Introduction to learning, multiple and nonparametric regression

Machine Learning

Jonas Striaukas

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Course details

Basic info:

- **My email**: js.fi@cbs.dk or jonas.striaukas@gmail.com
- **Lecture time**: TBA
- **Auditorium**: TBA
- **Office hours**: TBA
- **Course website**: https://jstriaukas.github.io/ml_course

Exam:

- **Structure**: TBA
- **When**: TBA

What I expect from you:

- Understand the concepts we learn in the class. In particular derivations of some simple theoretical results as well as full understanding of more complex theory.
- Be creative and active during class presentations.
- Work hard! And try to not miss classes...

Topics

- Introduction to learning, multiple and nonparametric regression
  - BLAH BLAH
- High-dimensional linear regression
  - BLAH BLAH
- High-dimensional regression properties and generalized linear models (GAMs)
  - BLAH BLAH
- Prediction, loss functions and M-estimators
  - BLAH BLAH
- Introduction to deep learning
  - BLAH BLAH
- Introduction to causal machine learning
  - BLAH BLAH

Nowadays, Big Data are ubiquitous: from the internet, biology and medicine to government, business, economy, finance....

Some quotes:

> "*There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days*", according to Eric Schmidt, the CEO of Google,in 2010.

> "*Data are becoming the new raw material of business*", according to Craig Mundie, Senior Advisor to the CEO at Microsoft

Examples in economics and finance:

- high-frequency financial assets data (e.g., stocks, bonds, fx, derivatives, ...);
- large panels of economic data (e.g., 131 macroeconomics time series (McCracken and Ng, 2015) with FRED MD database with monthly updates);
- spatial data (e.g., state-level data in US, euro area data);
- text-based data (e.g., newspaper articles, GDELT project; EC news data).

Problems associated with Big data:

- Data are collected from various sources and populations $\Longrightarrow$ **heterogeneity**;
- typically large numbers of variables are collected $\Longrightarrow$ some variables are **heavy-tailed**, i.e. have high kurtosis which is much higher than the normal distribution;
- incidental **endogeneity** due to high-dimensionality $\Longrightarrow$ huge impact on model selection and statistical inference (Fan and Liao, 2014);
- computation/optimization of model parameters $\Longrightarrow$ **convexity** so far is a way out to guarantee the stability of solutions with millions of parameters;
- noise accumulation and spurious correlation has a large impact on model selection $\Longrightarrow$ high-dimensional statistics/econometric methods as a solution to some of the problems

See Fan, Han, and Liu (2014) for an overview of how these intrinsic features of Big Data have significant impacts on the future developments of big data analysis techniques, from heterogeneity and heavy tailedness to endogeneity and measurement errors.

Learning, multiple and nonparametric regression

Spurious correlations – examples

Learning, multiple and nonparametric regression

Spurious correlations – some explanation

According to Bickel (2008), the main goals of high dimensional inferences are:

- to construct a method as effective as possible to predict future observations and;
- to gain insight into the relationship between features and responses for scientific purposes, as well as, hopefully, to construct an improved prediction method.