Introduction to ctsm-r

(Based on slides created by Rune Juhl)

Peder Bacher

DTU Compute, Dynamical Systems
Building 303B, Room 010
DTU - Technical University of Denmark
2800 Lyngby – Denmark
e-mail: pbac@dtu.dk

Summer school 2021:

Time Series Analysis - with a focus on Modelling and Forecasting in Energy Systems
Population and sample

(Infinit) Statistical population

Randomly selected

Mean \( \mu \)

Sample \( \{y_1, y_2, \ldots, y_n\} \)

Sample mean \( \bar{x} \)

Statistical Inference
Parameter estimation with example

Simplest example: a constant model for the mean

- Model
  \[ Y_i = \mu + \varepsilon_i \], where \( \varepsilon_i \sim N(0, \sigma^2) \) and i.i.d.

- i.i.d.: identically and independent distributed

- The parameters are: the mean \( \mu \) and the standard error \( \sigma \)

- We take a sample of \( n = 10 \) observations
  \[(y_1, y_2, \ldots, y_{10})\]
Likelihood

The likelihood is defined by the joint probability of the data

\[
L(\mu, \sigma) \equiv p(y_1, y_2, \ldots, y_{10} | \mu, \sigma)
\]

Hence, it’s a function of the two parameters (the sample is observed, so it is not varying). Due to independence

\[
= \prod_{i=1}^{10} p(y_i | \mu, \sigma)
\]

We assume in our model that the error \( \varepsilon_i = Y_i - \mu \) is normal distributed (Gaussian), so

\[
p(y_i | \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(y_i - \mu)^2}{2\sigma^2} \right)
\]

(1)
Maximum likelihood estimation

Parameter estimation

\[ \hat{\theta} = \arg \min_{\theta \in \Theta} \left( -\ln(L(\theta)) \right) \]

where \( \theta = (\mu, \sigma) \)
Maximum likelihood estimation

Parameter estimation

\[ \hat{\theta} = \arg \min_{\theta \in \Theta} (-\ln(L(\theta))) \]

where \( \theta = (\mu, \sigma) \)

Run the example in R
Likelihood for time correlated data

Given a sequence of measurements $\mathcal{Y}_N$

$$L(\theta) = p(\mathcal{Y}_N|\theta) = p(y_N, y_{N-1}, \ldots, y_0|\theta)$$

$$= \left( \prod_{k=1}^{N} p(y_k|\mathcal{Y}_{k-1}, \theta) \right) p(y_0|\theta)$$

Parameter estimation

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \left(-\ln(L(\theta))\right)$$
Likelihood for time correlated data

If Gaussian

\[ \hat{y}_{k|k-1} = E[y_k|\mathcal{Y}_{k-1}, \theta] \]
\[ R_{k|k-1} = V[y_k|\mathcal{Y}_{k-1}, \theta] \]
\[ \epsilon_k = y_k - \hat{y}_{k|k-1} \]

then the likelihood is

\[ L(\theta) = \left( \prod_{k=1}^{N} \frac{\exp\left(-\frac{1}{2} \epsilon_k^T R_{k|k-1}^{-1} \epsilon_k\right)}{\sqrt{|R_{k|k-1}| \sqrt{2\pi}^l}} \right) \]

Maximised using quasi Newton
Kalman filter

Parameter estimation with maximum likelihood

Predicted step
Based on e.g.
physical model

Prior knowledge
of state

Next timestep
$k \leftarrow k + 1$

Update step
Compare prediction
to measurements

Output estimate
of state

Measurements

Figure: "Basic concept of Kalman filtering" by Petteri Aimonen. Wikipedia
Introduction to grey-box modelling and **ctsmr**
Grey-box modelling

Figure: Ak et al. 2012
Grey-box modelling

Bridges the gap between physical and statistical modelling. THERE is a manual on ctsm.info
ctsmr

Continuous Time Stochastic Modelling in R
ctsmr

Continuous Time Stochastic Modelling in R

more correctly

Continuous-Discrete Time Stochastic Modelling in R
Grey-box modelling with ctsm-r

The model class

ctsmr implements a state space model with:

Continuous time stochastic differential system equations (SDE)

\[ dX_t = f(X_t, U_t, t, \theta)dt + g(X_t, U_t, t, \theta)dB_t \]

Discrete time measurement equations

\[ Y_{t_n} = h(X_{t_n}) + e_{t_n}, \quad e_{t_n} \in N(0, S(u_n, t_n, \theta)) \]

- Underlying physics (system, states) modelled using continuous SDEs.
- Some (or all) states are observed in discrete time.
Features in CTSM-R

- Automatic classification (LTI or NL)
- Symbolic differentiation replaced AD (NL only)
  (Jacobians are computed faster.)
- Finite difference approximation of gradients are computed in parallel.
- Scriptable! Run multiple model during the night. Possible to use compute cluster.
- Direct access to plotting facilities from the R framework.
Loading the library

The R package is called \texttt{ctsmr}

\texttt{R code}

\texttt{library(ctsmr)}
Loading the library

The R package is called **ctsmr**

**R code**

```r
library(ctsmr)
```

The model class is called **ctsm** - *Continuous Time Stochastic Model*.

**R code**

```r
MyModel <- ctsm$new()
```
Class.. huh?

- **ctsm** is a ReferenceClass.
- The functions are methods attached to the class.
Class.. huh?

- **ctsm** is a ReferenceClass.
- The functions are methods attached to the class.

### ctsm methods

**Specifying the model:**
- `addSystem()`
- `addObs()`
- `setVariance()`
- `addInput()`

**Estimate and prediction:**
- `setParameter()`
- `setOptions()`
- `estimate()`

### ctsmr defined functions

- `predict`
- `simulate`
- `filter.ctsmr`
- `smooth.ctsmr`
How to add System Equations

Use the `$addSystem` method to add a stochastic differential equation as a system equation.

**R code**

```r
MyModel$addSystem( dX ~ (mu*X-F*X/V)*dt + sig11*dw1)
MyModel$addSystem( dS ~ (-mu*X/Y+F*(SF-S)/V) * dt + sig22*dw2)
MyModel$addSystem( dV ~ F*dt + sig33*dw3 )
```

Pay attention to the `~`. Do not use `=`. The diffusion processes must be named `dw{n}`
How to add Observation Equations

Use the $addObs$ method to add a measurement/observation equation.

\[ Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} X \\ S \\ V \end{bmatrix} \]

R code

MyModel$addObs(y1 ~ X)
MyModel$addObs(y2 ~ S)
MyModel$addObs(y3 ~ V)

Pay attention to the $\sim$. Do not use $=.$
How to set the Variance structure of the Measurement Equations

The Example

Use the `$setVariance` method.

Example

```r
MyModel$setVariance(y1'y1 ~ s11)
```
How to set the Variance structure of the Measurement Equations

The Example

Use the `setVariance` method.

Example

```r
MyModel$setVariance(y1y1 ~ s11)
```

For \(y_1, y_2, y_3\) the size of the variance-covariance matrix is 3x3.

$$S = \begin{bmatrix} s_{11} & 0 & 0 \\ 0 & s_{22} & 0 \\ 0 & 0 & s_{33} \end{bmatrix}$$

R code

```r
MyModel$setVariance(y1y1 ~ s11)
MyModel$setVariance(y2 ~ s22)
MyModel$setVariance(y3^2 ~ s33)
```

Pay attention to the `~`. Do not use `=`.
Which variables are inputs?

Use the \$addInput method to specify which variable is an input and not a parameter.

R code

MyModel$addInput(F)
How to specify initial values, boundaries and prior standard deviance (for MAP)?

Use the \$setParameter method.

**R code**

```r
MyModel$setParameter(X = c(init=1,lb=0,ub=2), 
                      S0 = c(0.25,0,1))
MyModel$setParameter(V0 = c(1,lower=0,upperbound=2))
```

Pay attention to the \(=\). Do not use \(~\).

- Quite flexible.
- Named numbers (e.g. \(\text{init}=3\)) are processed first.
- Initial state values (e.g. \(X_0\)) can be named \(X_0\) or \(X\).
- \(\text{MyModel}\$\text{ParameterValues}\) contains the parsed values.
How to change filtering and numerical optimisation options (advanced)?

Use the `setOptions` method to change the options found in `MyModel$options`. 
Specify the data

cctsM expects a data.frame containing time and all inputs and outputs.

Example

```r
MyData <- data.frame(t = c(1,2,3), F = c(4,3,2), Y1 = c(7,6,5), Y2 = ...
```

Multiple independent datasets can be given as a list of data.frames.

Example

```r
AllMyData <- list(MyData1, MyData2, MyData2, ...
```
Estimate the parameters

To estimate the parameters run:

```r
fit <- MyModel$estimate(data = MyData)
```
Parameter inference

Like \texttt{lm()} use \texttt{summary()} on the fit for additional information.

- Parameter estimates alone:
  \texttt{fit}

- + standard deviance, t-statistics and p-values:
  \texttt{summary(fit)}

- + correlation of parameter estimates:
  \texttt{summary(fit, correlation=TRUE)}

- + additional information $\left( \frac{dF}{d\theta}, \frac{dPen}{d\theta} \right)$:
  \texttt{summary(fit, extended=TRUE)}
How to get k-step predictions

Use the predict function.

Usage

```
one.step.prediction <- predict(fit)
```

Available options:

- `n.ahead` number of steps ahead to predict.
- `newdata` to predict using a new dataset.
Diagnostics

- k-step predictions `predict`
- filtered states `filter.ctsmr`
- smoothed states `smooth.ctsmr`
- simulations `simulate`
- likelihood ratio tests
Example: Selecting a suitable grey-box model for the heat dynamics of a building
Test case: One floored 120 $m^2$ building

Objective

Find the best model describing the heat dynamics of this building
## Data

### Measurements of:

- $y_t$: Indoor air temperature
- $T_a$: Ambient temperature
- $\Phi_h$: Heat input
- $\Phi_s$: Global irradiance
Two big challenges when modelling with data

- **Model selection:** How to decide which model is most appropriate to use? We are looking for a model which gives us un-biased estimates of physical parameters of the system. This requires that the applied model is neither too simple nor too complex.

- **Model validation:** How to validate the performance of a dynamical model? We need to assess if the applied model fulfill assumptions of white-noise errors, i.e. that the errors show no lag-dependence.
Model selection

Likelihood ratio test: Test for model expansion

Say we have a model and like to find out if an expanded version will give a significantly better description of data

i.e. give an answer to: Should we use the expanded model instead of the one we have?

The likelihood ratio test

\[ \lambda(y) = \frac{L_{\text{sub}}(\hat{\theta}_{\text{mle,sub}})}{L(\hat{\theta}_{\text{mle}})} \]

can be applied to test for significant improvement of the expanded model (with maximum likelihood \( L_{\text{sub}}(\hat{\theta}_{\text{mle,sub}}) \)) over the sub-model (with maximum likelihood \( L(\hat{\theta}_{\text{mle}}) \))
Test for expansion

Simplest model

\[
\begin{array}{c}
\text{Interior} \quad T_i \\
C_i \\
\Phi_h \\
\text{Heater} \\
A_w \Phi_s \\
\text{Solar} \\
\text{Envelope} \quad R_{iw} \\
\text{Ambient} \\
+ \\
T_a \\
\end{array}
\]

First extension: building envelope part \((TiTe)\)

\[
\begin{array}{c}
\text{Interior} \quad T_i \\
C_i \\
\Phi_h \\
\text{Heater} \\
A_w \Phi_s \\
\text{Solar} \\
\text{Envelope} \quad R_{ie} \quad R_{ea} \\
\text{Ambient} \\
+ \\
T_a \\
\end{array}
\]
Test for expansion

Simplest model

First extension: indoor medium part (TiTm)
Test for expansion

Simplest model

First extension: sensor part (\(TiTs\))
Test for expansion

Simplest model

First extension: heater part \((TiTh)\)
Test for expansion

Simplest model

First extension: Which one??

$TiTe$, $TiTm$, $TiTs$, or $TiTh$?
### Log-likelihoods

<table>
<thead>
<tr>
<th>Model</th>
<th>Sub-model</th>
<th>$l(\theta; \mathcal{Y}_N)$</th>
<th>$m$</th>
<th>$T_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplest</td>
<td></td>
<td>$l(\theta; \mathcal{Y}_N)$</td>
<td>$m$</td>
<td>2482.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Expanded</td>
<td>$T_i$</td>
<td>3628.0</td>
<td>10</td>
<td>3639.4</td>
</tr>
<tr>
<td></td>
<td>$T_i T_e$</td>
<td>3884.4</td>
<td>10</td>
<td>3911.1</td>
</tr>
<tr>
<td></td>
<td>$T_i T_m$</td>
<td>3911.1</td>
<td>10</td>
<td>3911.1</td>
</tr>
<tr>
<td></td>
<td>$T_i T_s$</td>
<td>3911.1</td>
<td>10</td>
<td>3911.1</td>
</tr>
<tr>
<td></td>
<td>$T_i T_h$</td>
<td>3911.1</td>
<td>10</td>
<td>3911.1</td>
</tr>
</tbody>
</table>

### Likelihood-ratio test

<table>
<thead>
<tr>
<th>Sub-model</th>
<th>Model</th>
<th>$m - r$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_i$</td>
<td>$T_i T_h$</td>
<td>4</td>
<td>$&lt; 10^{-16}$</td>
</tr>
</tbody>
</table>
Identify the best physical model for the data

Simplest model

\[ C_i \quad \Phi_h \quad A_w \Phi_s \quad R_{ia} \quad \Theta_a \]

\[ T_i \quad T_h \quad T_s \quad T_a \]
Identify the best physical model for the data

Simplest model

Most complex model applied
Identify the best physical model for the data

Simplest model

The best model for the given data is probably in between

Most complex model applied
<table>
<thead>
<tr>
<th>Iteration</th>
<th>Start $I(\theta; \mathcal{Y}_N)$</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$2482.6$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$m = 6$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$Ti$</td>
<td>$TiTe$</td>
</tr>
<tr>
<td></td>
<td>$3628.0$</td>
<td>$3639.4$</td>
</tr>
<tr>
<td></td>
<td>$10$</td>
<td>$10$</td>
</tr>
<tr>
<td>2</td>
<td>$TiThTs$</td>
<td>$TiTmTh$</td>
</tr>
<tr>
<td></td>
<td>$4017.0$</td>
<td>$5513.1$</td>
</tr>
<tr>
<td></td>
<td>$14$</td>
<td>$14$</td>
</tr>
<tr>
<td>3</td>
<td>$TiTeThRia$</td>
<td>$TiTeThAe$</td>
</tr>
<tr>
<td></td>
<td>$5517.3$</td>
<td>$5520.5$</td>
</tr>
<tr>
<td></td>
<td>$15$</td>
<td>$15$</td>
</tr>
<tr>
<td>4</td>
<td>$TiTeThTsRia$</td>
<td>$TiTmTeThTs$</td>
</tr>
<tr>
<td></td>
<td>$5612.5$</td>
<td>$5612.9$</td>
</tr>
<tr>
<td></td>
<td>$19$</td>
<td>$22$</td>
</tr>
<tr>
<td>5</td>
<td>$TiTmTeThTsAe$</td>
<td>$TiTeThTsAeRia$</td>
</tr>
<tr>
<td></td>
<td>$5614.6$</td>
<td>$5614.7$</td>
</tr>
<tr>
<td></td>
<td>$23$</td>
<td>$20$</td>
</tr>
<tr>
<td>Iteration</td>
<td>Sub-model</td>
<td>Model</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>1</td>
<td>$Ti$</td>
<td>$TiTh$</td>
</tr>
<tr>
<td>2</td>
<td>$TiTh$</td>
<td>$TiTeTh$</td>
</tr>
<tr>
<td>3</td>
<td>$TiTeTh$</td>
<td>$TiTeThTs$</td>
</tr>
<tr>
<td>4</td>
<td>$TiTeThTs$</td>
<td>$TiTeThTsAe$</td>
</tr>
<tr>
<td>5</td>
<td>$TiTeThTsAe$</td>
<td>$TiTeThTsAeRia$</td>
</tr>
</tbody>
</table>
Model validation

How can the performance of a dynamical model be evaluated?

- We assume that the residuals are i.i.d and normal
- Auto-Correlation Function (ACF) and Cumulated Periodogram (CP) of the errors are the basic tools
- Time series plots of the inputs, outputs, and the errors are valuable for pointing out model deficiencies
Evaluate the simplest model

Inputs and residuals

ACF of residuals

Cumulated periodogram
Evaluate the model selected in step one
Evaluate the model selected in step two
Evaluate the model selected in step three

**Inputs and residuals**

- Time
- $T_i$, $\Phi_s$, $\Phi_h$, $\epsilon$

**ACF of residuals**

- Lag
- ACF

**Cumulated periodogram**

- Frequency
- Cumulated periodogram

DTU Compute
Evaluate the selected model in step four
Selected model
Selected model

Estimated parameters

\[
\begin{align*}
\hat{C}_i &= 0.0928 \quad \text{(kWh/°C)} \\
\hat{C}_e &= 3.32 \quad - \\
\hat{C}_h &= 0.889 \quad - \\
\hat{C}_s &= 0.0549 \quad - \\
\hat{R}_{ie} &= 0.897 \quad (\text{circC/kW}) \\
\hat{R}_{ea} &= 4.38 \\
\hat{R}_{ih} &= 0.146 \quad - \\
\hat{R}_{is} &= 1.89 \quad - \\
\hat{A}_w &= 5.75 \quad (\text{m}^2) \\
\hat{A}_e &= 3.87 \quad - \\
\end{align*}
\]

Estimated time constants

\[
\begin{align*}
\hat{\tau}_1 &= 0.0102 \quad \text{hours} \\
\hat{\tau}_2 &= 0.105 \quad - \\
\hat{\tau}_3 &= 0.788 \quad - \\
\hat{\tau}_4 &= 19.3 \quad - \\
\end{align*}
\]
Conclusions

- Applied Grey-box modelling, where a combination of \textit{prior physical knowledge} and \textit{data-driven modelling} is utilized

- Using a forward selection procedure with likelihood-ratio tests a suitable physical model is found

- The ability of the selected models to describe the heat dynamics are evaluated with the ACF, CP, and time series plots
Identifiability
Identifiability

Model identifiability is important for estimation in general (less important for prediction, very important for parameter interpretation).

There are two aspects of identifiability:

- **Structural identifiability**: the parameters in the model can never be estimated due to the structure of the model. Depends only on the model.

- **Practical identifiability**: there is not enough information in the data available to estimate the parameters in the model. Depends both on the model and the data.
Structural identifiability

State space model (innovation form)

\[
\frac{d\hat{X}(t)}{dt} = A\hat{X}(t) + BU(t) + K\epsilon(t)
\]
\[
Y(t) = C\hat{X}(t) + DU(t) + \epsilon(t)
\]

Apply the bilateral Laplace transformation (and after some voodoo)

\[
Y(s) = C(sI - A)^{-1}BU(s) + C(sI - A)^{-1}K\epsilon(s) + DU(s) + \epsilon(s)
\]
\[
= \left( C(sI - A)^{-1}B + D \right) U(s) + \left( C(sI - A)^{-1}K + I \right) \epsilon(s)
\]

Focus on the input related transfer function

\[
H_i(s) = C(sI - A)^{-1}B + D
\] (2)
Analyse the identifiability of an SDE model of a Wall

A lumped RC model of the wall

\[
\begin{align*}
\frac{dT_w}{dt} &= \frac{1}{C_w} \left( \frac{T_a - T_w}{R_{aw}} + \frac{T_i - T_w}{R_{wi}} \right) dt + d\omega_1(t) \\
\frac{dT_i}{dt} &= \frac{1}{C_i} \left( \frac{T_w - T_i}{R_{wi}} \right) dt + d\omega_2(t) \\
y_{t_k} &= T_{t_k} + \sigma_{t_k}
\end{align*}
\]
Transfer function

Apply equation ?? to obtain the input transfer function

\[ H_{\text{input}}(s) = \frac{1}{s^2 + \frac{R_{aw} C_i + C_i R_{wi} + R_{aw} C_w}{C_i C_w R_{aw} R_{wi}} \cdot s + \frac{1}{C_i C_w R_{aw} R_{wi}}} \]
Transfer function

Apply equation ?? to obtain the input transfer function

\[ H_{\text{input}}(s) = \frac{1}{s^2 + \frac{R_{\text{aw}} C_i + C_i R_{\text{wi}} + R_{\text{aw}} C_w}{C_i C_w R_{\text{aw}} R_{\text{wi}}} \cdot s + \frac{1}{C_i C_w R_{\text{aw}} R_{\text{wi}}} \]  

And compare it to

\[ H(s) = \frac{b_0}{s^2 + a_1 \cdot s + a_0} \]
Transfer function

Apply equation ?? to obtain the input transfer function

\[ H_{\text{input}}(s) = \frac{1}{s^2 + \frac{R_{aw} C_i C_w R_{aw} R_{wi}}{C_i C_w R_{aw} R_{wi}} \cdot s + \frac{1}{C_i C_w R_{aw} R_{wi}}} \]

And compare it to

\[ H(s) = \frac{b_0}{s^2 + a_1 \cdot s + a_0} \]

Only two independent equations

\[ a_0 = \frac{1}{C_i C_w R_{aw} R_{wi}} \]
\[ a_1 = \frac{R_{aw} C_i + C_i R_{wi} + R_{aw} C_w}{C_i C_w R_{aw} R_{wi}} \]
Fit all four parameters?

Solve two equations for four parameters.

\[
\begin{align*}
Ci &= Ci \\
Rwi &= Rwi \\
Cw &= -\frac{Ci}{C_i^2 R_{wi}^2 a_0 - a_1 C_i R_{wi} + 1} \\
R_{aw} &= -\frac{C_i^2 R_{wi}^2 a_0 - a_1 C_i R_{wi} + 1}{C_i^2 R_{wi} a_0}
\end{align*}
\]

Note: \( a_0 \) and \( a_1 \) are known when simulating data.
$C_w$ is a function of other parameters

Below is the feasible $C_w$ parameters: $C_w > 0$
Estimate two parameters

We can estimate two. So try fixing $R_{wi}$ and $R_{aw}$
Estimate two parameters

We can estimate two... So try fixing $R_{wi}$ and $R_{aw}$
Estimate two parameters

We can estimate two... So try fixing $C_w$ and $R_{aw}$
Estimate two parameters

We can estimate two... So try fixing $C_w$ and $R_{wi}$.
Estimate two parameters

We can estimate two.. So try fixing $R_{wi}$ and $C_i$
Estimate two parameters

We can estimate two. So try fixing $R_{aw}$ and $C_i$
Estimate two parameters

We can estimate two... So try fixing $C_i$ and $C_w$