Data-driven modelling in a nut shell

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 2023:

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

What is the meaning of all these different terms!

- **Statistics** is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data.
- **Machine learning** is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.

What is the meaning of all these different terms!

- **Statistics** is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data.
- **Machine learning** is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.
- **Computational statistics or statistical computing** is the interface between statistics and computer science.

What is the meaning of all these different terms!

- **Statistics** is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data.
- **Machine learning** is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.
- **Computational statistics or statistical computing** is the interface between statistics and computer science.
- Statistical learning, data mining, intelligent systems, predictive analysis, ...

What is the meaning of all these different terms!

- **Statistics** is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data.
- **Machine learning** is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.
- **Computational statistics or statistical computing** is the interface between statistics and computer science.
- Statistical learning, data mining, intelligent systems, predictive analysis, ...

For sure it's all AI ;-)

What is the meaning of all these different terms!

- **Statistics** is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data.
- **Machine learning** is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.
- **Computational statistics or statistical computing** is the interface between statistics and computer science.
- Statistical learning, data mining, intelligent systems, predictive analysis, ...

For sure it's all AI ;-)

- *Maybe what we do in the course:* **"data-driven modelling for energy engineering"**

Which modelling technique?

Depends on the application

- What is the objective:
  - learning about a phenomena, e.g. performance estimation
  - predictions
  - control
- Model structure from prior knowledge?
- Which data is at hand!

Which modelling technique?

Depends on the application

- What is the objective:
  - learning about a phenomena, e.g. performance estimation
  - predictions
  - control
- Model structure from prior knowledge?
- Which data is at hand!

Regression techniques

Linear regression, generalized linear model, mixed effects, maximum likelihood, robust estimation, ridge regression, lasso regression, kernels, spline basis, regression trees, neural networks, support vector machines, ...

Which modelling technique?

Depends on the application

- What is the objective:
  - learning about a phenomena, e.g. performance estimation
  - predictions
  - control
- Model structure from prior knowledge?
- Which data is at hand!

Regression techniques

Linear regression, generalized linear model, mixed effects, maximum likelihood, robust estimation, ridge regression, lasso regression, kernels, spline basis, regression trees, neural networks, support vector machines, ...

Statistical models for dynamical systems

ARMAX, SDE, grey-box: basically some Kalman filter is behind!

Which modelling technique?

Depends on the application

- What is the objective:
    - learning about a phenomena, e.g. performance estimation
    - predictions
    - control
- Model structure from prior knowledge?
- Which data is at hand!

Regression techniques

Linear regression, generalized linear model, mixed effects, maximum likelihood, robust estimation, ridge regression, lasso regression, kernels, spline basis, regression trees, neural networks, support vector machines, ...

Statistical models for dynamical systems

ARMAX, SDE, grey-box: basically some Kalman filter is behind!

Each technique have different pros and cons. Depends on the application, data, ...

Plot your data!

Plot, plot, plot

- Get insights by plotting, it's really the best way to understand your data...of course everything cannot be seen, but be creative!
- Plot all data and plot interesting details

Plots and base R functions

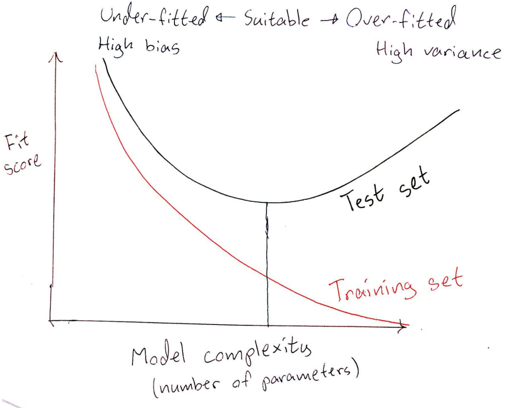
- Scatter plots (`plot`, `pairs`)
- Histograms (`hist`) and box-plots (`boxplot`)
- Time series plots (`ts.plot`), ACF (`acf`), CCF (`ccf`)

Model selection and the bias-variance tradeoff

We want to find a *suitable model* – the model which is neither too simple (under-fitted) nor too complex (over-fitted):

- Divide the data into a training set and a test set
- Define a fit score (smaller is better e.g. summed squared error)

How do we find the balance?



Model selection

How do we find the balance?

Model selection

How do we find the balance?

- Information criteria (AIC, BIC)
- Goodness of fit test (likelihood-ratio test, F-test)
- Cross-validation technique (n-fold CV)

Model selection

How do we find the balance?

- Information criteria (AIC, BIC)
- Goodness of fit test (likelihood-ratio test, F-test)
- Cross-validation technique (n-fold CV)

Model selection procedure

- Forward selection (start with a small model and extend)
- Backward selection (start with a large model and remove)

Model validation (use also while finding a model)

After fitting the model then analyse the residuals

If no patterns are left, hence we have white noise residual, then we are done!

Residuals analysis

Plot, plot, plot!

- Time series plots of residuals aligned with input series
- Scatter plots of residuals vs. inputs
- ACF and CCF

Model validation (use also while finding a model)

After fitting the model then analyse the residuals

If no patterns are left, hence we have white noise residual, then we are done!

Residuals analysis

Plot, plot, plot!

- Time series plots of residuals aligned with input series
- Scatter plots of residuals vs. inputs
- ACF and CCF

Forward selection

Fit a simple model, analyse the residuals: Can you see some patterns left related to some inputs? Improve the model and repeat...

- Good approach for modelling with new data
- Good approach for articles (you get a story)

Great, that sounds easy!

Well, with experience it helps, but it's never trivial!!

I think it can be challenging because of:

- Know the techniques
- Many choices to make!
- When is the model good enough!?
- We tend to make too complex models
- More like missing data, poor resolution, ...

Great, that sounds easy!

Well, with experience it helps, but it's never trivial!!

I think it can be challenging because of:

- Know the techniques
- Many choices to make!
- When is the model good enough!?
- We tend to make too complex models
- More like missing data, poor resolution, ...

Let's help each other!

Quick tour around in RStudio and R

Open RStudio

R basics

The workflow in R is to apply functions on data variables

```
## A data.frame has attributes in addition to its value
X <- data.frame(col1=1:10, col2=10:1)
X

## The variables has attributes (X is basically a list)
attributes(X)

## You can add attributes
attr(X, "element1") <- "value1"
attributes(X)

## Change the column names
names(X) <- c("a", "b")
X

## Check its class
class(X)

## Define a function (sum squared)
ss <- function(x){ sum(x^2) }

## Apply the function on each row
apply(X, 1, ss)
## or column
apply(X, 2, ss)
```

R basics

The list class is good (everything actually either a list or function in R)

```
L <- list()

## Put in everything you like
L$X <- X
L$text <- "Hello"
L$value <- 3
L

## Apply a function on each element
lapply(L, class)

## Load a package (more on R packages later)
require(parallel)
## Do this using all the cores of the processor
mclapply(L, class)
```

R packages (more than 10000 on cran!)

Popular packages:

- Great plotting: `ggplot2` and `plotly`
- Working with really large datasets: `data.table`
- Making browser apps: `shiny`
- Integrate c++ code easily: `Rcpp`
- Use Python in R or R in Python: Several R and Python packages

We will use different packages in the exercises

```
install.packages("Rcpp")
```

Search in R packages `https://www.r-pkg.org/`, try “splines” and “kernels”

R nice to know and tips

- Matlab - R reference guide
  `https://cran.r-project.org/doc/contrib/Hiebeler-matlabR.pdf`
- Debugging:
  - `browser()` and `traceback()`
  - Easy in RStudio to set breakpoint

R nice to know and tips

Speed issues

- For-loops can be really slow
- Testing speed:

```
## install.packages("microbenchmark")
library(microbenchmark)
x <- numeric(100)
microbenchmark(
    for(i in 1:100){x[i] <- i},
    sapply(1:100, function(i){ return(i) }),
    times=100
)
```

- If possible use `apply` functions, if recursive write it in c++

Install the ctsmr package

Tomorrow we will use the `ctsmrTMB` package and Thursday the onlineforecast package

If you did not install the *ctsmrTMB* package, do it today to check it works:
See instructions on `https://github.com/phillipbvetter/ctsmrTMB`.

Tomorrow Phillip who is making the package will present it!