiener process of dimension, d, with incremental covariance $\boldsymbol{Q}_t$
$\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$	Unknown parameter vector

The observation equation

The observations are in discrete time, functions of state, input, and parameters, and are subject to noise:

$$\boldsymbol{Y}_{t_i} = h(\boldsymbol{X}_{t_i}, \boldsymbol{U}_{t_i}, \theta) + \boldsymbol{e}_{t_i}$$

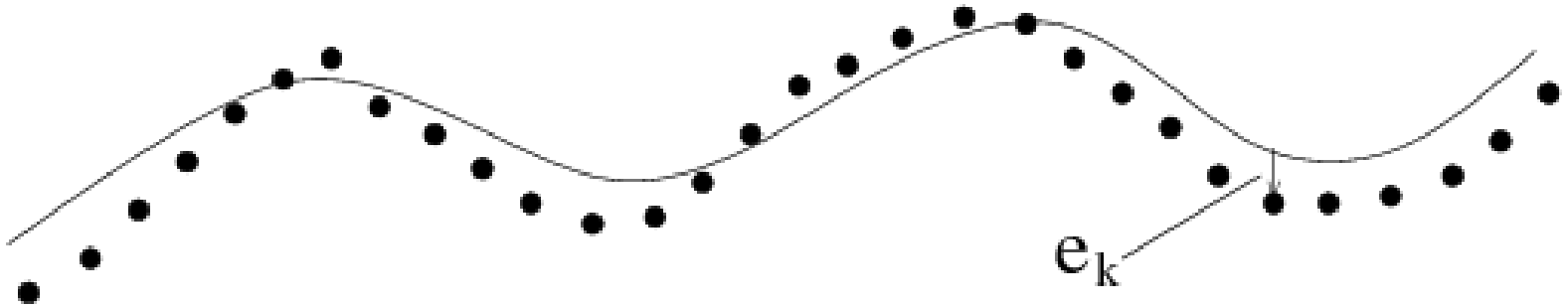
Notation

$\boldsymbol{Y}_{t_i} \in \mathbb{R}^m$	Observation vector
h	Observation function
$\boldsymbol{e}_{t_i} \in \mathbb{R}^m$	Gaussian white noise with covariance $\boldsymbol{\Sigma}_{t_i}$

Observations are available at the time points t_i : $t_1 < \ldots < t_i < \ldots < t_N$

$\boldsymbol{X}_0, \boldsymbol{W}_t, \boldsymbol{e}_{t_i}$ assumed independent for all $(t, t_i), t \neq t_i$

Stochastic Differential Equations (SDE's)



- The line demonstrates a model prediction, whereas the dots denote typical observations.
- Notice: Autocorrelated residuals are most often seen
  - this calls for SDE's.
- A situation as sketched above calls for using Stochastic Differential Equations (SDE's) as an alternative to Ordinary Differential Equations (ODE's).

ODE's - Characteristics

- Ordinary Differential Equations (ODE's) provide deterministic description of a system:

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t)dt \quad t \geq 0.$$

  where $\boldsymbol{f}$ is a deterministic function of the time t and the state $\boldsymbol{x}$.
- The solution to an ODE is a (deterministic) function
- For systems described by ODEs the future states can be predicted without any error!
- Parameters can calibrated using curve fitting methods (... but please check for uncorrelated residuals if you call it an estimate, if you are using statistical tests, or if you provide confidence intervals!).
- Consequently MLE and Prediction Error Methods are seldom the best methods for 'tuning the parameters'.

SDE's - Characteristics

- To describe the deviation between the ODE and the true variation of the state a system noise term is introduced, ie.

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{u}_t, t)\,\mathrm{d}t + G(\boldsymbol{X}_t, \boldsymbol{u}_t)\,\mathrm{d}\boldsymbol{W}_t,$$

- Reasons for including the system noise:
  1. Modelling approximations.
  2. Unrecognized inputs.
  3. Measurements of the input are noise corrupted.
- For an SDE's the solutions is a stochastic processes
- This implies that the future values are not know exactly (the outcomes are described a probability density function).
- Here proper statistical methods like MLE and Prediction Error Methods are appropriate for estimating the parameters – and we can easily test for hypotesis using statistical tests.

Advantages of Grey-box models (formulated as SDE's)

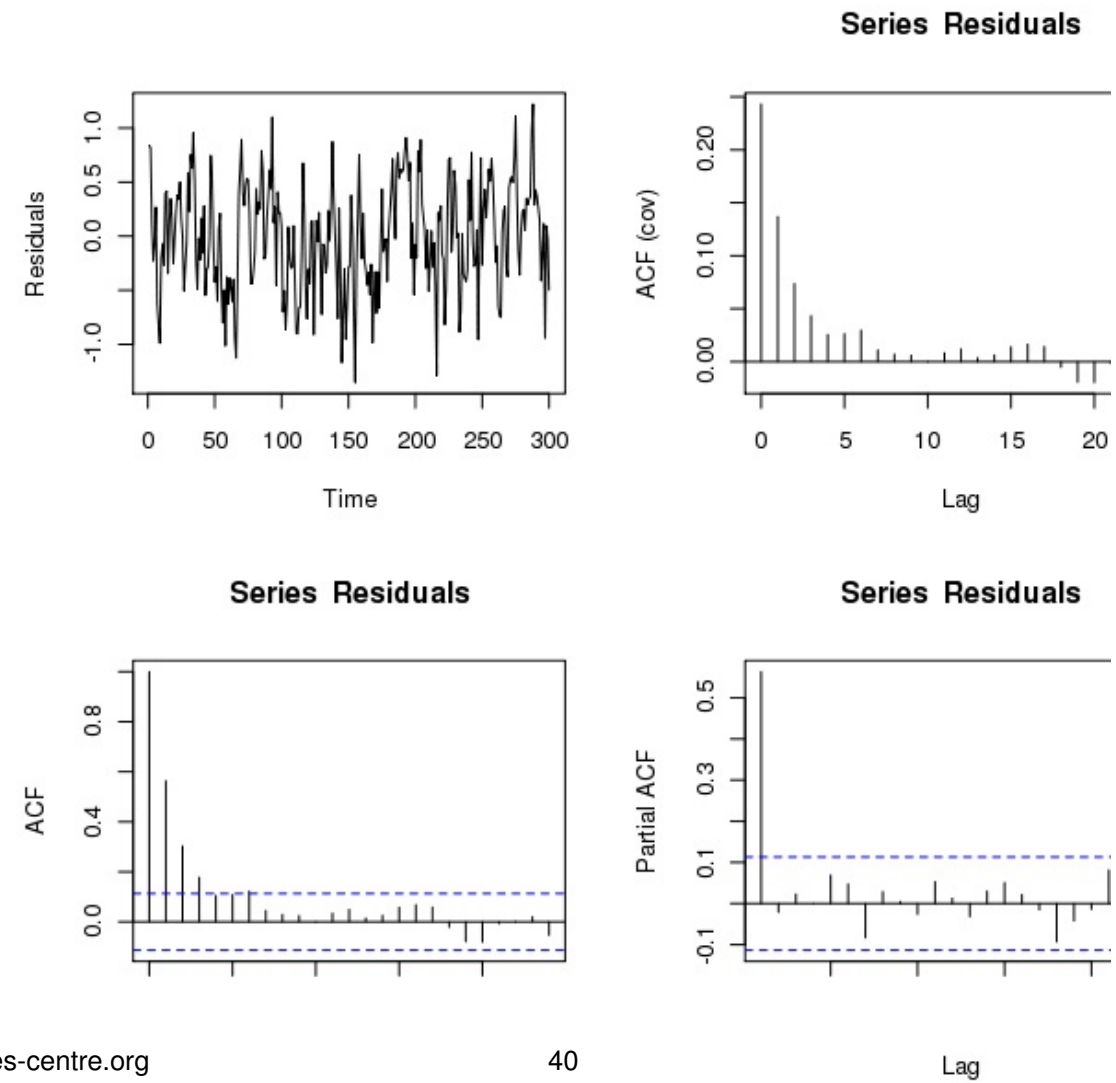
- Provides a decomposition of the total error into process error and measurement error.
- Facilitates use of statistical tools for model validation.
- Provides a systematic framework for pinpointing model deficiencies – will be demonstrated later on.
- Covariances of system error and measurement error are estimated.
- SDE based estimation gives more accurate and reliable parameter estimates than ODE based estimation.
- SDEs give more correct (more accurate and realistic) predictions and simulations.

Methods for Identification, Estimation and Model Validation

- **Model Identification:** See the next slides.
- **Parameter Estimation:**
  - Maximum Likelihood Methods
- **Model testing/selections:**
  - Test for significiant parameters (typically t-tests)
  - Test for model reductions (typically likelihood ratio tests)
  - Alternatively: Information criteria
- **Model Validation:**
  - Test whether the estimated model describes the data.
  - Autocorrelation functions – or Lag Dependent Functions.
  - Other classical methods ...

Identification of model order (here: number of states)

Use the **autocorrelation** and **partial autocorrelation** functions



Identification input variables and eg. time delays

Use **the Pre-whitening procedure and cross-correlation (pp. 223-226 in Time Series Analysis book)** or **Ridge regression (pp. 227-228 in TSA)**.

Identification of functional relations

Use **non-parametric methods (kernels, smoothing splines, etc.)** to estimate the **conditional mean** and the **conditional variance**.

- The conditional mean enters the drift term.
- The conditional variance enters the diffusion term.

Identification of Model Structure

- The **diffusion term** gives information for pinpointing model deficiencies.
- Assume that we based on 'large' values of relevant diffusion term(s) suspect $r \in \boldsymbol{\theta}$ to be a function of the states, input or time.
- Then consider the **extended state space model**:

$$\begin{aligned} \mathrm{d}\boldsymbol{X}_t &= f(\boldsymbol{X}_t, \boldsymbol{U}_t, \boldsymbol{\theta})\,\mathrm{d}t + G(\boldsymbol{X}_t, \boldsymbol{U}_t, \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{W}_t, \quad \boldsymbol{X}_{t_0} = \boldsymbol{X}_0 \\ \mathrm{d}r_t &= \mathrm{d}W_t^* \\ \boldsymbol{Y}_{t_i} &= h(\boldsymbol{X}_{t_i}, \boldsymbol{U}_{t_i}, \boldsymbol{\theta}) + \boldsymbol{e}_{t_i} \end{aligned} \tag{8}$$

which corresponds to a **random walk** description of r_t.

Identification of Model Structure

- Do we observe a significiant reduction of the relevant diffusion term(s)?
- In that case calculate the smoothed state estimate $\widehat{r}_{t|N}$ (use for instance the software tool CTSM-R - see slides by Rune Juhl).
- Plot $\widehat{r}_{t|N}$ versus the states, inputs and time.
- Identify a possible functional relationship.
- Build that functional relationship into the stochastic state space model.
- Estimate the model parameters and evaluate the improvement – using e.g. likelihood ratio tests.

This part of the lecture

- The likelihood principle
- Point estimation theory
- The likelihood function
- The score function
- The information matrix

The beginning of likelihood theory

- Fisher (1922) identified the likelihood function as the key inferential quantity conveying all inferential information in statistical modelling including the uncertainty
- The Fisherian school offers a Bayesian-frequentist compromise

A motivating example

Suppose we toss a thumbtack (used to fasten up documents to a background) 10 times and observe that 3 times it lands point up. Assuming we know nothing prior to the experiment, what is the probability of landing point up, θ?

- Binomial experiment with $y = 3$ and $n = 10$.
- P(Y=3;10,3,0.2) = 0.2013
- P(Y=3;10,3,0.3) = 0.2668
- P(Y=3;10,3,0.4) = 0.2150

A motivating example

By considering $P_\theta(Y = 3)$ to be a function of the unknown parameter we have the *likelihood function*:

$$L(\theta) = P_\theta(Y = 3)$$

In general, in a Binomial experiment with n trials and y successes, the likelihood function is:

$$L(\theta) = P_\theta(Y = y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

A motivating example

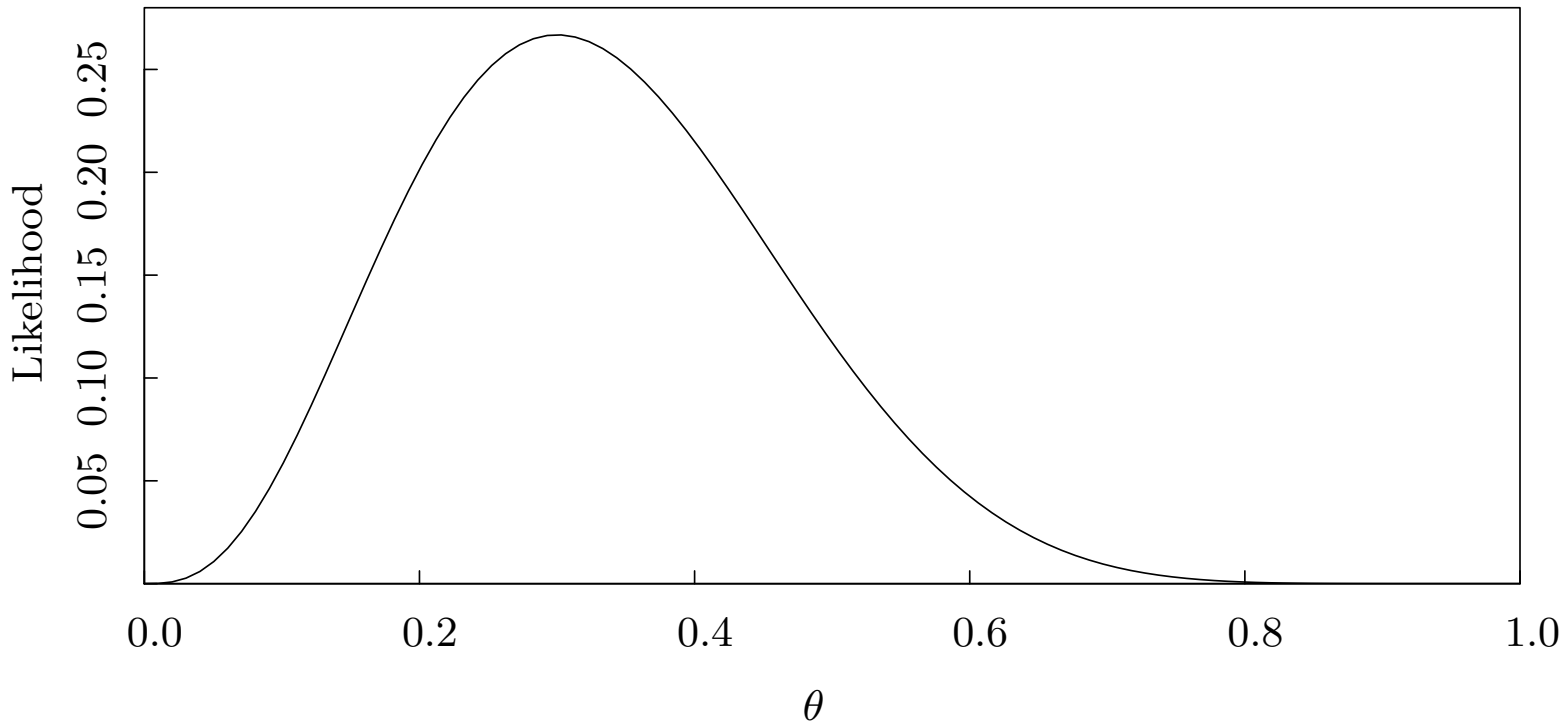


Figure: Likelihood function of the success probability θ in a binomial experiment with $n = 10$ and $y = 3$.

A motivating example

It is often more convenient to consider the log-likelihood function. The log-likelihood function is:

$$\log L(\theta) = y \log \theta + (n - y) \log(1 - \theta) + const$$

where *const* indicates a term that does not depend on θ.
By solving

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0$$

it is readily seen that the maximum likelihood *estimate* (MLE) for θ is

$$\widehat{\theta}(y) = \frac{y}{n} = \frac{3}{10} = 0.3$$

The likelihood principle

- Not just a method for obtaining a point estimate of parameters.
- It is the entire likelihood function that captures all the information in the data about a certain parameter.
- Likelihood based methods are inherently computational. In general numerical methods are needed to find the MLE.
- Today the likelihood principles play a central role in statistical modelling and inference.

Some syntax

- Multivariate random variable: $\boldsymbol{Y} = \{Y_1, Y_2, ..., Y_n\}^T$
- Observation set: $\{\boldsymbol{y} = y_1, y_2, \ldots, y_n\}^T$
- Joint density: $\{f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta^k}$
- Estimator (random) $\widehat{\boldsymbol{\theta}}(\mathbf{Y})$
- Estimate (number/vector) $\widehat{\boldsymbol{\theta}}(\mathbf{y})$

Point estimation theory

We will assume that the statistical model for $\boldsymbol{y}$ is given by parametric family of joint densities:

$$\{f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta^k}$$

Remember that when the n random variables are independent, the joint probability density equals the product of the corresponding marginal densities or:

$$f(y_1, y_2, \ldots y_n) = f_1(y_1) \cdot f_2(y_2) \cdot \ldots \cdot f_n(y_n)$$

Point estimation theory

Any estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ is said to be *unbiased* if

$$\mathrm{E}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$$

for all $\boldsymbol{\theta} \in \Theta^k$.

An estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ is said to be *uniformly minimum mean square error* if

$$\mathrm{E}\left[(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T\right] \leq \mathrm{E}\left[(\tilde{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T\right]$$

for all $\boldsymbol{\theta} \in \Theta^k$ and all other estimators $\tilde{\boldsymbol{\theta}}(\boldsymbol{Y})$.

Point estimation theory

- By considering the class of unbiased estimators it is most often not possible to establish a suitable estimator.
- We need to add a criterion on the variance of the estimator.
- A low variance is desired, and in order to evaluate the variance a suitable lower bound is given by the Cramer-Rao inequality.

Point estimation theory

Given the parametric density $f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta^k$, *for the observations* $\boldsymbol{Y}$. *Subject to certain regularity conditions, the variance of any unbiased estimator* $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ *of* $\boldsymbol{\theta}$ *satisfies the inequality*

$$\text{Var}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right] \geq \boldsymbol{i}^{-1}(\boldsymbol{\theta})$$

where $\boldsymbol{i}(\boldsymbol{\theta})$ *is the Fisher information matrix defined by*

$$\boldsymbol{i}(\boldsymbol{\theta}) = \text{E}\left[\left(\frac{\partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right]$$

and $\text{Var}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right] = \text{E}\left[(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T\right]$.

Point estimation theory

An unbiased estimator is said to be *efficient* if its covariance is equal to the Cramer-Rao lower bound.

The matrix Var $\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right]$ is often called a variance covariance matrix since it contains variances in the diagonal and covariances outside the diagonal. This important matrix is often termed the *Dispersion matrix*.

The likelihood function

- The likelihood function is built on an assumed parameterized statistical model as specified by a parametric family of joint densities for the observations $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)^T$.
- The *likelihood* of any specific value $\boldsymbol{\theta}$ of the parameters in a model is (proportional to) the probability of the actual outcome, $Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n$, calculated for the specific value $\boldsymbol{\theta}$.
- The likelihood function is simply obtained by considering the likelihood as a function of $\boldsymbol{\theta} \in \Theta^k$.

The likelihood function

Given the parametric density $f_{\boldsymbol{Y}}(\boldsymbol{y}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta^P$, for the observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ the *likelihood function for* $\boldsymbol{\theta}$ is the function

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = c(y_1, y_2, \ldots, y_n) f_Y(y_1, y_2, \ldots, y_n; \boldsymbol{\theta})$$

where $c(y_1, y_2, \ldots, y_n)$ is a constant.

The likelihood function is thus (proportional to) the joint probability density for the actual observations considered as a function of $\boldsymbol{\theta}$.

The log-likelihood function

- Very often it is more convenient to consider the *log-likelihood* function defined as

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \log(L(\boldsymbol{\theta}; \boldsymbol{y})).$$

- Sometimes the likelihood and the log-likelihood function will be written as $L(\boldsymbol{\theta})$ and $l(\boldsymbol{\theta})$, respectively, i.e. the dependency on $\boldsymbol{y}$ is suppressed.

The information matrix

The matrix

$$\boldsymbol{j}(\boldsymbol{\theta};\boldsymbol{y}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} l(\boldsymbol{\theta};\boldsymbol{y})$$

with the elements

$$\boldsymbol{j}(\boldsymbol{\theta};\boldsymbol{y})_{ij} = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\boldsymbol{\theta};\boldsymbol{y})$$

is called the *observed information* corresponding to the observation $\boldsymbol{y}$, evaluated in $\widehat{\boldsymbol{\theta}}$.

The observed information is thus equal to the Hessian (with opposite sign) of the log-likelihood function evaluated at $\boldsymbol{\theta}$. The Hessian matrix is simply (with opposite sign) the *curvature* of the log-likelihood function.

Example: Likelihood function for mean of normal distribution

An automatic production of a bottled liquid is considered to be stable. A sample of three bottles were selected at random from the production and the volume of the content volume was measured. The deviation from the nominal volume of 700.0 ml was recorded.

The deviations (in ml) were 4.6; 6.3; and 5.0.

Example: Likelihood function for mean of normal distribution

First a *model* is formulated

i Model: C+E (center plus error) model, $Y = \mu + \epsilon$

ii Data: $Y_i = \mu + \epsilon_i$

iii Assumptions:

- Y_1, Y_2, Y_3 are independent
- $Y_i \sim \mathrm{N}(\mu, \sigma^2)$
- σ^2 is known, $\sigma^2 = 1$,

Thus, there is only one unknown model parameter, $\mu_Y = \mu$.

Example: Likelihood function for mean of normal distribution

The joint probability density function for Y_1, Y_2, Y_3 is given by

$$f_{Y_1,Y_2,Y_3}(y_1, y_2, y_3; \mu) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y_1-\mu)^2}{2}\right] \times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y_2-\mu)^2}{2}\right] \times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y_3-\mu)^2}{2}\right]$$

which for every value of μ is a function of the three variables y_1, y_2, y_3.

Remember that the normal probability density is: $f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$

Example: Likelihood function for mean of normal distribution

Now, we have the observations, $y_1 = 4.6$; $y_2 = 6.3$ and $y_3 = 5.0$, and establish the likelihood function

$$\begin{aligned} L_{4.6,6.3,5.0}(\mu) &= f_{Y_1,Y_2,Y_3}(4.6, 6.3, 5.0; \mu) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(4.6-\mu)^2}{2}\right] \\ &\times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(6.3-\mu)^2}{2}\right] \\ &\times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(5.0-\mu)^2}{2}\right] \end{aligned}$$

The function depends only on μ.
Note that the likelihood function expresses the infinitesimal probability of obtaining the sample result $(4.6, 6.3, 5.0)$ as a function of the unknown parameter μ.

Example: Likelihood function for mean of normal distribution

Reducing the expression one finds

$$L_{4.6,6.3,5.0}(\mu) = \frac{1}{(\sqrt{2\pi})^3} \exp\left[-\frac{1.58}{2}\right] \exp\left[-\frac{3(5.3-\mu)^2}{2}\right]$$
$$= \frac{1}{(\sqrt{2\pi})^3} \exp\left[-\frac{1.58}{2}\right] \exp\left[-\frac{3(\bar{y}-\mu)^2}{2}\right]$$

which shows that (except for a factor not depending on μ), the likelihood function does only depend on the observations (y_1, y_2, y_3) through the average $\bar{y} = \sum y_i/3$.

Example: Likelihood function for mean of normal distribution

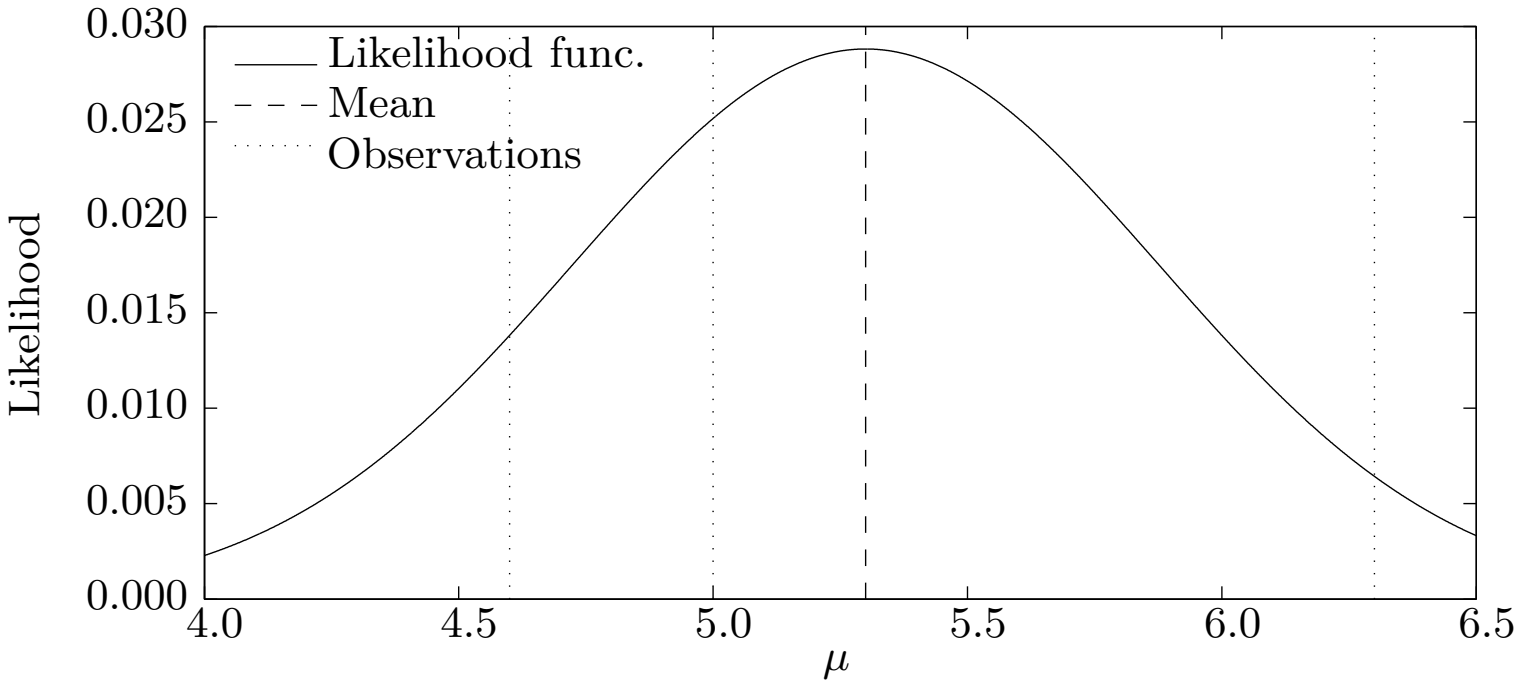


Figure: The likelihood function for μ given the observations $y_1 = 4.6$; $y_2 = 6.3$ and $y_3 = 5.0$.

Invariance property

Assume that $\widehat{\boldsymbol{\theta}}$ is a maximum likelihood estimator for $\boldsymbol{\theta}$, and let $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$ denote a one-to-one mapping of $\Omega \subset \mathbb{R}^k$ onto $\Psi \subset \mathbb{R}^k$. Then the estimator $\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})$ is a maximum likelihood estimator for the parameter $\boldsymbol{\psi}(\boldsymbol{\theta})$.

The principle is easily generalized to the case where the mapping is not one-to-one.

Distribution of the ML estimator

We assume that $\widehat{\boldsymbol{\theta}}$ is consistent. Then, under some regularity conditions,

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \rightarrow \mathrm{N}(0, \boldsymbol{i}(\boldsymbol{\theta})^{-1})$$

where $\boldsymbol{i}(\boldsymbol{\theta})$ is the expected information or the information matrix.

The results can be used for inference under very general conditions. As the price for the generality, the results are only asymptotically valid.

- Asymptotically the variance of the estimator is seen to be equal to the Cramer-Rao lower bound for any unbiased estimator.
- The practical significance of this result is that the MLE makes efficient use of the available data for large data sets.

Distribution of the ML estimator

In practice, we would use

$$\widehat{\theta} \sim \mathrm{N}(\theta, \boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}}))$$

where $\boldsymbol{j}(\widehat{\boldsymbol{\theta}})$ is the observed (Fisher) information.

This means that asymptotically

i) $E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$

ii) $D[\widehat{\boldsymbol{\theta}}] = \boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}})$

Distribution of the ML estimator

- The standard error of $\widehat{\theta}_i$ is given by

$$\widehat{\sigma}_{\hat{\theta}_i} = \sqrt{\mathrm{Var}_{ii}[\widehat{\theta}]}$$

  where $\mathrm{Var}_{ii}[\widehat{\theta}]$ is the i'th diagonal term of $\boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}})$
- Hence we have that an estimate of the dispersion (variance-covariance matrix) of the estimator is

$$D[\widehat{\boldsymbol{\theta}}] = \boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}})$$

- An estimate of the uncertainty of the individual parameter estimates is obtained by decomposing the dispersion matrix as follows:

$$D[\widehat{\boldsymbol{\theta}}] = \widehat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{\theta}}} \boldsymbol{R} \widehat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{\theta}}}$$

  into $\widehat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{\theta}}}$, which is a diagonal matrix of the standard deviations of the individual parameter estimates, and $\boldsymbol{R}$, which is the corresponding correlation matrix. The value R_{ij} is thus the estimated correlation between $\widehat{\theta}_i$ and $\widehat{\theta}_j$.

The Wald Statistic

A test of an individual parameter

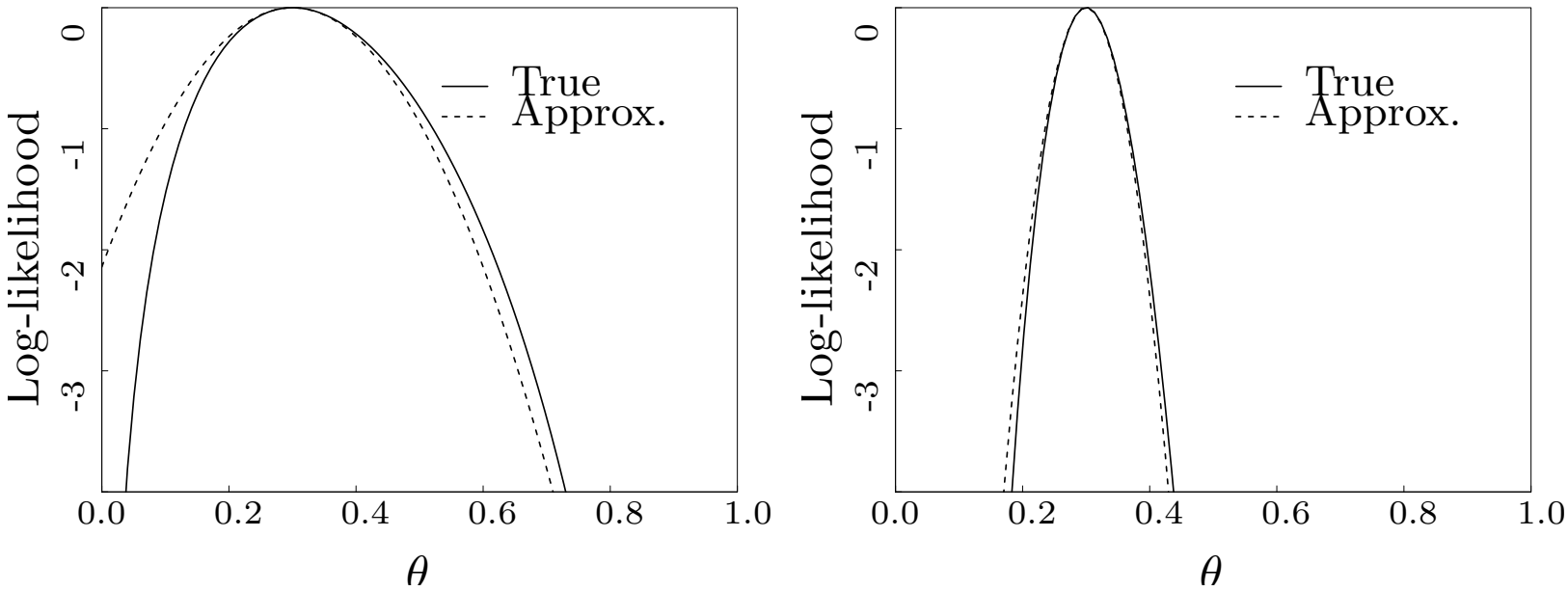
$$\mathcal{H}_0 : \theta_i = \theta_{i,0}$$

is given by the *Wald statistic*:

$$Z_i = \frac{\widehat{\theta}_i - \theta_{i,0}}{\widehat{\sigma}_{\hat{\theta}_i}}$$

which under $\mathcal{H}_0$ is approximately N$(0, 1)$-distributed.

Example: Quadratic approximation of the log-likelihood



(a) $n = 10, y = 3$ **(b)** $n = 100, y = 30$

Figure: Quadratic approximation of the log-likelihood function.

Likelihood ratio tests

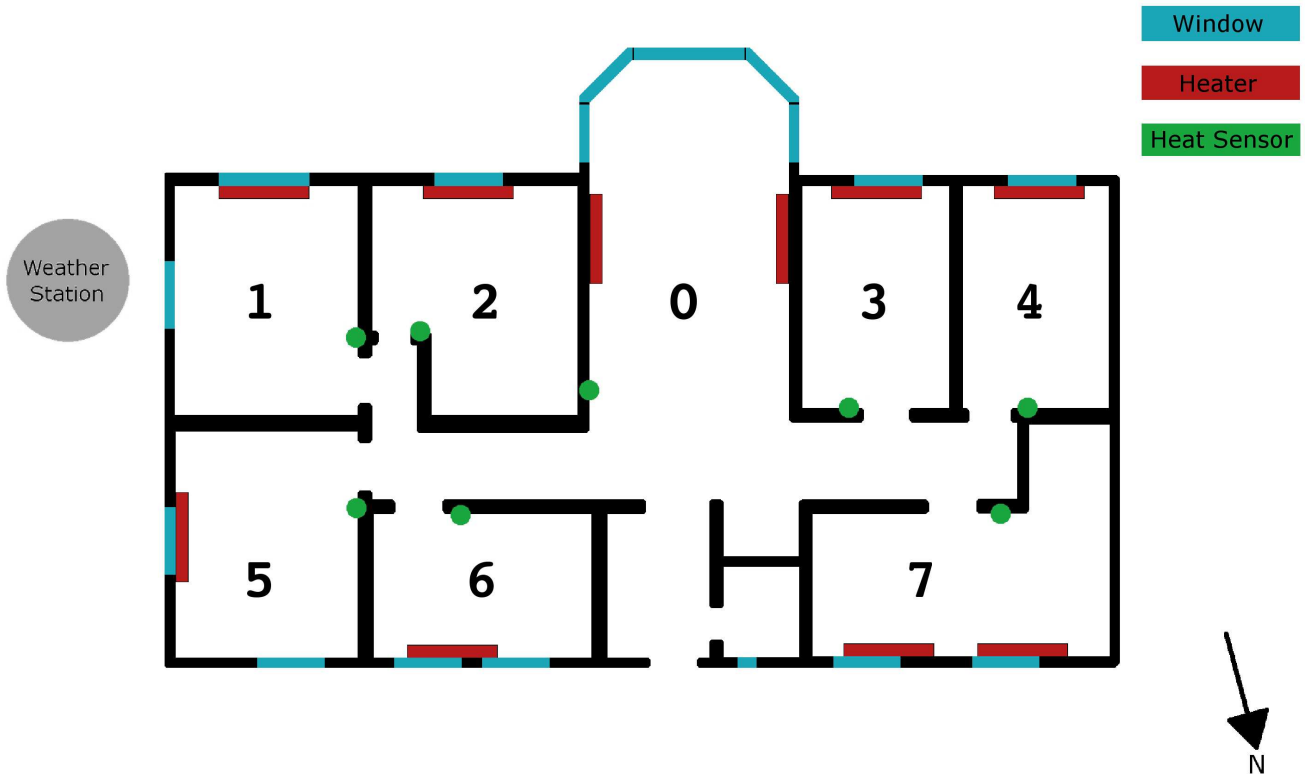
- Want to know the distribution of D when assuming $\mathcal{H}_0$ (model B).
- It is sometimes possible to calculate the exact distribution. This is for instance the case for the General Linear Model for Gaussian data.
- In most cases, however, we must use following important result regarding the asymptotic behavior.

The random variable $D = 2(\ell_A(\widehat{\boldsymbol{\theta}_A}, \boldsymbol{Y}) - \ell_B(\widehat{\boldsymbol{\theta}_B}, \boldsymbol{Y}))$ *converges in law to a* χ^2 *random variable with* $f = (\dim(\Omega_A) - \dim(\Omega_B))$ *degrees of freedom, i.e.,*

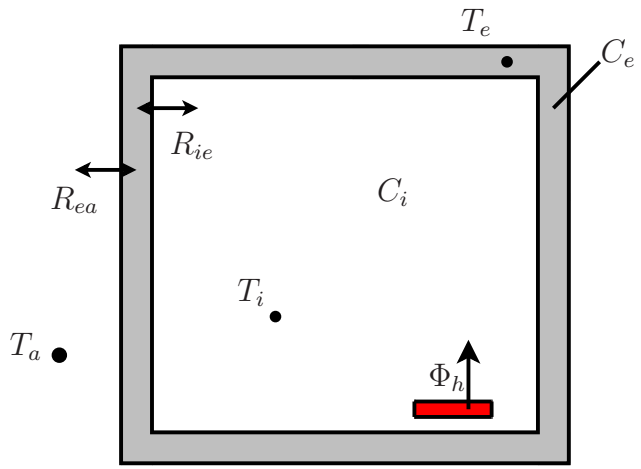
$$D \rightarrow \chi^2(f)$$

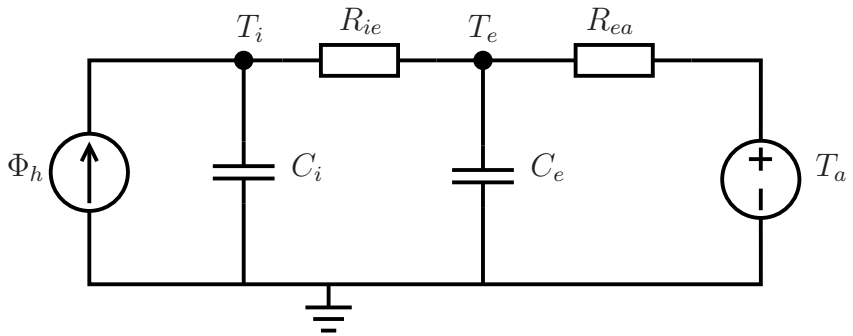
under $\mathcal{H}_0$.

Flexhouse layout

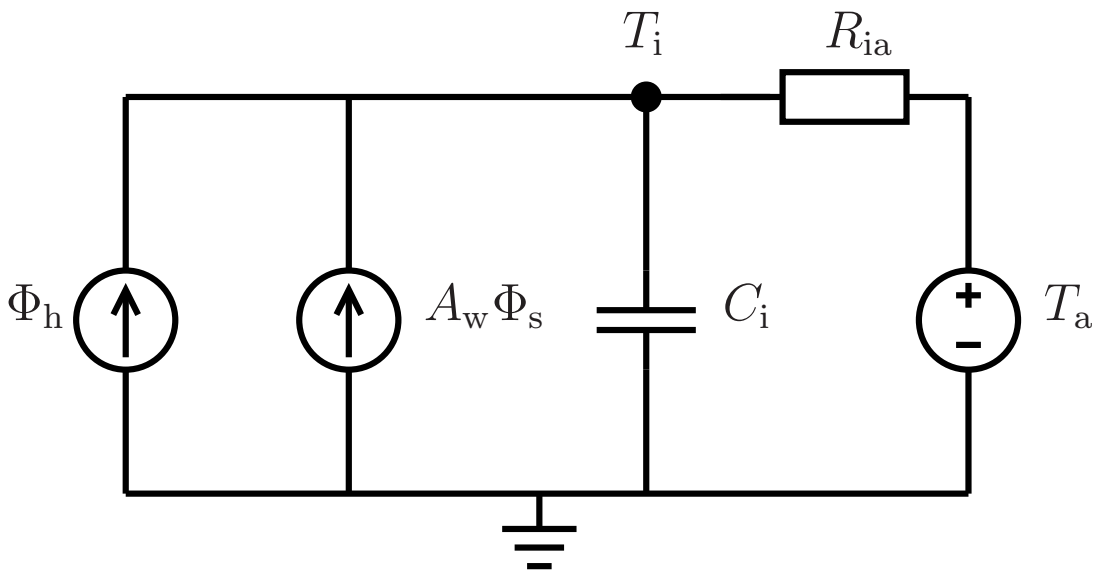


RC-diagram ofte used for illustrating linear models





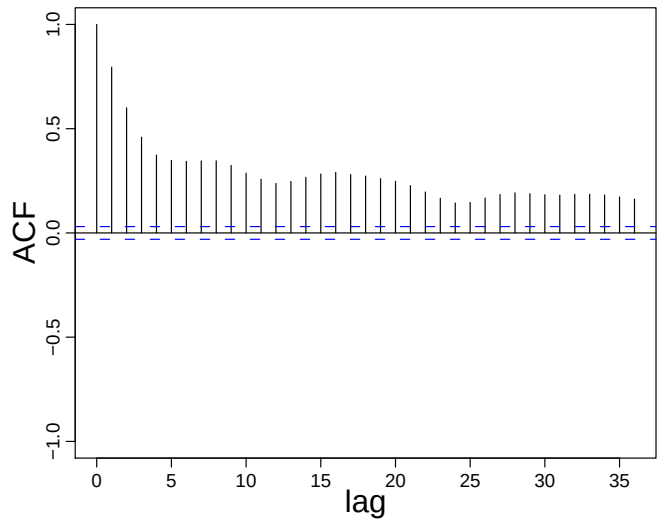
Model A

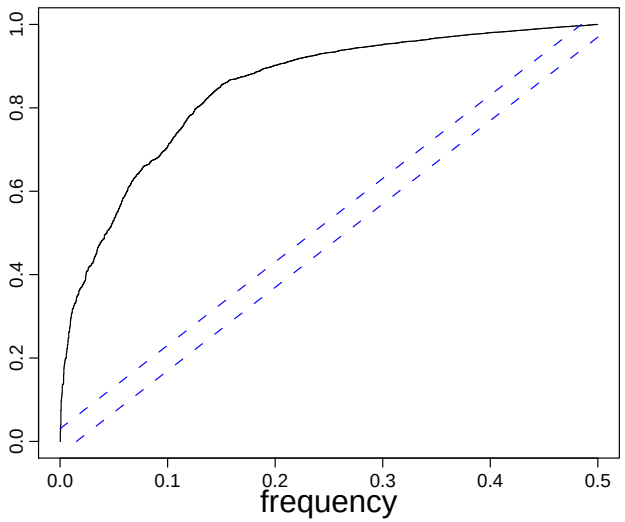


A_w is the effective window area.

Model Validation for Model A

Autocorrelation function and Periodogram for the residuals.

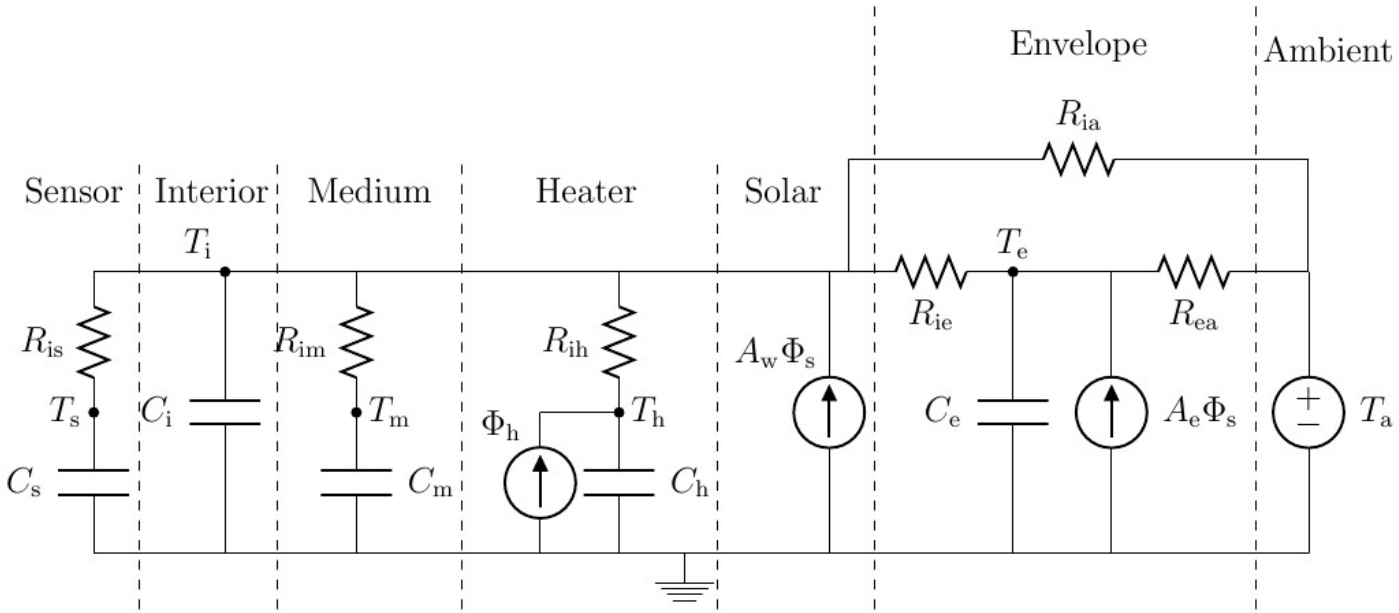




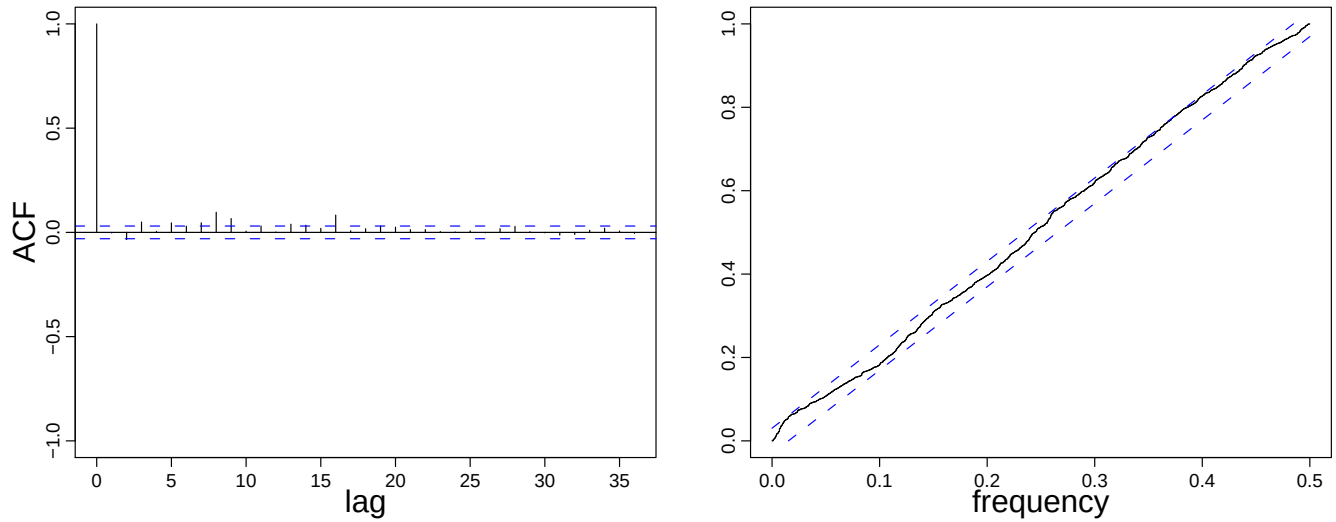
Model is seen not to be adequate.

Model E

After some steps: (Notice that eg. the electric heating system is included)



Model Validation for Model E.



It is concluded that the model is adequate.

Continuous Time Stochastic Modelling (CTSM-R)

- The parameter estimation is performed by using the software CTSM-R.
- The software has been developed at DTU Compute
- Download from www.ctsm.info (see also slides by Rune Juhl)
- The program returns the uncertainty of the parameter estimates as well.

The estimation procedure (CTSM-R)

CTSM-R is based on

- The Extended Kalman Filter
- Approximate likelihood estimation

The estimation procedure (CTSM-R)

CTSM-R is based on

- The Extended Kalman Filter
- Approximate likelihood estimation

and provides eg.

- Likelihood testing for nested models
- Calculations of smoothen state $E[\boldsymbol{X}_t|\mathcal{Y}_T]$
- Calculations of k-step predictions $E[\boldsymbol{X}_t|\mathcal{Y}_{t-k}]$.
- Calculations of noise free simulations $E[\boldsymbol{X}_t|\mathcal{Y}_{t_0}]$

The continuous-discrete time stochastic state space formulation

General formulation

$$
\begin{aligned}
d\boldsymbol{x}_t &= \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)dt + \boldsymbol{\sigma}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)d\boldsymbol{w}_t \\
\boldsymbol{y}_k &= \boldsymbol{h}(\boldsymbol{x}_{t_k}, \boldsymbol{u}_{t_k}, \boldsymbol{\theta}, \boldsymbol{e}_k, t_k),
\end{aligned}
$$

where

- $\boldsymbol{x}_t$ is the continuous time state and $\boldsymbol{y}_k \in \mathbb{R}^l$ is the discrete time observations.
- $\boldsymbol{u}_t \in \mathbb{R}^r$ is the inputs
- $\boldsymbol{\theta} \in \mathbb{R}^p$ is a parameter vector
- $\boldsymbol{e}_k \in \mathbb{R}^l$ is a random observation error.

The estimation procedure (CTSM) - Limitations

Most general set up in CTSM

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)dt + \boldsymbol{\sigma}(\boldsymbol{u}_t, \boldsymbol{\theta}, t)d\boldsymbol{w}_t$$
$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_{t_k}, \boldsymbol{u}_{t_k}, \boldsymbol{\theta}, t_k) + \boldsymbol{e}_k,$$

where

- $\boldsymbol{\sigma} \in \mathbb{R}^{n \times n}$ is a quadratic matrix, independent of the state
- $\boldsymbol{e}_k \sim N(\boldsymbol{0}, \boldsymbol{S}_k(\boldsymbol{\theta}, \boldsymbol{u}_k))$ is a Gaussian random variable.

This limitation can in most cases be solved using the Lamperti Transformation - see the reference list later on.

Transformation of the State Space 1

Consider the system equation

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)dt + \boldsymbol{\sigma}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t)\boldsymbol{R}(\boldsymbol{u}_t, \boldsymbol{\theta}, t)d\boldsymbol{w}_t,$$

where $\boldsymbol{R}(\boldsymbol{u}_t, \boldsymbol{\theta}, t) \in \mathbb{R}^{n \times n}$ is any matrix function and $\boldsymbol{\sigma} \in \mathbb{R}^{n \times n}$ is a diagonal matrix

$$\sigma_{ii}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}, t) = \sigma_i(x_{i,t}, \boldsymbol{u}_t, \boldsymbol{\theta}, t)$$

Transformation of the State Space 2

Choose the transformation

$$z_{i,t} = \psi^i(x_{i,t}, \boldsymbol{u}_t, \boldsymbol{\theta}, t) = \int \frac{d\xi}{\sigma_i(\xi, \boldsymbol{u}_t, \boldsymbol{\theta}, t)}\Bigg|_{\xi=x_i},$$

Transformation of the State Space 2

Choose the transformation

$$z_{i,t} = \psi^i(x_{i,t}, \boldsymbol{u}_t, \boldsymbol{\theta}, t) = \int \frac{d\xi}{\sigma_i(\xi, \boldsymbol{u}_t, \boldsymbol{\theta}, t)}\bigg|_{\xi=x_i},$$

then by Itô's lemma z_i is also an Itô process given by

$$\begin{aligned} dz_{i,t} =& \frac{\partial}{\partial t}\psi^i(\cdot, t)dt + \frac{f_i(\cdot)}{\sigma_i(\cdot)}dt - \frac{1}{2}\sigma_i(\cdot)\sum_{j=1}^{n}[\boldsymbol{R}(\cdot)]^2_{i,j}\,dt \\ &+ \sum_{j=1}^{n}[\boldsymbol{R}(\cdot)]_{i,j}\,dw_j, \end{aligned}$$

where the diffusion term is now independent of the state z_i.

Summary

By applying the grey box modelling approach

- physical/prior knowledge and information in data are combined, ie. we have bridged the gap between physical and statistical modelling.
- many statistical, mathematical and physical methods for model validation and structure modification become available.
- parameter estimates have physical significance - seldom the case for black box models.
- we obtain more accurate predictions and more realistic prediction intervals.
- we obtain realistic simulations (ODE based models do not provide a reasonable framework for simulations)

Some References - Generic

- H. Madsen: *Time Series Analysis*, Chapman and Hall, 392 pp, 2008.
- H. Madsen and P. Thyregod (2011): *An Introduction to General and Generalized Linear Models*, Chapman and Hall, 340 pp.
- J.M. Morales, A.J. Conejo, H. Madsen, P. Pinson, M. Zugno: *Integrating Renewables in Electricity Markes*, Springer, 430 pp., 2013.
- H.Aa. Nielsen, H. Madsen: *A generalization of some classical time series tools*, Computational Statistics and Data Analysis, Vol. 37, pp. 13-31, 2001.
- H. Madsen, J. Holst: *Estimation of Continuous-Time Models for the Heat Dynamics of a Building*, Energy and Building, Vol. 22, pp. 67-79, 1995.
- P. Sadegh, J. Holst, H. Madsen, H. Melgaard: *Experiment Design for Grey Box Identification*, Int. Journ. Adap. Control and Signal Proc., Vol. 9, pp. 491-507, 1995.
- P. Sadegh, L.H. Hansen, H. Madsen, J. Holst: *Input Design for Linear Dynamic Systems using Maxmin Criteria*, Journal of Information and Optimization Sciences, Vol. 19, pp. 223-240, 1998.
- J.N. Nielsen, H. Madsen, P.C. Young: *Parameter Estimation in Stochastic Differential Equations; An Overview*, Annual Reviews in Control, Vol. 24, pp. 83-94, 2000.
- N.R. Kristensen, H. Madsen, S.B. Jørgensen: *A Method for systematic improvement of stochastic grey-box models*, Computers and Chemical Engineering, Vol 28, 1431-1449, 2004.
- N.R. Kristensen, H. Madsen, S.B. Jørgensen: *Parameter estimation in stochastic grey-box models*, Automatica, Vol. 40, 225-237, 2004.
- J.B. Jørgensen, M.R. Kristensen, P.G. Thomsen, H. Madsen: *Efficient numerical implementation of the continuous-discrete extended Kalman Filter*, Computers and Chemical Engineering, 2007.
- K.R. Philipsen, L.E. Christiansen, H. Hasman, H. Madsen: *Modelling conjugation with stochastic differential equations*, Journal of Theoretical Biology, Vol. 263, pp. 134-142, 2010.

Some References - Heat Dynamics of Buildings

- H. Madsen, J. Holst: *Estimation of Continuous-Time Models for the Heat Dynamics of a Building*, Energy and Building, Vol. 22, pp. 67-79, 1995.
- B. Nielsen, H. Madsen: *Identification of Transfer Functions for Control of Greenhouse Air Temperature*. J. Agric. Engng. Res., Vol. 60, pp. 25-34. 1995.
- K.K. Andersen, H. Madsen, L. Hansen: *Modelling the heat dynamics of a building using stochastic differential equations*, Energy and Buildings, Vol. 31, pp. 13-24, 2000.
- K.K. Andersen, O.P. Palsson, H. Madsen, L.H. Knudsen: *Experimental design and setup for heat exchanger modelling*, International Journal of Heat Exchangers, Vol. 1, pp. 163-176, 2001.
- H.Aa. Nielsen, H. Madsen: *Modelling the heat consumption in district heating systems using a grey-box approach*, Energy and Buildings, Vol. 38, pp. 63-71, 2006.
- M.J. Jiménez, H. Madsen: *Models for Describing the Thermal Characteristics of Building Components*, Building and Energy, Vol. 43, pp. 152-162, 2008.
- M.J. Jiménez, H. Madsen, K.K. Andersen: *Identification of the Main Thermal Characteristics of Building Components using MATLAB* , Building and Energy, Vol. 43, pp. 170-180, 2008.
- N. Friling, M.J. Jimenez, H. Bloem, H. Madsen: *Modelling the heat dynamics of building integrated and ventilated photovoltaic modules*, Energy and Building, Vol. 41(10), pp. 1051-1057, 2009.
- P. Bacher, H. Madsen: *Identifying suitable models for the heat dynamics of buildings*, Vol. 43, pp. 1511-1522, 2011.
- O. Corradi, H. Ochesenfeld, H. Madsen, P. Pinson, *Controlling Electricity Consumption by Forecasting its Response to Varying Prices*, IEEE Transactions, Vol. 8, pp. 421-429, 2013.
- P.H. Delff Andersen, A. Iversen, H. Madsen, C. Rode: *Dynamic Modeling of Presence of Occupants using Inhomogeneous Markov Chains*, Energy and Buildings, Vol. 69, pp. 213-223, 2014.
- I. Naveros, P. Bacher, D.P. Ruiz, M.J. Jimenez, H. Madsen: *Setting up and validating a complex model for a simple homogeneous wall*, Energy and Buildings, Vol. 70,pp. 303-317, 2014.
- **See www.henrikmadsen.org for more articles and for downloads**