Introduction to learning, multiple and nonparametric regression

Machine Learning

Jonas Striaukas

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/teaching

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, active during class presentations and work hard! And try **not** to miss classes...

Machine learning, computing, etc.

"The purpose of computing is insight, not numbers."

Richard Hamming

Topics of the course

- Introduction to learning, multiple and nonparametric regression
  - Least squares estimator, nonparameteric estimation, empirical risk analysis.
- Penalized least squares
  - Penalized least squares, optimization and implementation techniques, structured nonparameteric models and structured penalization.
- Penalized least squares: properties
  - BLAH BLAH
- Prediction, loss functions and M-estimators
  - BLAH BLAH
- Introduction to deep learning
  - BLAH BLAH
- Introduction to causal machine learning
  - BLAH BLAH

Learning, multiple and nonparametric regression

Big data

Nowadays, Big Data are ubiquitous: from the internet, biology and medicine to government, business, economics, 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.
- "*Big data is not about the data*", according to Gary King of Harvard University.

Do we need ML or even AI to understand economics and/or finance data?

- **Yes!** ML is not that different from classical econometrics... "Black-box" deep learning is not that black box after all...

Learning, multiple and nonparametric regression

Big data – examples

Big data 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 FRED MD database with monthly updates, McCracken and Ng (2016));
- spatial data (e.g., state-level data in US, euro area data);
- text-based data (e.g., newspaper articles, GDELT project; EC news data);
- ... .

Learning, multiple and nonparametric regression

Impact of Big data & dimensionality

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;
- noise accumulation and spurious correlation has a large impact on model selection $\Longrightarrow$ high-dimensional statistics methods.

For curious students: see Fan, Han, and Liu (2014) for an overview of how these features impacts the developments of big data analysis techniques.

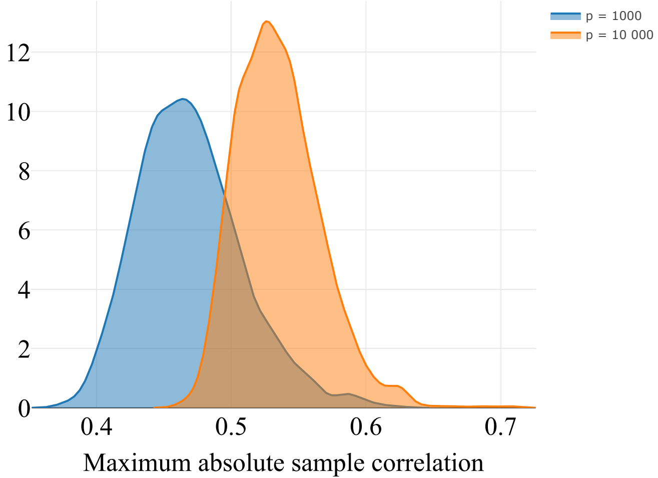
Spurious correlation refers to the observation that two variables have zero population correlation, but in the finite samples the correlation is high.

Consider the following experiment:

- $Z_j \sim N(0,1)$, $Z_j \in \mathbf{R}^n$, $j = \{1, \dots, p+1\}$;
- set $n = 50$ and $p = \{10^3, 10^4\}$;
- compute
$$\hat{r} = \max_{j \geq 2} |\widehat{\text{corr}}(Z_1, Z_j)|,$$
where $\widehat{\text{corr}}(Z_1, Z_j)$ is the sample correlation between Z_1 and Z_j;
- repeat 1000 times.

Learning, multiple and nonparametric regression

Spurious correlations – plots



The main goals of high dimensional inferences are (see Fan and Lv (2008), Bickel (2008)):

- Prediction: to construct a method as effective as possible to predict future observations and;
- (Causal) inference: to gain insight into the relationship between features and responses for scientific purposes, as well as, hopefully, to construct an improved prediction method useful for (economic) policy.

Multiple linear regression

Statistical learning theory

Consider a multiple linear regression model:

$$Y = \sum_{j \in [p]} \beta_j X_j + \varepsilon, \qquad (1)$$

where

- Y – response or dependent variable;
- X_j – variables are often called explanatory variables or covariates or independent variables;
- intercept can be included in the model by including a unit vector as one of covariates;
- β_j – regression coefficients;
- ε is the error term, some assumptions:
  - "random error" ε is often assumed has zero mean;
  - $\mathbb{E}(\varepsilon|X) = 0$ – uncorrelated with covariates X, which is referred to as *exogenous* variables (can also assume less restrictive assumptions).

Multiple linear regression

Statistical learning theory

Given observed sample $\{X_{ij}, Y_i : i \in [n], j \in [p]\}$, where $[p] \triangleq \{1, \ldots, p\}$, we have

$$Y_i = \sum_{j \in [p]} \beta_j X_{ij} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i X_i) = 0. \tag{2}$$

Classical estimator used to fit the model (dates back to Gauss and Legendre in the 19^{th} century): least squares.

Construct residuals:

$$r_i = Y_i - \sum_{j \in [p]} \beta_j X_{ij}. \tag{3}$$

Under classical assumptions, the least squares solves for $\beta = (\beta_1, \ldots, \beta_p)^\top$ by minimizing:

$$\arg\min_{\beta \in \mathbf{R}^p} \sum_{i \in [n]} r_i^2 = \arg\min_{\beta \in \mathbf{R}^p} \sum_{i \in [n]} \Big(Y_i - \sum_{j \in [p]} \beta_j X_{ij}\Big)^2,$$

$$\ell(\beta) \triangleq \sum_{i \in [n]} r_i^2 \text{ (definition for later).}$$

Multiple linear regression

Notation

Denote by:

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ – response vector;
- $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})^\top$ – covariate j vector;
- $\mathbf{X} = (\mathbf{X}_1^\top, \ldots, \mathbf{X}_p^\top)$ – covariate matrix, also known as the design matrix;
- $\beta = (\beta_1, \ldots, \beta_p)^\top$ – regression coefficient vector;
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ – regression error term vector.

Our linear model can be written in a matrix notation:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon. \tag{4}$$

We minimize:

$$\ell(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_n^2 = \langle \mathbf{Y} - \mathbf{X}\beta, \mathbf{Y} - \mathbf{X}\beta \rangle / n.$$

Multiple linear regression

Analysis

$$\ell(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_n^2 = \langle \mathbf{Y} - \mathbf{X}\beta, \mathbf{Y} - \mathbf{X}\beta \rangle / n.$$

Taking derivative of $\ell(\beta)$ w.r.t. β, equating to zero and rearranging, we obtain what is called normal equations, i.e.:

$$\mathbf{X}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{X}\beta.$$

Assume $n \leqslant p$, $\mathbf{X}^\top \mathbf{X}$ is an invertable matrix. Multiply both sides from the left by $(\mathbf{X}^\top \mathbf{X})^{-1}$, we obtain the solution for β as:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

QUESTION: What if $p \leqslant n$? Can we still write down the solution?

© Jonas Striaukas

Multiple linear regression

Analysis

$$\ell(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_n^2 = \langle \mathbf{Y} - \mathbf{X}\beta, \mathbf{Y} - \mathbf{X}\beta \rangle / n.$$

Taking derivative of $\ell(\beta)$ w.r.t. β, equating to zero and rearranging, we obtain what is called normal equations, i.e.:

$$\mathbf{X}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{X}\beta.$$

Assume $n \leqslant p$, $\mathbf{X}^\top \mathbf{X}$ is an invertable matrix. Multiply both sides from the left by $(\mathbf{X}^\top \mathbf{X})^{-1}$, we obtain the solution for β as:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

QUESTION: What if $p \leqslant n$? Can we still write down the solution?

Answer: Yes. For some conformable vector $\mathbf{w}$, we can write it down as:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-} \mathbf{X}^\top \mathbf{Y} + [I - \underbrace{(\mathbf{X}^\top \mathbf{X})^{-} \mathbf{X}^\top \mathbf{X}}_{\substack{= I \text{ in case } n > p \\ \text{and } \mathbf{X} \text{ columns are} \\ \text{linearly independent.}}}]\mathbf{w}$$

Multiple linear regression

Projection matrix

Theorem

Define $\mathbf{P} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$. *Then, for* $j \in [p]$, *we have*

$$\mathbf{P}\mathbf{X}_j = \mathbf{X}_j$$

and

$$\mathbf{P}^2 = \mathbf{P} \quad \textit{or} \quad \mathbf{P}(\mathbf{I}_n - \mathbf{P}) = 0_n.$$

That is, $\mathbf{P}$ *is a projection matrix onto the space spanned by the columns of* $\mathbf{X}$.

Proof.

First, it is easy to see that for any $\mathbf{X}$:

$$\mathbf{P}\mathbf{X} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X} = \mathbf{X}.$$

Taking $\mathbf{X} = \mathbf{P}$ proves the second equality. □

© Jonas Striaukas

Multiple linear regression

Gauss Markov theorem

Suppose we have a linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

with

- *exogeneity* — $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$;
- *homoscedasticity* — $\mathrm{Var}(\varepsilon|\mathbf{X}) = \sigma^2$.

Theorem

Under linear model, exogeneity and homoscedasticity, if follows that:

i) *unbiasedness* — $\mathbb{E}(\hat{\beta}|\mathbf{X}) = \beta$;

ii) *conditional standard errors* — $\mathrm{Var}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$;

iii) *BLUE — the least squares estimator is Best Linear Unbiased Estimator.*

Note: the linearity of least squares estimator is not strictly necessary.

© Jonas Striaukas

Multiple linear regression

Weighted least-squares

We can relax constant variance assumption, i.e., homoskedasticity. Assuming uncorrelated errors, we can allow for 'observation specific' variances:

$$\mathbf{Y} = \mathbf{X} + \varepsilon, \qquad \mathrm{Var}(\varepsilon_i | \mathbf{X}_i) = \sigma^2 v_i$$

where σ^2 remains unknown, while $v_i \in \mathbf{R}$ are *known* positive constants.

Let $Y_i^* = v_i^{-1/2} Y_i$, $X_{ij}^* = v_i^{-1/2} X_{ij}$ and $\varepsilon_i^* = v_i^{-1/2} \varepsilon_i$.
The weighted least squares estimator is then:

$$\begin{aligned}
\hat{\beta}^{\mathrm{wls}} &= \arg\min_{\beta \in \mathbf{R}^p} \sum_{i \in [n]} (Y_i^* - \sum_{j \in [p]} \beta_j X_{ij}^*)^2 \\
&= \arg\min_{\beta \in \mathbf{R}^p} \sum_{i \in [n]} v_i^{-1} (Y_i - \sum_{j \in [p]} \beta_j X_{ij})^2,
\end{aligned}$$

still *BLUE* for β.

Multiple linear regression

Weighted least-squares – example of the data

Multiple linear regression

Weighted least-squares

In general, if the error terms are correlated, we can deal with it by assuming some (known) positive definite matrix $\mathbf{W}$ such that

$$\mathbf{Y} = \mathbf{X} + \varepsilon, \qquad \mathrm{Var}(\varepsilon|\mathbf{X}) = \sigma^2\mathbf{W}.$$

Similarly, we may write

$$\mathbf{Y}^* = \mathbf{W}^{-1/2}\mathbf{Y}, \quad \mathbf{X}^* = \mathbf{W}^{-1/2}\mathbf{X}, \quad \varepsilon^* = \mathbf{W}^{-1/2}\varepsilon.$$

The residual sum of squares is then

$$\ell(\beta) = \|\mathbf{Y}^* - \mathbf{X}^*\beta\|^2 = (\mathbf{Y} - \mathbf{X}\beta)^\top\mathbf{W}^{-1}(\mathbf{Y} - \mathbf{X}\beta),$$

and the *general* least squares estimator is:

$$\hat{\beta}^{\text{gls}} = (\mathbf{X}^\top\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}^{-1}\mathbf{Y}.$$

The estimator is, again, *BLUE* for β.

© Jonas Striaukas

Multiple linear regression

Box-Cox transformations

Multiple linear regression

Summary of multiple linear regression

Multiple linear regression can only fit linear relationships — can be very restrictive in some cases!

We can build the model by applying some transformations of the original covariates, augment the model with these new/transformed covariates in the multiple linear regression. In a nonparametric model

we do not assume the functional form of the regression, i.e., we model

$$\mathbf{Y} = f(\mathbf{X}) + \varepsilon.$$

Essentially, in simple terms, we machine learning is about trying to find and fit $f(.)$ that generalizes the relationship between Y's and X's as much as possible. Thus, the model will appear quite often during the course.

BICKEL, P. J. (2008): "Discussion on the paper by Fan and Lv," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5).

FAN, J., F. HAN, AND H. LIU (2014): "Challenges of big data analysis," National science review, 1(2), 293–314.

FAN, J., AND Y. LIAO (2014): "Endogeneity in high dimensions," Annals of Statistics, 42(3), 872.

FAN, J., AND J. LV (2008): "Sure independence screening for ultrahigh dimensional feature space," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), 849–911.

MCCRACKEN, M. W., AND S. NG (2016): "FRED-MD: A monthly database for macroeconomic research," Journal of Business & Economic Statistics, 34(4), 574–589.