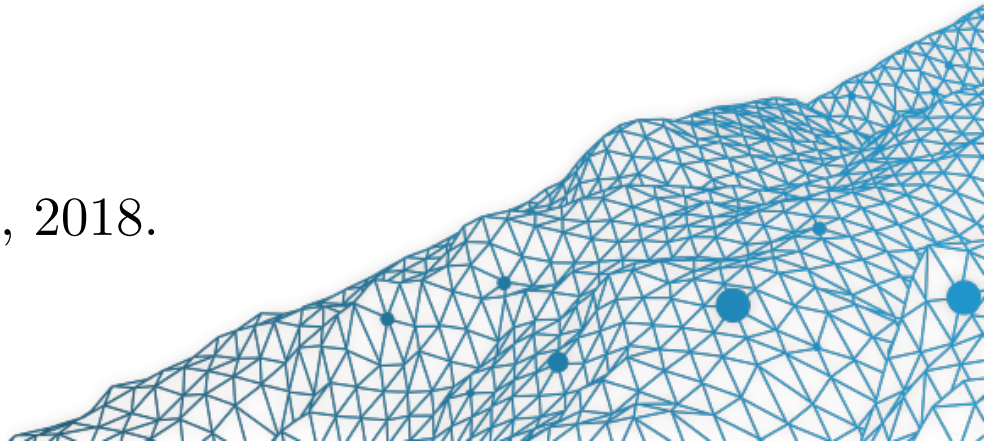
Big Data for Economics # 1

A. Charpentier (Université de Rennes 1)

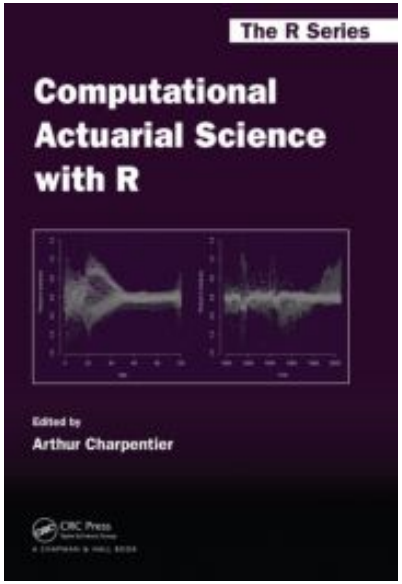
https://github.com/freakonometrics/ub

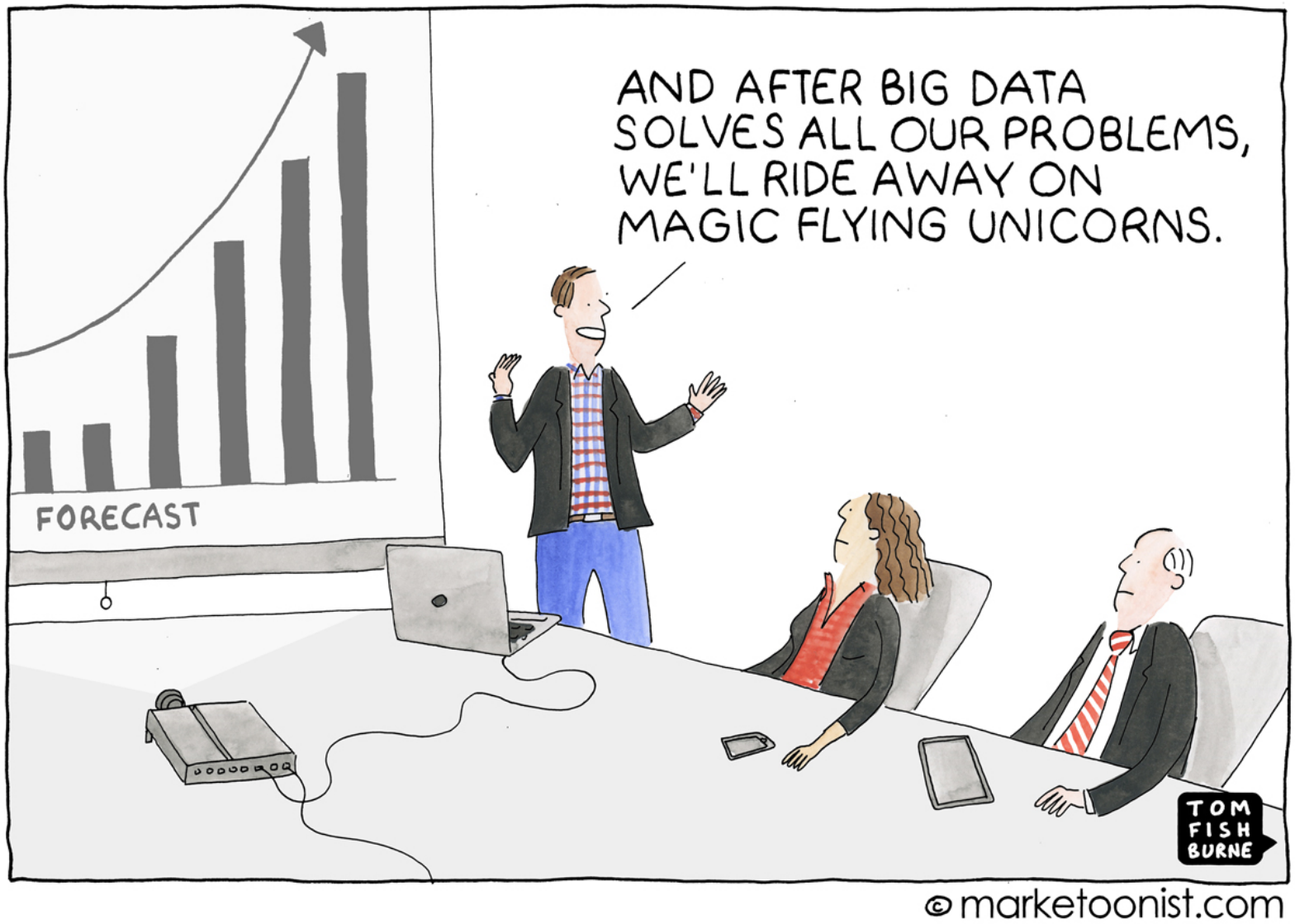
UB School of Economics Summer School, 2018.



A. Charpentier (Université de Rennes 1)

Professor Economics Department, Université de Rennes 1
Director Data Science for Actuaries Program, Institute of Actuaries
(previously Actuarial Sciences, UQàM & ENSAE Paristech
actuary in Hong Kong, IT & Stats FFA)
PhD in Statistics (KU Leuven), Fellow of the Institute of Actuaries
MSc in Financial Mathematics (Paris Dauphine) & ENSAE
Research Chair :
ACTINFO (valorisation et nouveaux usages actuariels de l'information)
Editor of the freakonometrics.hypotheses.org's blog
Editor of Computational Actuarial Science, CRC
Author of Mathématiques de l'Assurance Non-Vie (2 vol.), Economica




AND AFTER BIG DATA SOLVES ALL OUR PROBLEMS, WE'LL RIDE AWAY ON MAGIC FLYING UNICORNS.
FORECAST
TOM FISH BURNE
© marketoonist.com

Agenda

Lecture 1 Introduction : Why Big Data brings New Questions ?

Lecture 2 Simulation Based Techniques & Bootstrap

Lecture 3 Loss Functions : from OLS to Quantile Regression

Lecture 4 Nonlinearities and Discontinuities

Lecture 5 Cross-Validation and Out-of-Sample diagnosis

Lecture 6 Variable and model selection

Lecture 7 New Tools for Classification Problems

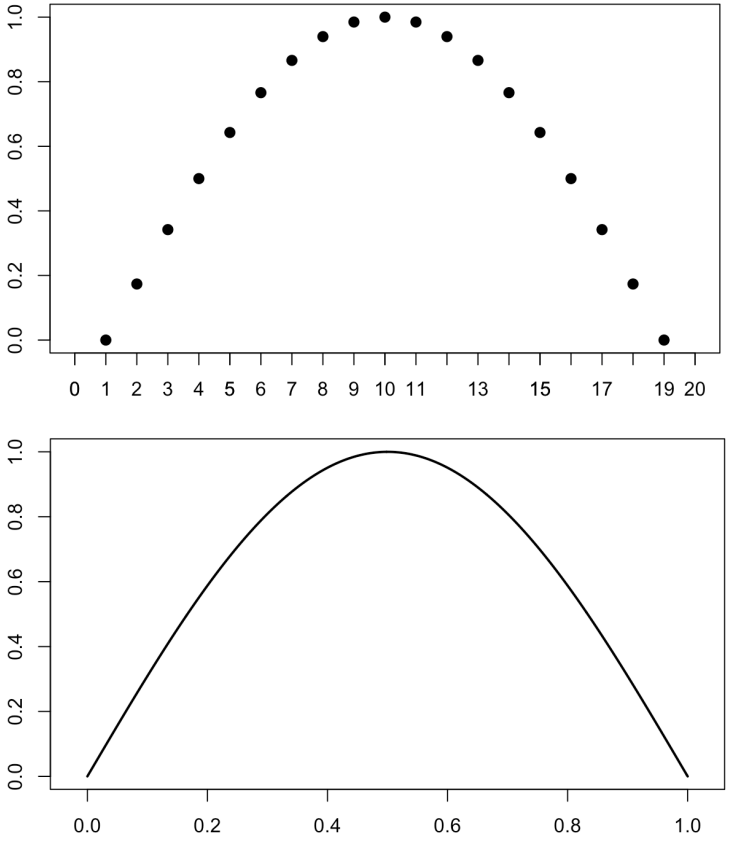
Lecture 8 New Tools for Time Series & Forecasting

#1 Introduction : Why Big Data brings New Questions ?

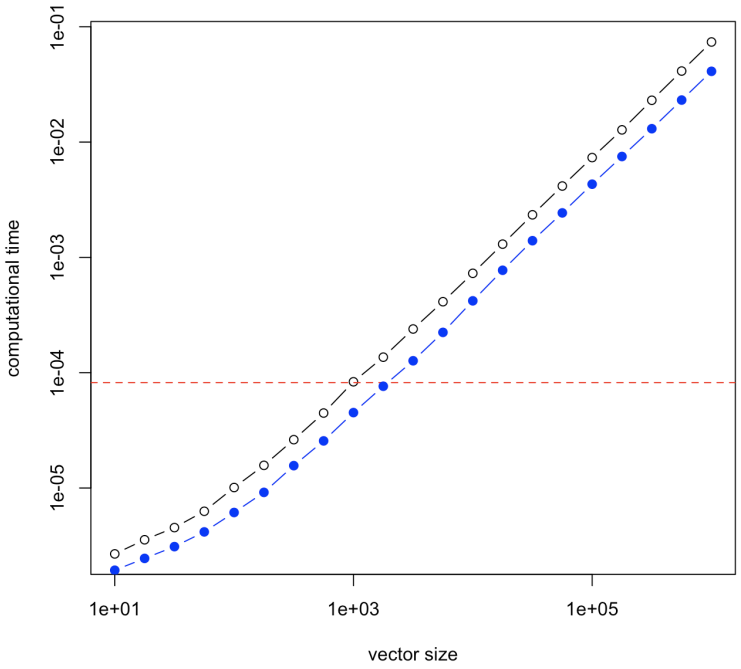
Small versus Big Data ? Discrete versus Continous Models

Consider a sequence $\boldsymbol{x} = \{x_1, \cdots, x_n\}$.
To find the maximum, use `max()`
with computational time in $O(n)$.

Consider a function $t \mapsto f(t)$
To find the maximum, use `optim()`
If $t^\star = \underset{t \in [0,1]}{\text{argmax}}\{f(t)\}$
First order condition is $\nabla f(t^\star) = 0$.



Small versus Big Data ? Discrete versus Continous Models



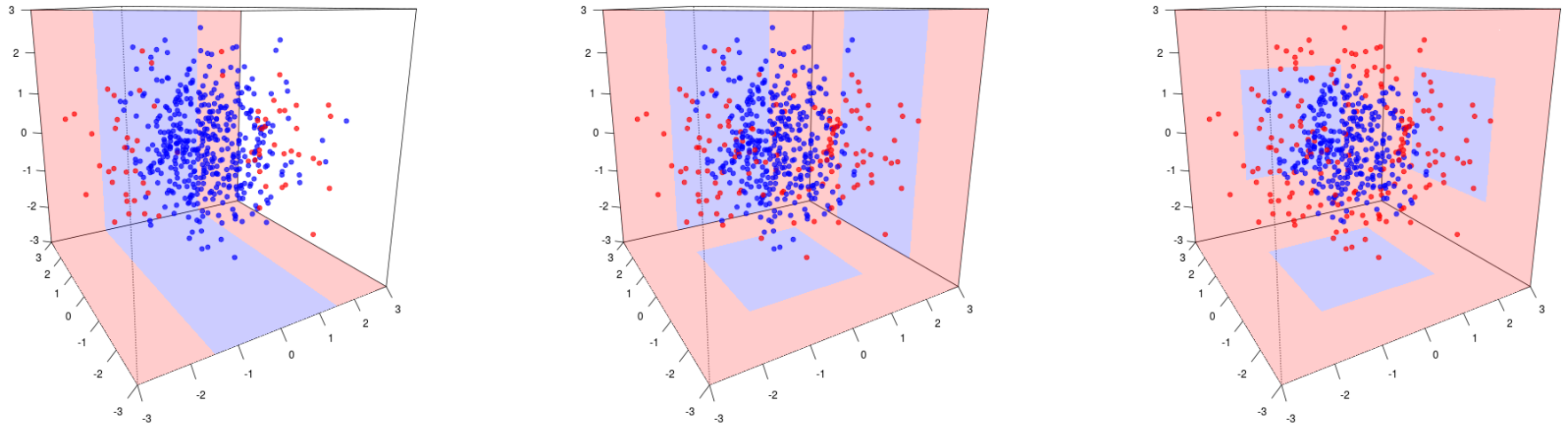
Comparing computational time of `max()`
some home made function,
and the continuous version (horizontal line)
with `optim()`

Here when $n = 1,000$ continuous time is more effective...

(see high frequency models in finance)

Curse of Dimensionality

The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality.



see Rose (2016, The End of Average: How We Succeed in a World That Values Sameness)

Big Data / New Data ? Text Based Data

- text analytics, web crawling and graph mining

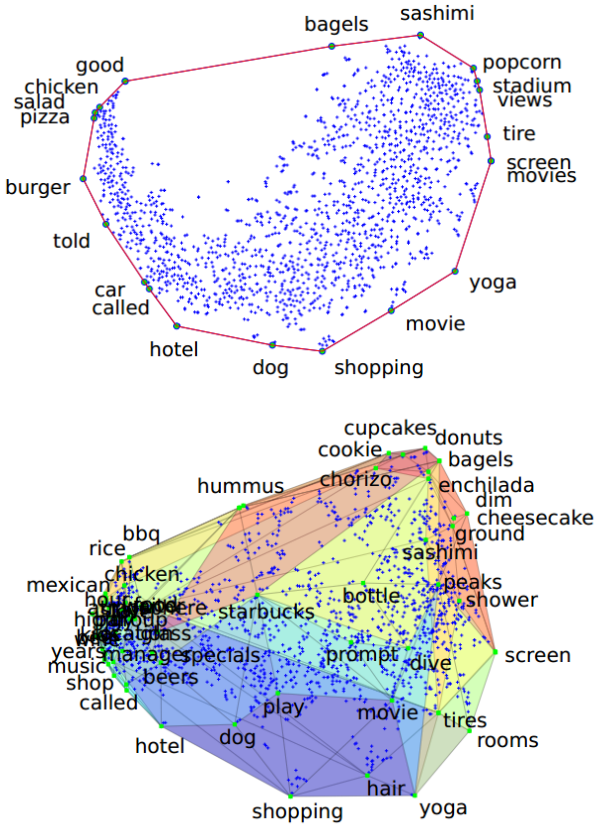
e.g. yelp.com review corpus

(see Lee & Mimmo, 2014 Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference)

index i is a review

variable x_k indicates if review contains kth word

(e.g. yoga, dog or bbq)



Big Data / New Data ? Text Based Data

- co-clustering and text mining

Simultaneous clustering of rows and columns in a matrix

Seminar *noun* 'sem.I.na:[r]

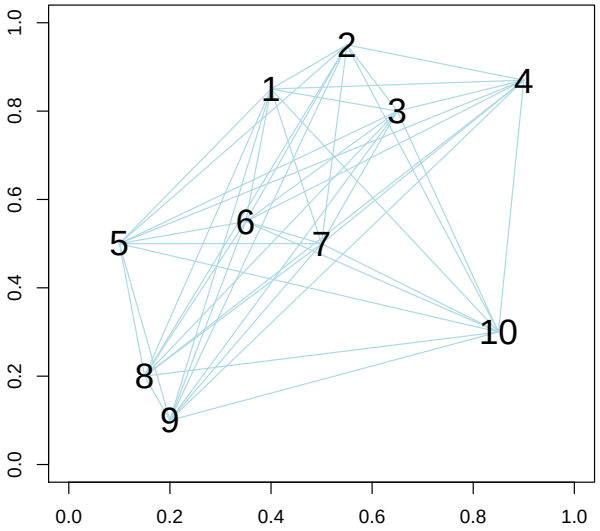
an occasion when a teacher or expert and a **group** of people meet to study and discuss something

via dictionary.cambridge.org

a small **group** of students, as in a university, engaged in advanced study and original research under a member of the faculty and meeting regularly to exchange information and hold discussions

via dictionary.reference.com

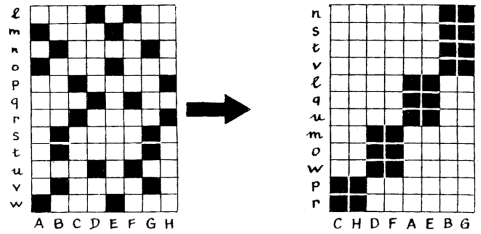
Unsupervised Techniques (Clusters)



Distance matrix $D_{i,j} = D(\boldsymbol{x}_{c_i}, \boldsymbol{x}_{c_j})$

the distance is between clusters, not (only) individuals, e.g.

$$D(\boldsymbol{x}_{c_1}, \boldsymbol{x}_{c_2}) = \begin{cases} \min\limits_{i \in c_1, j \in c_2} \{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\} \\ d(\overline{\boldsymbol{x}}_{c_1}, \overline{\boldsymbol{x}}_{c_2}) \\ \max\limits_{i \in c_1, j \in c_2} \{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\} \end{cases}$$

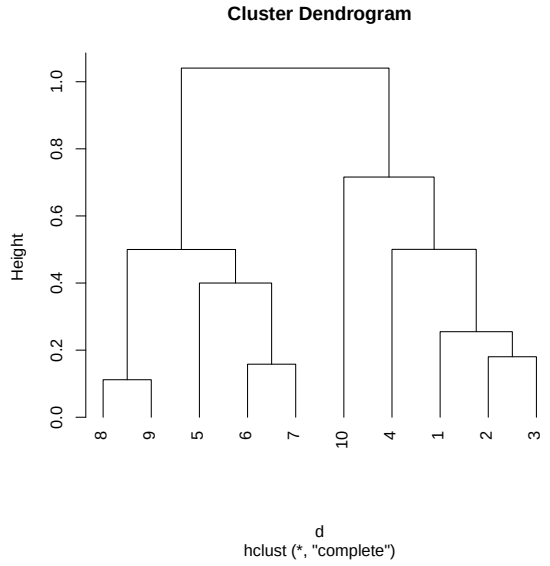


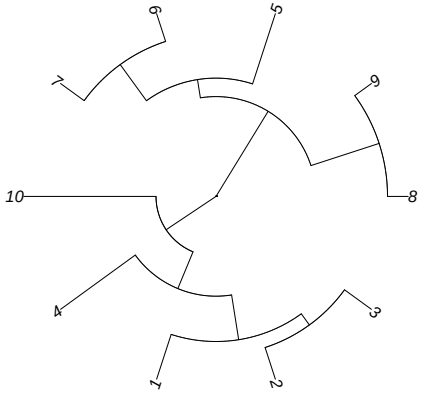
for some (standard) distance d, e.g. Euclidean (ℓ_2), Manhattan (ℓ_1), Jaccard, etc. See also Bertin (1967, Sémiologie graphique).

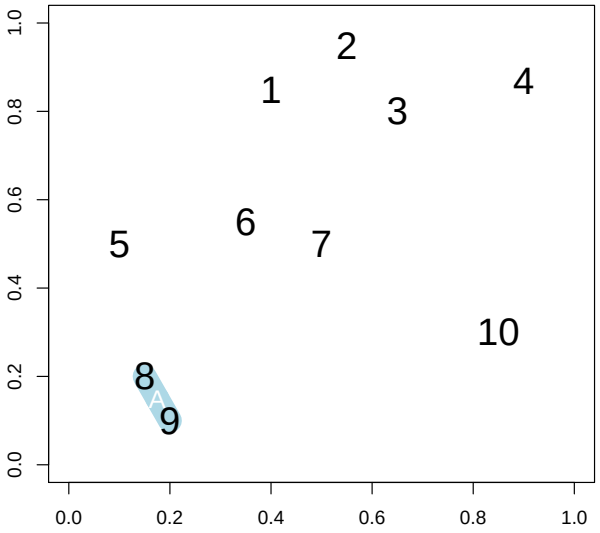
Unsupervised Techniques (Clusters)

Distance matrix $D_{i,j} = D(\boldsymbol{x}_{c_i}, \boldsymbol{x}_{c_j})$

The standard output is usually a dendrogram.



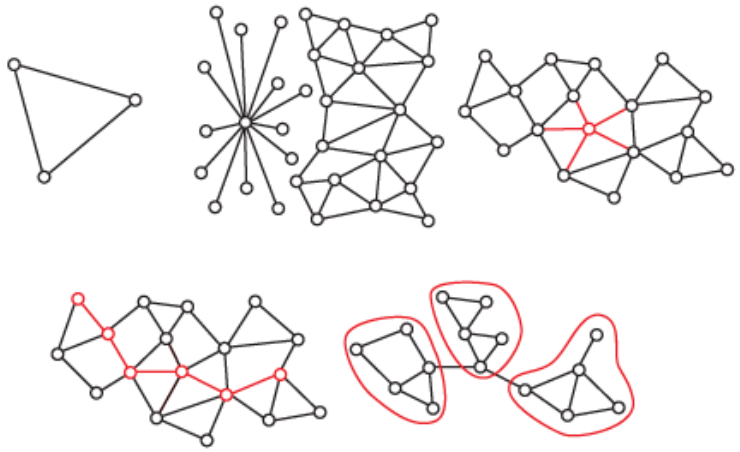




Big Data / New Data ? Network Data

Classical (individual based) data in econonmetrics $(y_i, \boldsymbol{x}_i)$, $(y_j, \boldsymbol{x}_j)$, etc supposed to be independent

Individuals are nodes of a network v_i, v_j, etc, that can be connected $e_{i,j} = 1$ or not $e_{i,j} = 0$.



See Easley, D. & Kleinberg, J. (2010) Networks, Crowds, and Markets: Reasoning About a Highly Connected World Cambridge University Press, Jackson, M. (2008). Social and Economic Networks, Wasserman, S. & Faust, K. (1994) Social Network Analysis : Methods and Applications, Christakis & Fowler (2009, Connected : The Surprising Power of Our Social Networks and How They Shape Our Lives) or Can & Alatas (2017, Big Social Network Data and Sustainable Economic Development)

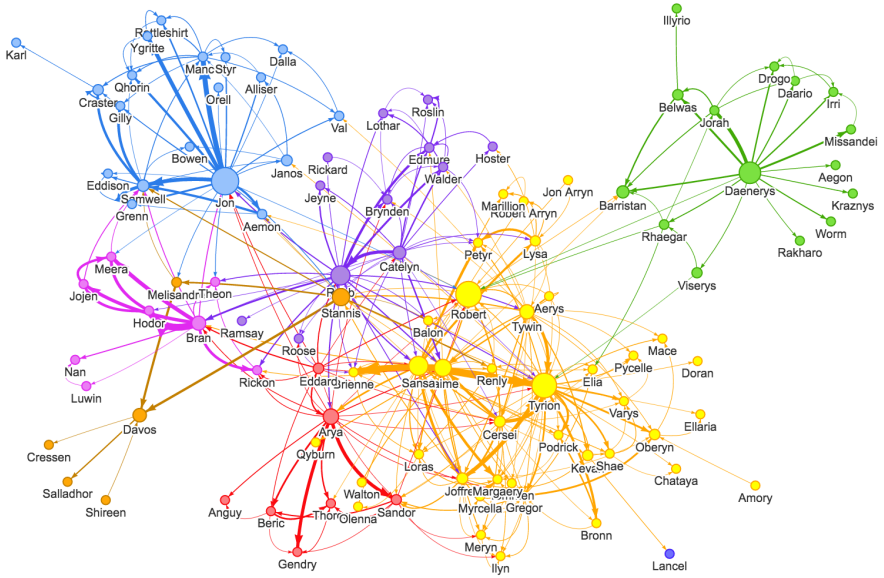
Working with networks can be complicated, and conter-intuitive

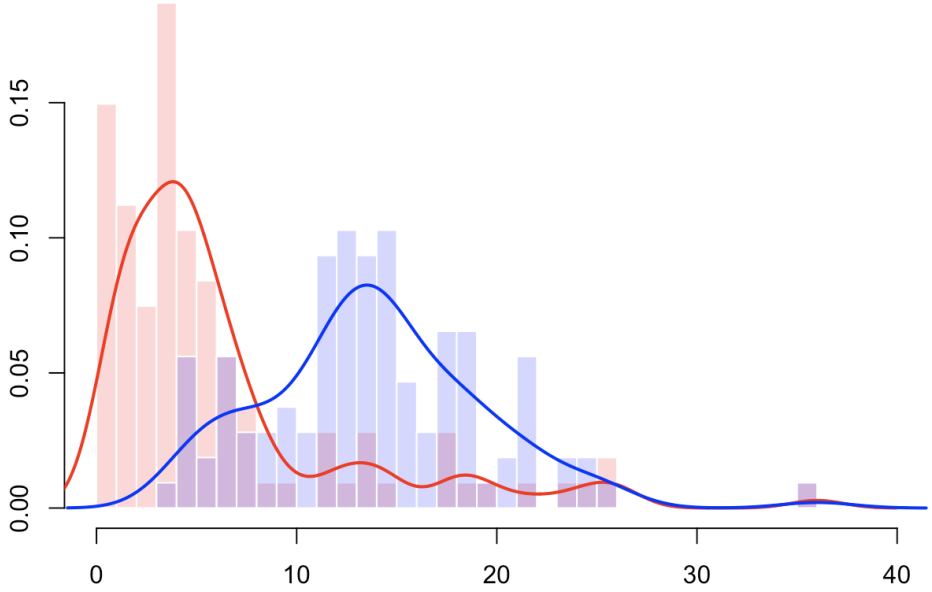
Big Data / New Data ? Network Data

- Friendship paradox

People on average have fewer friends than their friends (popular people are over-represented in the views of others)

See Game of Thrones network





Big Data / New Data ? Network Data

Consider a vertex $v \in V$, in the undirected graph $G = (V, E)$, and let $d(v)$ denote the number of edges touching it (i.e. v has $d(v)$ friends).

The average number of friends of a random person in the graph is

$$\mu = \frac{1}{n_V} \sum_{v \in V} d(v) = \frac{2n_E}{n_V}$$

The average number of friends that a typical friend has is

$$\frac{1}{n_V} \sum_{v \in V} \left(\frac{1}{d(v)} \sum_{v' \in E_v} d(v') \right)$$

But

$$\sum_{v \in V} \left(\frac{1}{d(v)} \sum_{v' \in E_v} d(v') \right) = \sum_{v, v' \in G} \left(\frac{d(v')}{d(v)} + \frac{d(v)}{d(v')} \right)$$

Big Data / New Data ? Network Data

$$= \sum_{v,v' \in G} \left(\frac{d(v')^2 + d(v)^2}{d(v)d(v')} \right) = \sum_{v,v' \in G} \left(\frac{(d(v') - d(v))^2}{d(v)d(v')} + 2 \right) > \sum_{v,v' \in G} (2) = \sum_{v \in V} d(v)$$

Thus,

$$\frac{1}{n_V} \sum_{v \in V} \left(\frac{1}{d(v)} \sum_{v' \in E_v} d(v') \right) > \frac{1}{n_V} \sum_{v \in V} d(v)$$

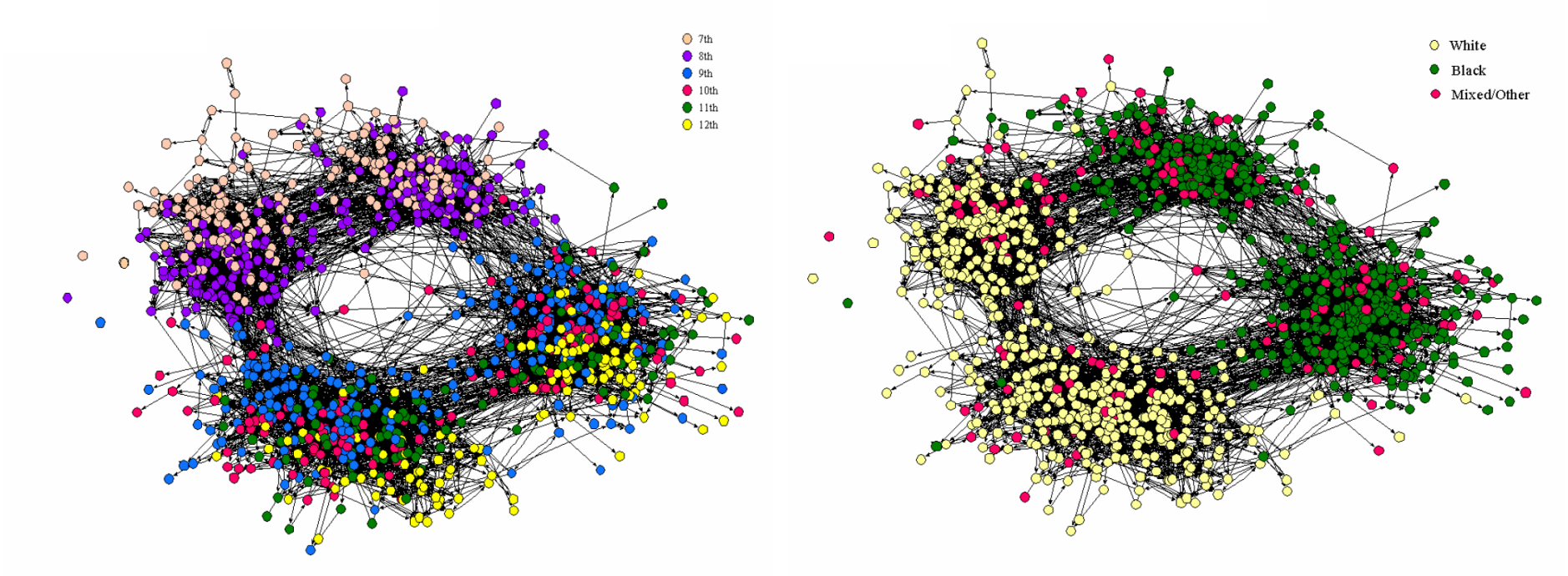
Remark This can be related to the variance decomposition $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ thus

$$\frac{\mathbb{E}[X^2]}{\mathbb{E}[X]} = \mathbb{E}[X] + \frac{\text{Var}[X]}{\mathbb{E}[X]} > \mathbb{E}[X]$$

(Jensen inequality). See Feld (1991) and Zuckerman & Jost (2001)

Big Data / New Data ? Network Data

- Homophily, "*birds of a feather flock together*"



from Moody (2001) Race, School Integration and Friendship Segregation in America

Big Data / New Data ? Network Data

- Peer effect

see Angrist (2014, The perils of peer effects)

Table 1. Students' accuracy in perceiving the drinking norm at their school (comparing actual with perceived number of alcoholic drinks consumed the last time students "partied"/socialized)

Actual school norm (median drinks)	Accuracy of perceived drinking norm: Underestimate of ≥3 drinks (%)	Underestimate of 1-2 drinks (%)	Accurate estimate (%)	Overestimate of 1-2 drinks (%)	Overestimate of ≥3 drinks (%)	Total	*N* of respondents	*N* of schools
0	NA	NA	20.6	19.5	59.9	=100%	1,891	4
1	NA	10.5	3.8	28.5	57.2	=100%	2,526	6
2	NA	7.5	8.1	30.6	53.7	=100%	8,345	14
3	3.8	6.4	13.5	37.5	38.7	=100%	18,859	35
4	3.1	12.3	12.6	37.0	34.9	=100%	20,353	38
5	4.3	15.8	20.6	24.1	35.3	=100%	11,481	20
6	6.9	23.2	15.0	23.5	31.5	=100%	8,912	12
7	5.7	23.3	9.7	23.6	37.8	=100%	352	1
Total schools	3.4	11.8	13.8	31.9	39.1	=100%	72,719	130

Source : Perkins, Haines & Rice (2005, Misperceiving the college drinking norm and related problems)

Big Data / New Data ? Tensor Data

In the (standard) linear regression,

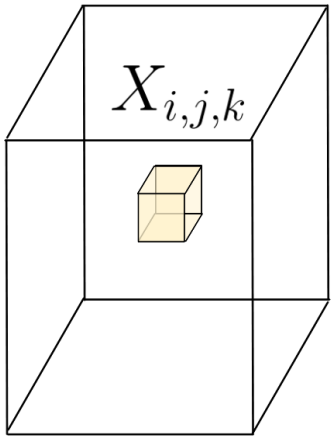
$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,k} \\ \vdots & & \vdots \\ X_{n,1} & \cdots & X_{n,k} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\boldsymbol{Y}$	$\boldsymbol{X}$	$\boldsymbol{\beta}$	$\boldsymbol{\varepsilon}$
$n \times 1$	$n \times k$	$k \times 1$	$n \times 1$

But there are many other application, e.g. in medical imaging

- electroencephalography (EEG, 2D matrix)
- anatomical magnetic resonance images (MRI, 3D array)
- functional magnetic resonance images (fMRI, 4D array)

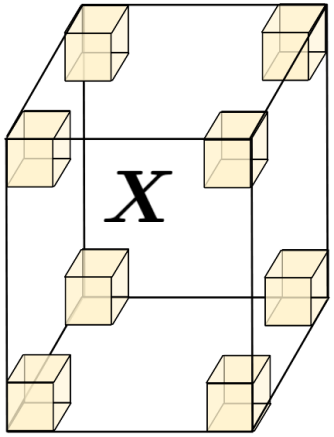
Big Data / New Data ? Tensor Data cannot turn an array into vectors: on MRI data, on a 128 resolution, means 2,097,152 regression parameters.



Consider a d-dimensional tensor parameter, $\boldsymbol{X} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$

$$\boldsymbol{Y} = \langle \boldsymbol{B}, \boldsymbol{X} \rangle + \boldsymbol{\varepsilon}$$

Problem of dimension: can be avoided with Tucker decomposition.



For a rank-one tensor,

$$\boldsymbol{X} = [X_{i,j,k}] = \boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{c} = [a_i b_j c_k]$$

Big Data / New Data ? Tensor Data

but one might consider the decomposition of rank L,

$$\boldsymbol{X} \sim \sum_{\ell=1}^{L} \lambda_\ell \boldsymbol{a}_\ell \circ \boldsymbol{b}_\ell \circ \boldsymbol{c}_\ell = [\boldsymbol{\lambda} \oplus \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]$$

where $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are the factor matrices

$$\boldsymbol{A} = [a_{i,\ell}] \ \boldsymbol{B} = [b_{j,\ell}] \text{ and } \boldsymbol{C} = [c_{k,\ell}]$$

so that

$$X_{i,j,k} = \sum_{\ell=1}^{L} \lambda_\ell a_{i,\ell} b_{j,\ell} c_{k,\ell}$$

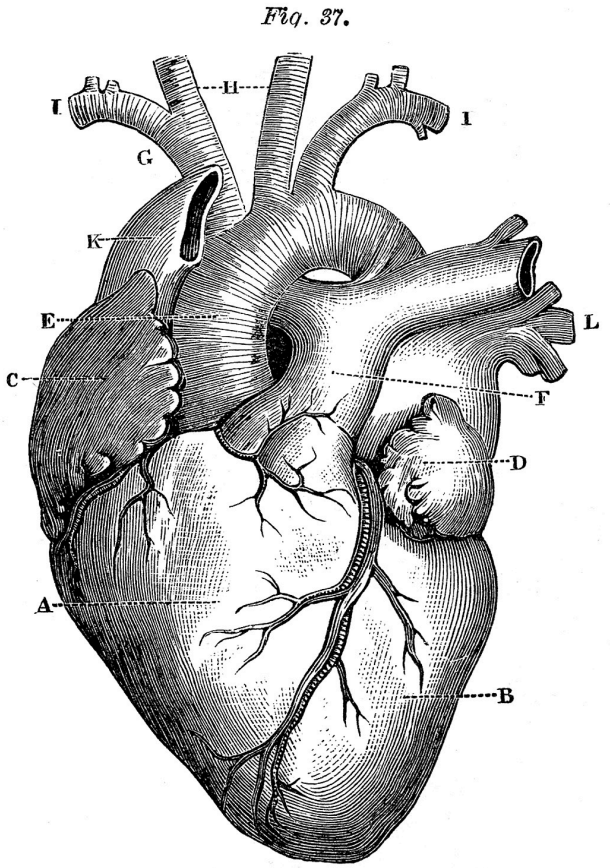
This is the so-called Tucker-decomposition.

```
library(rTensor)
?tucker
```

Toy Datasets for the Course

Myocardial infarction of patients admitted in E.R.

- heart rate (FRCAR),
- heart index (INCAR)
- stroke index (INSYS)
- diastolic pressure (PRDIA)
- pulmonary arterial pressure (PAPUL)
- ventricular pressure (PVENT)
- lung resistance (REPUL)
- death or survival




```
> myocarde = read.table("http://freakonometrics.free.fr/myocarde.csv"
    ,head=TRUE, sep=";")
```

Toy Datasets for the Course

```
> myocarde = read.table("http://freakonometrics.free.fr/myocarde.csv"
    ,head=TRUE, sep=";")
> head(myocarde)
  FRCAR INCAR INSYS PRDIA PAPUL PVENT REPUL   PRONO
1    90  1.71  19.0    16  19.5  16.0   912 SURVIE
2    90  1.68  18.7    24  31.0  14.0  1476  DECES
3   120  1.40  11.7    23  29.0   8.0  1657  DECES
4    82  1.79  21.8    14  17.5  10.0   782 SURVIE
5    80  1.58  19.7    21  28.0  18.5  1418  DECES
6    80  1.13  14.1    18  23.5   9.0  1664  DECES
> myocarde$PRONO = (myocarde$PRONO=="SURVIE")*1
```

Toy Datasets for the Course

Simulated data, $y \in \{0, 1\}$

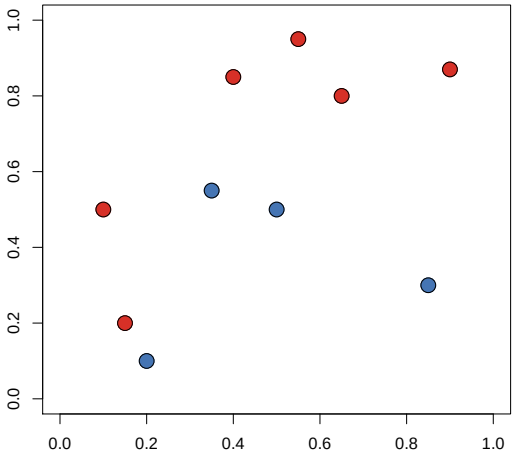
$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp[\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}]}{1 + \exp[\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}]}$$

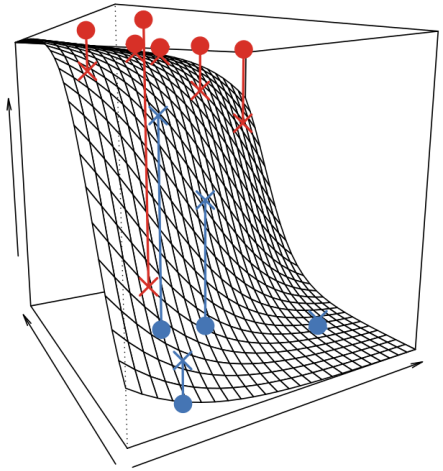
Inference using maximum likelihood techniques

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\left\{\sum_{i=1}^{n} \log[\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x})]\right\}$$

and the score model is then

$$s(\boldsymbol{x}) = \frac{\exp[\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}]}{1 + \exp[\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}]}$$





```r
x1 = c(.4,.55,.65,.9,.1,.35,.5,.15,.2,.85)
x2 = c(.85,.95,.8,.87,.5,.55,.5,.2,.1,.3)
y  = c(1,1,1,1,1,0,0,1,0,0)
df = data.frame(x1=x1,x2=x2,y=as.factor(y))
```

Machine Learning & Econometrics : The Two Cultures

Varian (2014, The Probabilistic Approach in Econometrics)

Google

Definitions

Machine learning, data mining, predictive analytics, etc. all use data to predict some variable as a function of other variables.

- May or may not care about insight, importance, patterns
- May or may not care about *inference*---how y changes as some x changes

Econometrics: Use statistical methods for prediction, inference, *causal* modeling of economic relationships.

- Hope for some sort of insight, inference is a goal
- In particular, *causal* inference is goal for decision making

Econometrics in a "Big Data" Context ?

Here data means $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{X}$ a $n \times p$ matrix.

n large means " asymptotic" theorems can be invoked $(n \to \infty)$

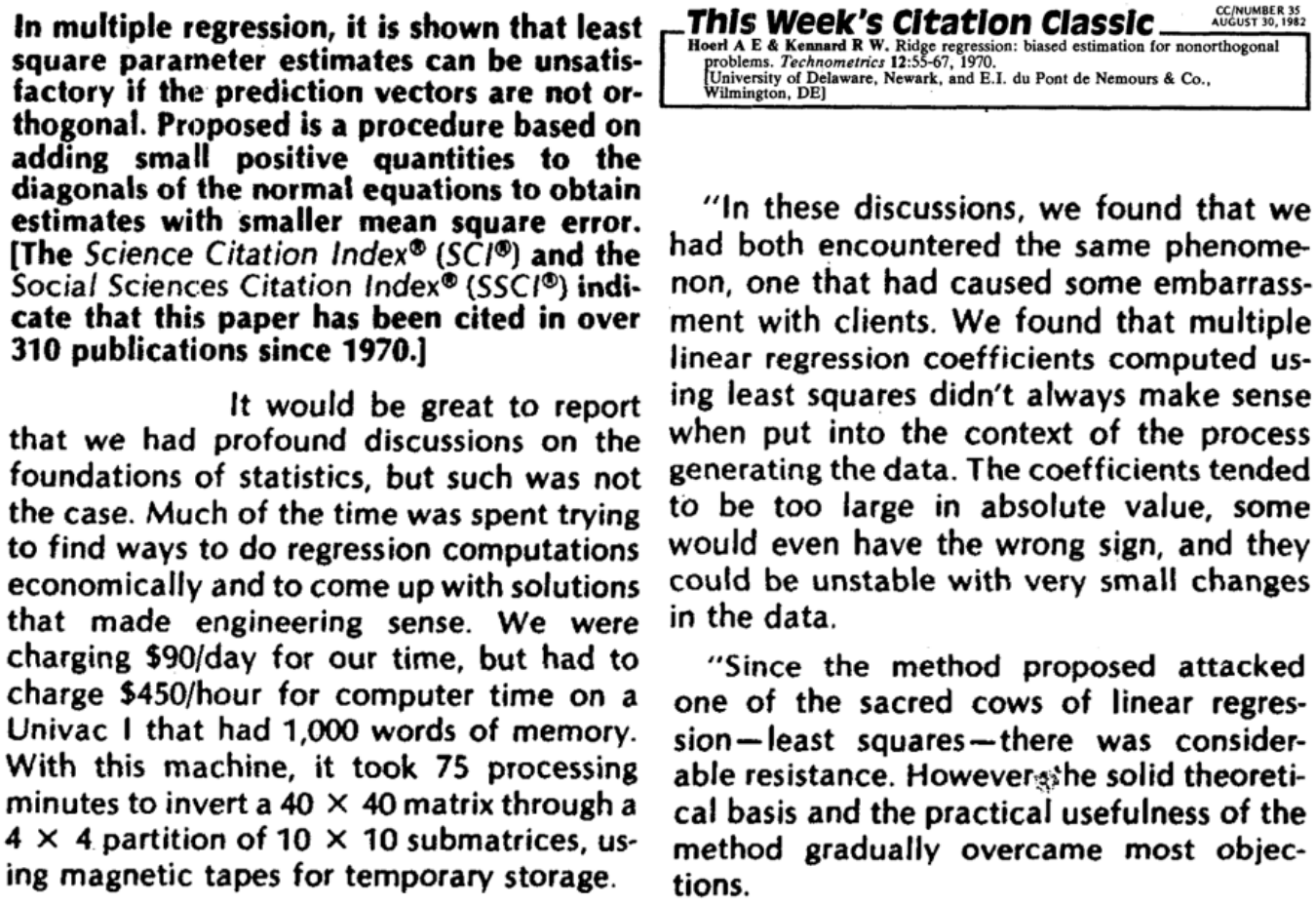
Portnoy (1988, Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity) proved that MLE estimator is asymptotically Gaussian as long as $p^2/n \to 0$.

There might be High-dimensional issues if $p > \sqrt{n}$.

Nevertheless, there might be sparsity issues in high dimension (see Hastie, Tibshirani & Wainwright (2015, Statistical Learning with Sparsity)) : a sparse statistical model has only a small number of nonzero parameters or weights

First order condition $\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) = \boldsymbol{0}$ is based on QR decomposition (can be computationaly intensive).

Computational Aspects of Econometrics



In multiple regression, it is shown that least square parameter estimates can be unsatisfactory if the prediction vectors are not orthogonal. Proposed is a procedure based on adding small positive quantities to the diagonals of the normal equations to obtain estimates with smaller mean square error. [The *Science Citation Index*® (*SCI*®) and the *Social Sciences Citation Index*® (*SSCI*®) indicate that this paper has been cited in over 310 publications since 1970.]

This Week's Citation Classic CC/NUMBER 35 AUGUST 30, 1982

Hoerl A E & Kennard R W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12:55-67, 1970.
[University of Delaware, Newark, and E.I. du Pont de Nemours & Co., Wilmington, DE]

It would be great to report that we had profound discussions on the foundations of statistics, but such was not the case. Much of the time was spent trying to find ways to do regression computations economically and to come up with solutions that made engineering sense. We were charging $90/day for our time, but had to charge $450/hour for computer time on a Univac I that had 1,000 words of memory. With this machine, it took 75 processing minutes to invert a 40 × 40 matrix through a 4 × 4 partition of 10 × 10 submatrices, using magnetic tapes for temporary storage.

"In these discussions, we found that we had both encountered the same phenomenon, one that had caused some embarrassment with clients. We found that multiple linear regression coefficients computed using least squares didn't always make sense when put into the context of the process generating the data. The coefficients tended to be too large in absolute value, some would even have the wrong sign, and they could be unstable with very small changes in the data.

"Since the method proposed attacked one of the sacred cows of linear regression—least squares—there was considerable resistance. However, the solid theoretical basis and the practical usefulness of the method gradually overcame most objections.

Arthur Hoerl in 1982, back on Hoerl & Kennard (1970) on Ridge regression.

Data & Models

We cannot differentiate data and model that easily...

After an operation, should I stay at hospital, or go back home ?

as in Angrist & Pischke (2008, Mostly Harmless Econometrics),

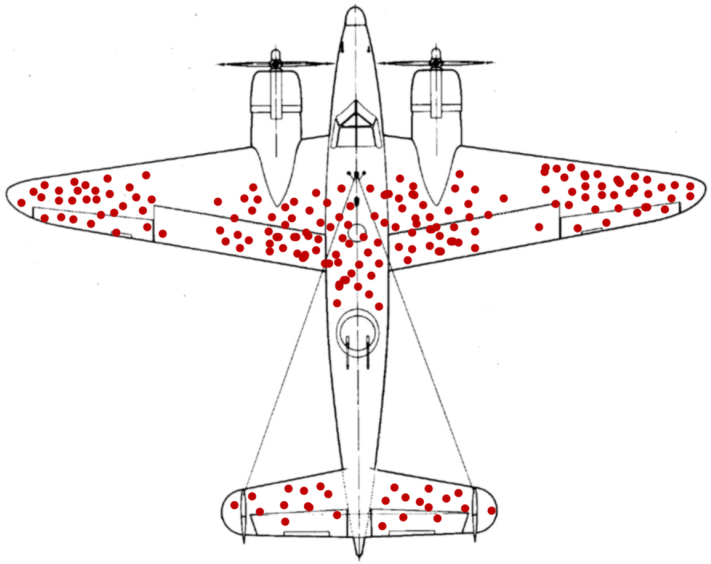
(health | hospital) − (health | stayed home)

should be written

(health | hospital) − (health | *had* stayed home)

+ (health | *had* stayed home) − (health | stayed hom

Need randomization to solve selection bias.



see also Mangel & Samaniego (1984, Abraham Wald's Work on Aircraft Survivability).

Back on the history of the "regression"

Galton (1870, Heriditary Genius, 1886, Regression towards mediocrity in hereditary stature) and Pearson & Lee (1896, On Telegony in Man, 1903 On the Laws of Inheritance in Man) studied genetic transmission of characterisitcs, e.g. the heigth.

On average the child of tall parents is taller than other children, but less than his parents.

"*I have called this peculiarity by the name of regression*", Francis Galton, 1886.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.

Table 8.1. Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

Height of the mid-parent in inches	Height of the adult child													Total no. of adult children	Total no. of mid-parents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7		
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	5	5	10	4	9	2	2	45	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

Source: Galton (1886a).

Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties: he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

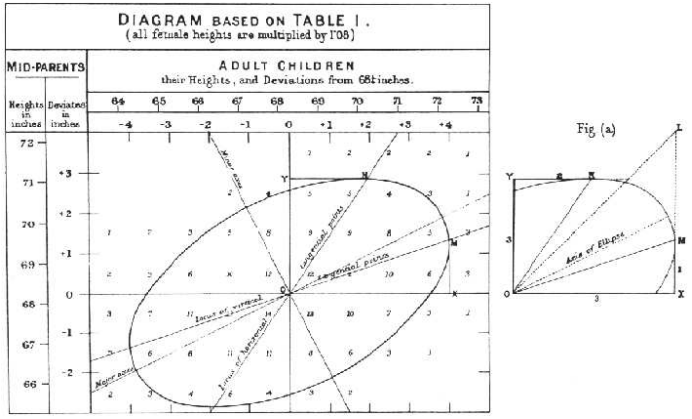
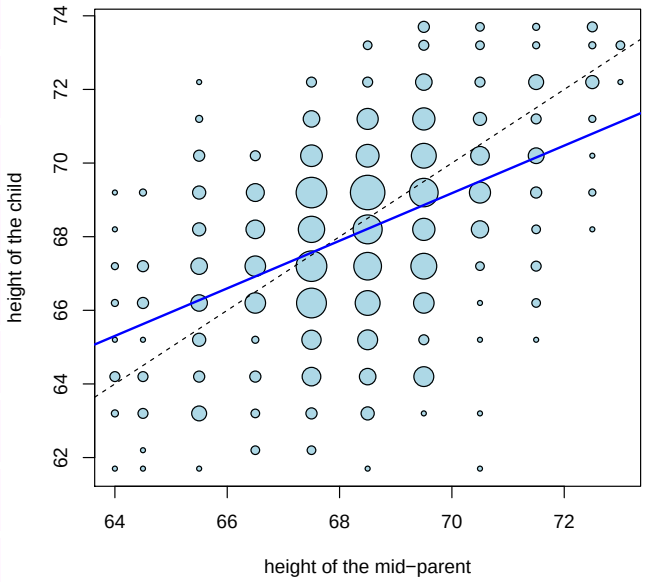


Figure 8.7. Galton's smoothed rendition of Table 8.1, with one of the "concentric and similar ellipses" drawn in. The geometric relationship of the two regression lines to the ellipse is also shown. (From Galton, 1886a.)

Back on the history of the "regression"

```
> library(HistData)
> attach(Galton)
> Galton$count <- 1
> df <- aggregate(Galton, by=list(parent,
    child), FUN=sum)[,c(1,2,5)]
> plot(df[,1:2],cex=sqrt(df[,3]/3))
> abline(a=0,b=1,lty=2)
> abline(lm(child~parent,data=Galton))
>  coefficients(lm(child~parent,data=Galton)
   )[2]
   parent
0.6462906
```



It is more an autoregression issue here :

if $Y_t = \phi Y_{t-1} + \varepsilon_t$, then $\text{cor}[Y_t, Y_{t+h}] = \phi^h \to 0$ as $h \to \infty$.

Mathematical Statistics Courses in a Nutshell - Non-Parametric Approach

Consider a sample $\{y_1, y_2, \cdots, y_n\}$. Its empirical cumulative distribution function is

$$\widehat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,y]}(y_i).$$

```
> F = ecdf(Y)
> F(180)
[1] 0.855
```

From Kolmogorov-Smirnov theorem $\lim_{n\to\infty} \widehat{F}_n(y) = F(y)$, while Glivenko-Cantelli theorem, states that the convergence in fact happens uniformly

$$\|\widehat{F}_n - F\|_\infty = \sup_{y\in\mathbb{R}} \left|\widehat{F}_n(y) - F(y)\right| \xrightarrow{\text{a.s.}} 0.$$

Mathematical Statistics Courses in a Nutshell - Non-Parametric Approach

Furthermore, pointwise, $\widehat{F}_n(y)$ has asymptotically normal distribution with the standard $\sqrt{n}$ rate of convergence:

$$\sqrt{n}\big(\widehat{F}_n(y) - F(y)\big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, F(y)\big(1-F(y)\big)\Big).$$

Let $\widehat{Q}_n$ denote the pseudo-inverse of $\widehat{F}_n$. Note that $\forall u \in (0,1)$, $\exists i$ such that $\widehat{Q}_n(u) = y_i$. More specifically, if $y_{1:n} \leq y_{2:n} \leq \cdots \leq y_{n:n}$,

$$\widehat{Q}_n(u) = y_{i:n} \text{ where } i - 1 \leq \frac{u}{n} < i.$$

Proposition Generating numbers from distribution $\widehat{F}_n$ means draw randomly, with replacement, uniformly, in $\{y_1, \cdots, y_n\}$.

Mathematical Statistics Courses in a Nutshell - Parametric Approach

The philosophy of the classical parametric paradigm is based on the following three beliefs:

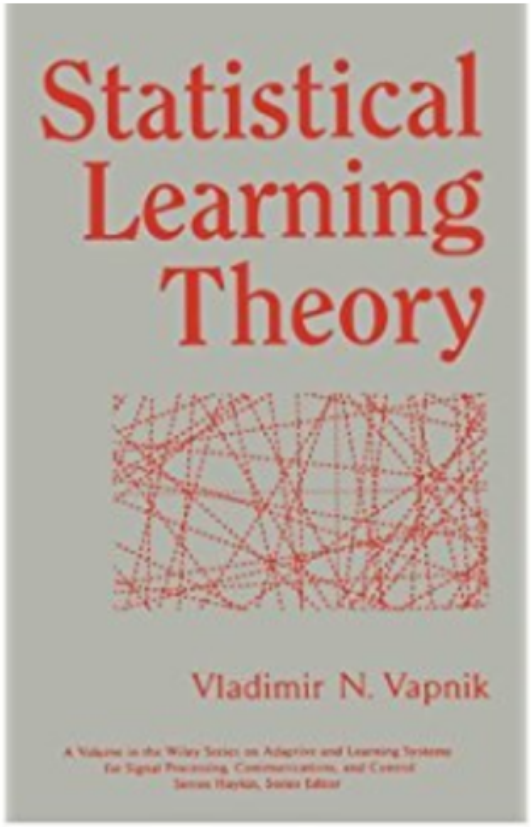
1. *To find a functional dependency from the data, the statistician is able to define a set of functions, linear in their parameters, that contain a good approximation to the desired function. The number of free parameters describing this set is small.*

2. *The statistical law underlying the stochastic component of most real-life problems is the normal law.*

3. *The induction engine in this paradigm—the maximum likelihood method —is a good tool for estimating parameters.*

Note that these three beliefs were also supported by the philosophy:

If there exists a mathematical proof that some method provides an asymptotically optimal solution, then in real life this method will provide a reasonable solution for a small number of data samples.



Vapnik (1998, Statistical Learning Theory)

Mathematical Statistics Courses in a Nutshell - Parametric Approach

Consider observations $\{y_1, \cdots, y_n\}$ from iid random variables $Y_i \sim F_{\boldsymbol{\theta}}$ (with "density" $f_{\boldsymbol{\theta}}$).

Likelihood is

$$\mathcal{L}(\boldsymbol{\theta}\,;\boldsymbol{y}) \mapsto \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(y_i)$$

Maximum likelihood estimate is

$$\widehat{\boldsymbol{\theta}}^{\text{mle}} \in \arg\max_{\boldsymbol{\theta}\in\Theta}\{\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})\}$$

ON AN ABSOLUTE CRITERION FOR FITTING FREQUENCY CURVES.

By *R. A. Fisher*, Gonville and Caius College, Cambridge.

theoretical curve, so the probability of any particular set of θ's is proportional to P, where

$$\log P = \sum_{1}^{n} \log f.$$

The most probable set of values for the θ's will make P a maximum.

If a continuous curve is observed—*e.g.*, the period during which a barometer is above any level during the year is a continuous function from which may be derived the relative frequency with which it stands at any height—we should use the expression

$$\log P = \int_{-\infty}^{\infty} y \log f \, dx.$$

Fisher (1912, On an absolute criterion for fitting frequency curves).

Mathematical Statistics Courses in a Nutshell - Parametric Approach

$\boldsymbol{y} = \{y_1, \cdots, y_n\}$ are observations, realizations from a collection of i.i.d. random variables $\{Y_1, \cdots, Y_n\}$. Assume that $Y_i \sim F_\theta$.

θ is the true, unknown, value of the parameter of the (true) distribution

An estimate $\widehat{\theta}$ is based on some statistic t (or some algorithm) so that $\widehat{\theta} = t(\boldsymbol{y})$

E.g. $\widehat{\theta} = \overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$

Hence $\widehat{\theta}$ is realization of random variable $\widehat{\Theta} = t(\boldsymbol{Y})$

The bias of $\widehat{\theta}$ is $\mathbb{E}_{F_\theta}\left[\widehat{\Theta}\right] - \theta$

The variance of $\widehat{\theta}$ is $\mathrm{Var}_{F_\theta}\left[\widehat{\Theta}\right] = \mathbb{E}_{F_\theta}\left[\left(\widehat{\Theta} - \mathbb{E}_{F_\theta}\left[\widehat{\Theta}\right]\right)^2\right]$

Mathematical Statistics Courses in a Nutshell - Parametric Approach

Under standard assumptions (Identification of the model, Compactness, Continuity and Dominance), the maximum likelihood estimator is consistent $\widehat{\boldsymbol{\theta}}^{\text{mle}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}$. With additional assumptions, it can be shown that the maximum likelihood estimator converges to a normal distribution

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}^{\text{mle}} - \boldsymbol{\theta}\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1})$$

where I is Fisher information matrix (i.e. $\widehat{\boldsymbol{\theta}}^{\text{mle}}$ is asymptotically efficient).

Eg. if $Y \sim \mathcal{N}(\theta, \sigma^2)$, $\log \mathcal{L} \propto -\sum_{i=1}^{n} \left(\frac{y_i - \theta}{\sigma} \right)^2$, and $\widehat{\theta}^{\text{mle}} = \overline{y}$ (see also method of moments).

$$\max \left\{ \log \mathcal{L} \right\} \Longleftrightarrow \min \left\{ \sum_{i=1}^{n} \varepsilon_i^2 \right\} = \text{least squares}$$

Mathematical Statistics Courses in a Nutshell

Classical tools : plug-in principle the empirical version of $\mathbb{E}[Y]$ is $\frac{1}{n}\sum_{i=1}^{n} y_i$. Thus, the empirical version of the variance $\mathbb{E}[(Y-\mathbb{E}[Y])^2]$ is

$$s^2(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \mathbb{E}[Y]\big)^2$$

and the plug-in version is

$$\widehat{s}^2(\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}\big(y_i - \overline{y}\big)^2$$

Similarly, $\text{Var}[\overline{y}] = n^{-1}\text{Var}[Y]$ and the plug-in estimator is

$$\widehat{\text{Var}}[\overline{y}] = \frac{1}{nn-1}\sum_{i=1}^{n}\big(y_i - \overline{y}\big)^2$$

Mathematical Statistics Courses in a Nutshell

Taylor approximation (or Delta-method) is based on

$$s(\widehat{\theta}) \approx s(\theta) + s'(\theta) \cdot [\widehat{\theta} - \theta]$$

which implies that $\text{Var}[s(\widehat{\theta})] \approx (s'(\theta))^2 \text{Var}[\widehat{\theta}]$. Then plug-in $\widehat{\theta}$ for unknown θ. Works well on large sample (as $n \to \infty$).

A pivotal statistics t has a distribution which **does not** depend on the underlying (unknown) probability distribution. Consider two samples $\{x_1, \cdots, x_{n_x}\}$ from a $\mathcal{N}(\mu_x, \sigma)$ and $\{y_1, \cdots, y_{n_y}\}$ from a $\mathcal{N}(\mu_y, \sigma)$. We want to test $H_0 : \mu_x = \mu_y$.

Idea 1: use $\widehat{\theta}_1 = \overline{x} - \overline{y}$, and

$$\widehat{\theta}_1 \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right) \text{ under } H_0$$

Naturally plug-in the unbiased estimator of σ^2,

$$\widehat{\sigma}^2 = \frac{1}{n_x + n_y - 2}\left(\sum_{i=1}^{n_x}(x_i - \overline{x})^2 + \sum_{i=1}^{n_y}(y_i - \overline{y})^2\right)$$

Mathematical Statistics Courses in a Nutshell - Parametric Approach

Idea 2: use $\widehat{\theta}_2 = \widehat{s}^{-1}(\overline{x} - \overline{y})$, were

$$\widehat{s}^2 = \widehat{\sigma}^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right)$$

so that $\widehat{\theta}_2 \sim \mathcal{S}td(n_x + n_y - 2)$.

MLE has interesting properties : in large sample, MLE is nearly unbiased, with the least possible variance (Cramér Rao bound)

Mathematical Statistics Courses in a Nutshell - Bayesian Approach

One can also define Empirical Bayes. Consider counts y_i so that $Y_i \sim \mathcal{P}(\theta_i)$, i.e.

$$p_{\theta_i}(y) = \mathbb{P}[Y_i = y] = e^{-\theta_i}\frac{\theta_i^y}{y!},\ y \in \mathbb{N}.$$

Let g denote the prior density of θ, so that

$$\mathbb{E}[\theta|x] = \frac{\int \theta p_\theta(x) g(\theta) d\theta}{\int p_\theta(x) g(\theta) d\theta} = (y+1)\frac{f(y+1)}{f(Y)}$$

where $f(y) = \int p_\theta(y) g(\theta) d\theta$ is the marginal density of Y

Thus, if $\widehat{p}_y$ is the empirical frequency of type y, $\widehat{\mathbb{E}}[\theta|y] = (y+1)\widehat{p}_{y+1}/\widehat{p}_y$.

Robbin's formula - from Robin(1956, An Empirical Bayes Approach to Statistics) - no need for a prior.

Mathematical Statistics Courses in a Nutshell - Shrinkage

Consider a Bayesian model, $Y|\mu \sim \mathcal{N}(\mu, 1)$ where the prior distribution of μ is $\mathcal{N}(M, A)$.

The posterior of μ is $\mu|y\mathcal{N}(M + B(y - M), A/(A + 1))$, thus Bayes estimator is $\widehat{\mu}^{\text{Bayes}} = M + B(y - M)$

Recall that $\widehat{\mu}^{\text{mle}} = y$

The expected squared error is respectively

$$\mathbb{E}\left[\left(\widehat{\mu}^{\text{Bayes}} - \mu\right)^2\right] = \frac{A}{A+1}$$

$$\mathbb{E}\left[\left(\widehat{\mu}^{\text{mle}} - \mu\right)^2\right] = 1$$

Consider now the case $Y_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$ where the prior distribution of μ_i is $\mathcal{N}(M, A)$. Set $\boldsymbol{\mu} = (\mu_1, c \cdots, \mu_n)$ and $\boldsymbol{y} = (y_1, c \cdots, y_n)$.

Mathematical Statistics Courses in a Nutshell - Shrinkage

Then $\widehat{\mu}_i^{\text{Bayes}} = M + B(y_i - M)$ or with a vector notation

$$\widehat{\boldsymbol{\mu}}^{\text{Bayes}} = \boldsymbol{M} + B(\boldsymbol{y} - \boldsymbol{M})$$

$$\widehat{\boldsymbol{\mu}}^{\text{Bayes}} = \boldsymbol{y}$$

Thus

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{\mu}}^{\text{Bayes}} - \boldsymbol{\mu}\right\|^2\right] = n\frac{A}{A+1} \text{ and } \mathbb{E}\left[\left\|\widehat{\boldsymbol{\mu}}^{\text{mle}} - \boldsymbol{\mu}\right\|^2\right] = n$$

Let $B = A/(A+1)$ then

$$\widehat{B} = 1 - \frac{n-3}{S} \text{ with } S = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

Is an unbiased estimator of B (as long as $n > 3$) James-Stein estimator is

$$\widehat{\boldsymbol{\mu}}^{\text{JS}} = \widehat{\boldsymbol{M}} + \widehat{B}(\boldsymbol{y} - \widehat{\boldsymbol{M}})$$

Mathematical Statistics Courses in a Nutshell - Shrinkage

It can be seen as some "empirical Bayes" estimate.

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{JS}}-\boldsymbol{\mu}\right\|^2\right]=nB+3(1-B)$$

With $n=20$ and $A=1$,

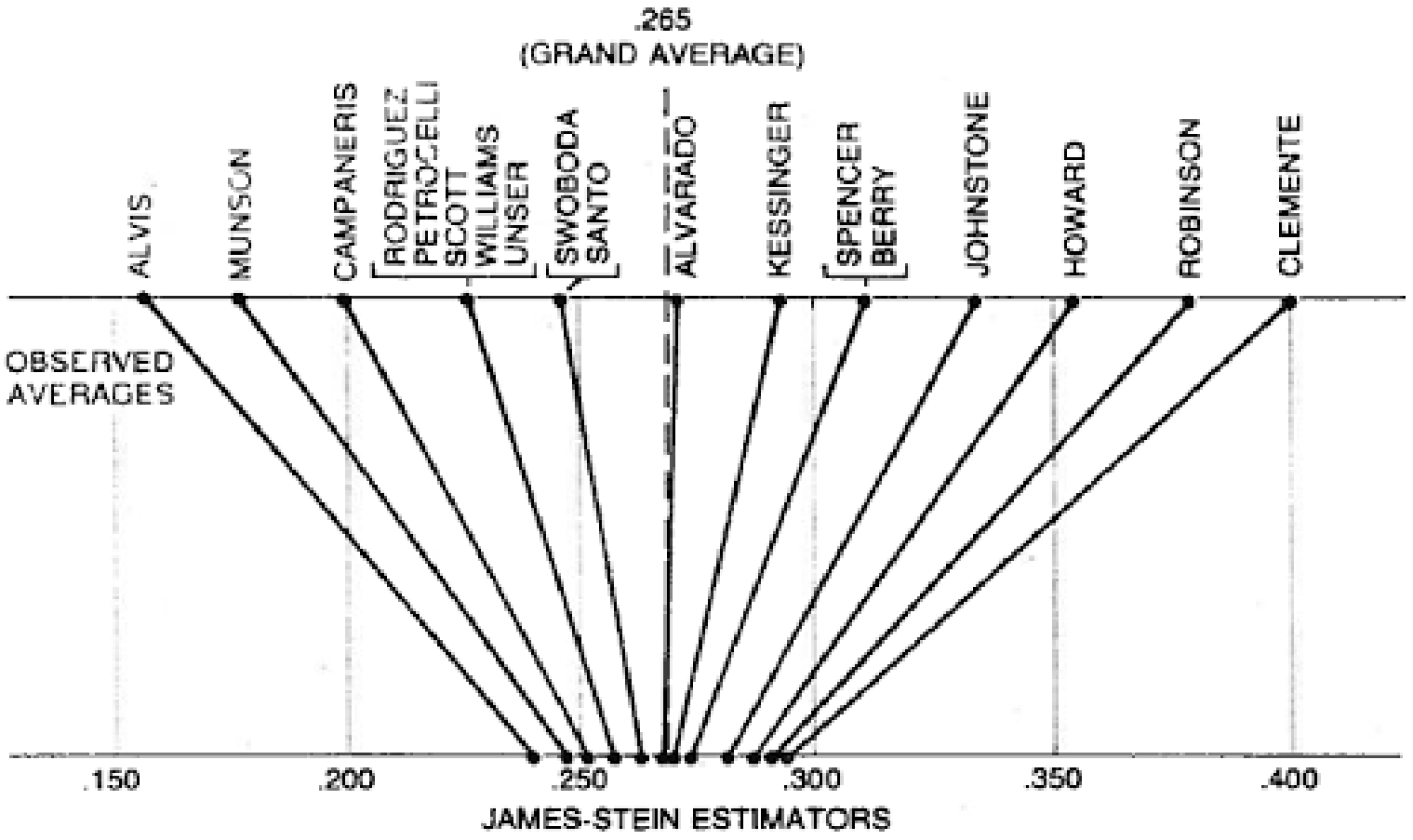
$$\underbrace{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{Bayes}}-\boldsymbol{\mu}\right\|^2\right]}_{=10}<\underbrace{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{JS}}-\boldsymbol{\mu}\right\|^2\right]}_{=11.5}<\underbrace{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{mle}}-\boldsymbol{\mu}\right\|^2\right]}_{=20}$$

More generally, if $Y_i|\mu_i\sim\mathcal{N}(\mu_i,1)$ and $n\geq 4$, then

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{JS}}-\boldsymbol{\mu}\right\|^2\right]<\underbrace{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\mu}}^{\mathsf{mle}}-\boldsymbol{\mu}\right\|^2\right]}_{=n}$$

Thus $\widehat{\boldsymbol{\mu}}^{\mathsf{mle}}$ is inadmissible since its squared error risks dominates $\widehat{\boldsymbol{\mu}}^{\mathsf{JS}}$ whatever $\boldsymbol{\mu}$ is !

Mathematical Statistics Courses in a Nutshell - Shrinkage



Mathematical Statistics Courses in a Nutshell - Shrinkage

Recall that for a linear regression model $\widehat{\boldsymbol{\beta}}^{\text{ols}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$.

Define $\widehat{\boldsymbol{\beta}}^{\text{ridge}} = (\boldsymbol{X}^{t}ranspose\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda\mathbb{I})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\text{ols}}$.

One can prove that

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \text{argmin}\big\{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2\big\}$$

i.e. penalized least squares / penalized likelihood / regularization.

Set

$$\widehat{\boldsymbol{\beta}}^{\text{JS}} = \left(1 - \frac{(p-2)\sigma^2}{\widehat{\boldsymbol{\beta}}^{\text{ols}\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\text{ols}}}\right)\widehat{\boldsymbol{\beta}}^{\text{ols}}$$

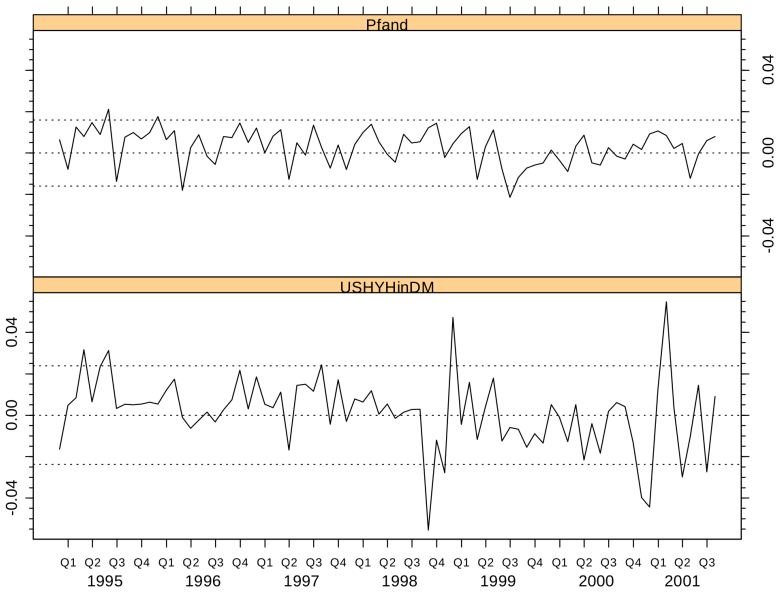
Observe that

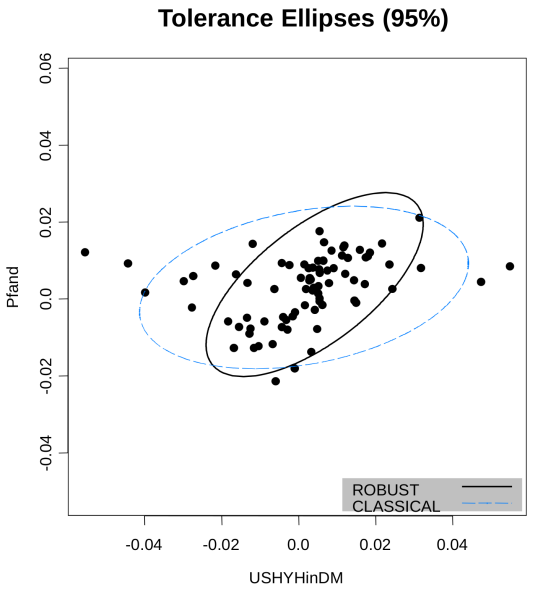
$$\mathbb{E}\left[\left\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\text{JS}} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2\right] < p\sigma^2$$

Mathematical Statistics Courses in a Nutshell - Robust Approach

Robust inference is important in real life applications (see Hubert, Rousseeuw & van Aelst (2004, Robustness))

How robust is that estimator ? See Martin (2014) on financial time series





MLE (or classical) correlation estimator $\widehat{\theta} \sim 30\%$ while $\widehat{\theta}^{\text{mcd}} \sim 65\%$

Mathematical Statistics Courses in a Nutshell - Robust Approach

(here fast minimum covariance determinant - MCD - see Rousseeuw (1984), but see also the Stahel-Donoho estimator (SDE) is due to Stahel (1981) and Donoho (1982), while the OGK estimator was proposed by Maronna & Zamar (2002).)

Let θ denote a parameter, such that $\theta = T(F)$ for instance $\theta = \int x dF(x)$. The empirical version is $\widehat{\theta} = T(\widehat{F})$, i.e. in the context of the mean,

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The influence function of $T(F)$, evaluated at point y is

$$IF(y) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\big(T((1-\epsilon)F + \epsilon\delta_y) - T(F)\big)$$

Mathematical Statistics Courses in a Nutshell - Robust Approach

$$IF(y) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\big(T((1-\epsilon)F + \epsilon\delta_y) - T(F)\big)$$

$IF(y)$ measures the differential effect of modifying F when we put additional probability on y E.g. if $\theta = \mathbb{E}[Y]$, $IF(y) = y - \theta$.

$$\underbrace{\widehat{\theta}}_{T(\widehat{F})} = \underbrace{\theta}_{T(F)} + \frac{1}{n}\sum_{i=1}^{n} IF(y_i)$$

Further, $\text{Var}[\widehat{\theta}] = \dfrac{1}{n}\text{Var}[IF(y)]$

Econometrics Courses in a Nutshell

Haavelno (1944, The Probabilistic Approach in Econometrics)

THE PROBABILITY APPROACH
IN ECONOMETRICS

By

Trygve Haavelmo

RESEARCH ASSOCIATE
COWLES COMMISSION FOR
RESEARCH IN ECONOMICS

Supplement to Econometrica, Volume 12, July, 1944

THE ECONOMETRIC SOCIETY
The University of Chicago
Chicago 37, Illinois

CHAPTER III

Stochastical Schemes as a Basis for Econometrics

As far as is known, the scheme of probability and random variables is, at least for the time being, the only scheme suitable for formulating such theories. We may have objections to using this scheme, but among these objections there is at least one that can be safely dismissed, viz., the objection that the scheme of probability and random variables is not general enough for application to economic data. Since, however, this is apparently not commonly accepted by economists we find ourselves justified in starting our discussion in this chapter with a brief outline of the modern theory of stochastical variables, with particular emphasis on certain points that seem relevant to economics.

The more recent developments in statistical theory are based upon the so-called modernized classical theory of probability. Here "probability" is defined as an absolutely additive and nonnegative *set-function*,[1] satisfying certain formal properties.[2]

Let us first take an example to illustrate this probability concept.

Data $(y_i, \boldsymbol{x}_i)$ are seen as realizations of (iid) random varibles $(Y, \boldsymbol{X})$ on some probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$.

Econometrics Courses in a Nutshell

Consider a linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, with matrix notation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Assume

- correct specification
- exogeneity, i.e. $\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0$. Thus, residuals are centered $\mathbb{E}[\varepsilon] = 0$ and covariates are uncorrelated with the errors $\mathbb{E}[\boldsymbol{X}^\top \varepsilon] = \boldsymbol{0}$
- covariates are linearly independent, i.e. $\mathbb{P}[\text{rank}(\boldsymbol{X}) = p] = 1$
- spherical errors, i.e. $\text{Var}[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \sigma^2\mathbb{I}$. Thus, residuals are homoscedasticity - $\text{Var}[\varepsilon_i|\boldsymbol{X}] = \sigma^2$ - and non-correlated $\mathbb{E}[\varepsilon_i\varepsilon_j|\boldsymbol{X}] = 0$, $\forall i \neq j$.
- gaussian errors, i.e. $\boldsymbol{\varepsilon}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathbb{I})$

$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ is the least-square estimator of $\boldsymbol{\beta}$, obtained as

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\left\{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2\right\}.$$

Inference in the Linear Model

Consider a linear model $y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i$, with matrix notation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Assume

- correct specification
- exogeneity, i.e. $\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0$. Thus, residuals are centered $\mathbb{E}[\varepsilon] = 0$ and covariates are uncorrelated with the errors $\mathbb{E}[\boldsymbol{X}^\mathsf{T}\varepsilon] = \boldsymbol{0}$
- covariates are linearly independent, i.e. $\mathbb{P}[\text{rank}(\boldsymbol{X}) = p] = 1$
- spherical errors, i.e. $\text{Var}[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \sigma^2\mathbb{I}$. Thus, residuals are homoscedasticity - $\text{Var}[\varepsilon_i|\boldsymbol{X}] = \sigma^2$ - and non-correlated $\mathbb{E}[\varepsilon_i\varepsilon_j|\boldsymbol{X}] = 0$, $\forall i \neq j$.
- gaussian errors, i.e. $\boldsymbol{\varepsilon}|\boldsymbol{X} \sim \mathcal{N}$

$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$ is the least-square estimator of $\boldsymbol{\beta}$, obtained as

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\left\{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2\right\}.$$

Inference in the Linear Model

Under the Gaussian assumption, $\widehat{\boldsymbol{\beta}}$ is also the maximum-likelihood estimator (MLE) of $\boldsymbol{\beta}$.

Under the exogeneity assumption, $\widehat{\boldsymbol{\beta}}$ is the solution of $\mathbb{E}[\boldsymbol{x}_i[y_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}]] = \mathbf{0}$, i.e. it is also the the Generalized method of moments estimator (GMM) of $\boldsymbol{\beta}$.

Observe furthermore that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$: it is linear in $\boldsymbol{y}$.

$$\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}\right)^2$$ is the least-square estimator of σ^2.

Under the exogeneity assumption, OLS estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are unbiased, i.e.

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}|\boldsymbol{X}] = \boldsymbol{\beta} \text{ and } \mathbb{E}[\widehat{\sigma}^2|\boldsymbol{X}] = \sigma^2$$

Furthermore, the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\text{Var}[\widehat{\boldsymbol{\beta}}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}.$$

One can prove that $\text{Cov}[\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2|\boldsymbol{X}] = \mathbf{0}$.

Inference in the Linear Model

From Gauss-Markov theorem, with spherical residuals (errors should be uncorrelated and homoscedastic), $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) (in the sense that $\text{Var}[\widetilde{\boldsymbol{\beta}}|\boldsymbol{X}] - \text{Var}[\widehat{\boldsymbol{\beta}}|\boldsymbol{X}]$ is a non-negative definite matrix for any unbiased estimator $\widetilde{\boldsymbol{\beta}}$ linear in $\boldsymbol{y}$, i.e. $\widetilde{\boldsymbol{\beta}} = \boldsymbol{M}\boldsymbol{y}$).

Assuming normality of the residuals, we can prove that $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1})$

This estimator reaches the Cramér-Rao bound for the model, and thus is optimal in the class of all unbiased estimators (linear and non-linear).

Furthermore, $\widehat{\sigma}^2 \sim \dfrac{\sigma^2}{n-p} \cdot \chi^2_{n-p}$. Even if it is not optimal, there are no unbiased estimators of σ^2 with variance smaller.

Without normality assumption, $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically normal, $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1})$, as $n \to \infty$.

Similarly, one can prove that $\widehat{\sigma}^2 \xrightarrow{\mathcal{L}} \mathcal{N}(\sigma^2, \mathbb{E}[\varepsilon^4]\sigma^4)$, as $n \to \infty$.

Bayesian Linear Model

Consider a linear regression model, $Y = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta} + \varepsilon$, with some Gaussian i.i.d. noise.

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^n} \exp\Big[- \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\Big]$$

Set $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$, which satisfies

$$\big[\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\big]^\mathsf{T}\big[\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\big] = 0$$

Consider a diffuse prior $\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta})\pi(\sigma^2)$ with $\pi(\boldsymbol{\beta}) \propto$ constant and $\pi(\sigma^2) = 1/\sigma^2$, i.e. $\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$

First, let's condition on σ^2, then marginalize and focus just on $\boldsymbol{\beta}$, so that

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2) \propto \exp\left[-\frac{1}{2\sigma^2}\big((n-k)s^2 + [\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}]^\mathsf{T}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}][\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}]\big)\right]$$

i.e.

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2) \propto \exp\left[-\frac{1}{2}[\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}]^\mathsf{T}\left[\sigma^2[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\right]^{-1}[\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}]\big)\right]$$

Bayesian Linear Model

which is a Gaussian distribution with mean $\widehat{\boldsymbol{\beta}}$ and variance matrix $\sigma^2[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}$

Hence, Bayes estimator for various symmetric loss function is the MLE.

If we marginalize, i.e. $\pi(\boldsymbol{\beta}|\boldsymbol{y}) = \displaystyle\int_{\mathbb{R}_+} \pi(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})d\sigma^2$ We can easily prove that

$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto \left[(n-k)s^2 + [\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]^\mathsf{T}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}][\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}]\right]^{-n/2}$ which is the kernel of a Student-t distribution. On the other hand

$$\pi(\sigma^2|\boldsymbol{y}) = \int_{\mathbb{R}^k} \pi(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})d\boldsymbol{\beta}$$

We can easily prove that

$$\pi(\sigma^2|\boldsymbol{y}) \propto \sigma^{-(n-k+1)} \exp\left(-\frac{(n-k)s^2}{2\sigma^2}\right)$$

which is the kernel of a Inverted Gamma distribution.

Bayesian Linear Model

Hence

$$\mathbb{E}[\sigma^2|\boldsymbol{y}] = s^2(n-k)\Gamma\left(\frac{n-k-1}{2}\right)/\Gamma\left(\frac{n-k}{2}\right) \ (\to s^2 \text{ as } n \to \infty)$$

while

$$\text{Var}[\sigma^2|\boldsymbol{y}] = \left[\frac{(n-k)s^2}{n-k-2}\right] - \mathbb{E}[\sigma^2|\boldsymbol{y}]^2$$

If we consider a conjugate prior $\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2)$

Here $\pi(\boldsymbol{\beta}|\sigma^2)$ is a (conditional Gaussian distribution, while $\pi(\sigma^2)$ is an inverted Gamma distribution. More precisely

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}(\boldsymbol{b}, \sigma^2 A^{-1})$$

Bayesian Linear Model

One can prove that the conditional posterior distribution for $\boldsymbol{\beta}$ is a Gaussian distribution,

$$\boldsymbol{\beta}|\sigma^2,\boldsymbol{y}\sim\mathcal{N}\big(\widetilde{\boldsymbol{\beta}},\sigma^2[A+\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\big)$$

where

$$\widetilde{\boldsymbol{\beta}}=[A+\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\big[A\boldsymbol{b}+\boldsymbol{X}^\mathsf{T}\boldsymbol{y}\big]$$

If we marginalize, i.e.

$$\pi(\boldsymbol{\beta}|\boldsymbol{y})=\int_{\mathbb{R}_+}\pi(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})d\sigma^2$$

We can easily prove that, if σ_0^2 is the mean of the prior distribution of σ^2

$$\pi(\boldsymbol{\beta}|\boldsymbol{y})\propto\Big[(n+\sigma_0^2-k)c^2+[\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}]^\mathsf{T}[A+\boldsymbol{X}^\mathsf{T}\boldsymbol{X}][\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}]\Big]^{-n+\sigma_0^2+k/2}$$

(for some constant c) which is the kernel of a Student-t distribution.

OIn the other hand

$$\pi(\sigma^2|\boldsymbol{y})=\int_{\mathbb{R}^k}\pi(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})d\boldsymbol{\beta}$$

Bayesian Linear Model

We can easily prove that

$$\pi(\sigma^2|\boldsymbol{y}) \propto \sigma^{-(n+\sigma_0^2-k+1)} \exp\left(-\frac{(n+\sigma_0^2-k)c^2}{2\sigma^2}\right)$$

which is the kernel of a Inverted Gamma distribution.

One can also write

$$\widetilde{\boldsymbol{\beta}} = \left[\frac{1}{\sigma^2}A + \frac{1}{\sigma^2}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\right]^{-1}\left[\frac{1}{\sigma^2}A\boldsymbol{b} + \frac{1}{\sigma^2}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right]$$

which is a (matrix base) weighted average of $\boldsymbol{b}$ (priori mean) and $\widehat{\boldsymbol{\beta}}$ (MLE).

Further $\widetilde{\boldsymbol{\beta}}$ even if $\text{rank}(\boldsymbol{X}) < k$ (as soon as A is positive definite).

This is Ridge estimator. Stein and Theil estimates are other examples of mixed estimators.

Econometrics Courses in a Nutshell : Interpretation

```
> runif(30)
library("DALEX")
apartments_lm_model <- lm(m2.price ~ construction.year + surface +
    floor +
                          no.rooms + district, data = apartments)
summary(apartments_lm_model)

Coefficients:
                    Estimate  Std. Error t value Pr(>|t|)
(Intercept)       5020.1391    682.8721   7.352 4.11e-13 ***
construction.year   -0.2290      0.3483  -0.657   0.5110
surface            -10.2378      0.5778 -17.720  < 2e-16 ***
floor              -99.4820      3.0874 -32.222  < 2e-16 ***
no.rooms           -37.7299     15.8440  -2.381   0.0174 *
```

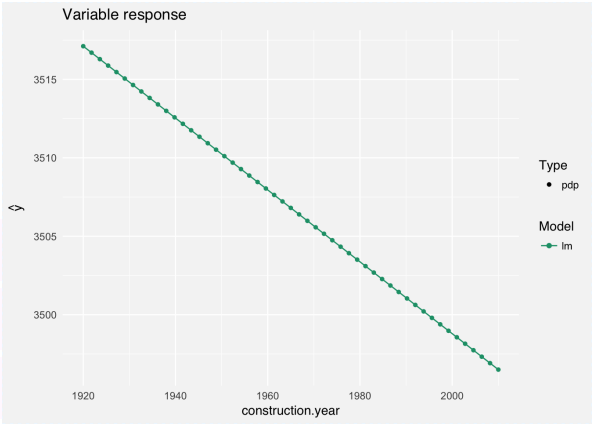
Econometrics Courses in a Nutshell : Interpretation

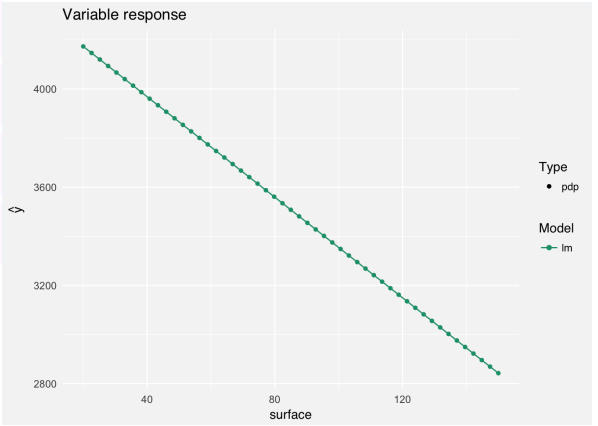
```
districtBielany          17.2144     40.4502    0.426    0.6705
districtMokotow         918.3802     39.4386   23.286   < 2e-16 ***
districtOchota          926.2540     40.5279   22.855   < 2e-16 ***
districtPraga           -37.1047     40.8930   -0.907    0.3644
districtSrodmiescie    2080.6110     40.0149   51.996   < 2e-16 ***
districtUrsus            29.9419     39.7249    0.754    0.4512
districtUrsynow         -18.8651     39.7565   -0.475    0.6352
districtWola            -16.8912     39.6283   -0.426    0.6700
districtZoliborz        889.9735     40.4099   22.024   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

library(auditor)
```

Econometrics Courses in a Nutshell : Interpretation

Impact of a continuous covariate

```
sv_lm  <- variable_response(explainer_lm,
    variable =  "construction.year", type = "
    pdp")
plot(sv_lm)
sv_lm  <- variable_response(explainer_lm,
    variable =  "surface", type = "pdp")
plot(sv_lm)
```

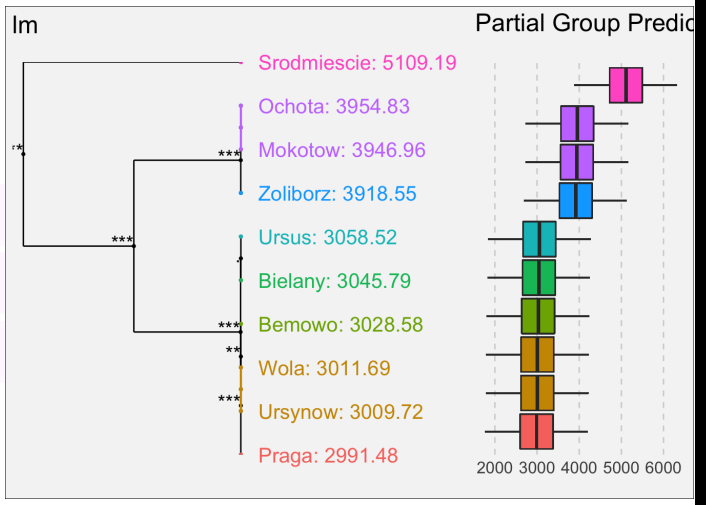




Econometrics Courses in a Nutshell : Interpretation

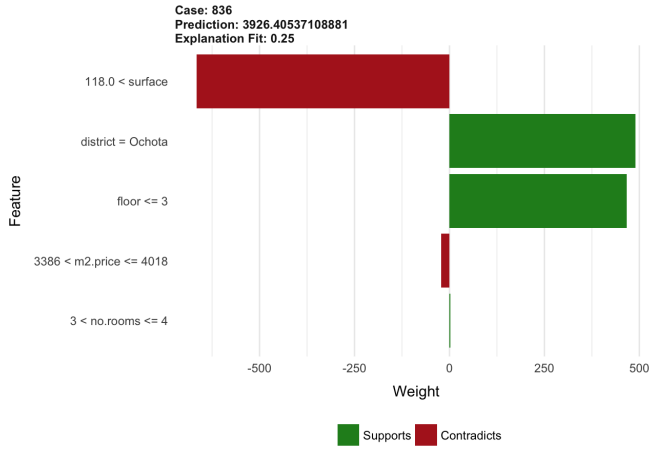
Impact of a discrete (qualitative) covariate

```
1 svd_lm  <- variable_response(explainer_lm,
      variable = "district", type = "factor")
2 plot(svd_lm)
```



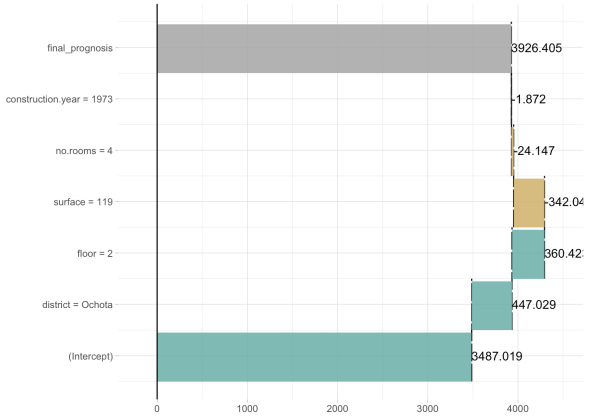
Econometrics Courses in a Nutshell : Interpretation

```
1 library(mlr)
2 apartments_task <- makeRegrTask(data =
      apartments, target = "m2.price")
3 apartments_lm_mlr <- mlr::train("regr.lm",
      apartments_task)
4 library(lime)
5 explained_prediction <- apartments[836, ]
6 lime_explainer <- lime(apartments, model =
      apartments_lm_mlr)
7 lime_explanation <- lime::explain(apartments
      [836, ], explainer = lime_explainer, n_
      features = 5)
8 plot_features(lime_explanation)
```



Econometrics Courses in a Nutshell : Interpretation

```
1 library(breakDown)
2 predict.function <- function(model, new_
      observation) stats::predict.lm(model,
      newdata=new_observation)
3 explain_1 <- broken(apartments_lm_model,
      apartments[836, ], data = apartments,
      predict.function = predict.lm, direction =
      "down")
4 plot(explain_1)
```



Econometrics Courses in a Nutshell - Parallel Computations

Consider the case where datasets are located on various servers, and cannot be downloaded (e.g. hospitals), but one can run functions and obtain outputs.
see Wolfson *et. al* (2010, Data Shield)
or http://www.datashield.ac.uk/

Consider a regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Use the QR decomposition of $\boldsymbol{X}$, $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{Q}$ is an orthogonal matrix $\boldsymbol{Q}^\mathsf{T}\boldsymbol{Q} = \mathbb{I}$. Then

$$\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} = \boldsymbol{R}^{-1}\boldsymbol{Q}^\mathsf{T}\boldsymbol{y}$$

Econometrics Courses in a Nutshell - Parallel Computations

Consider m blocks - map part

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix} \text{ and } \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_1^{(1)}\boldsymbol{R}_1^{(1)} \\ \boldsymbol{Q}_2^{(1)}\boldsymbol{R}_2^{(1)} \\ \vdots \\ \boldsymbol{Q}_m^{(1)}\boldsymbol{R}_m^{(1)} \end{bmatrix}$$

Econometrics Courses in a Nutshell - Parallel Computations

Consider the QR decomposition of $\boldsymbol{R}^{(1)}$ - step 1 of reduce part

$$\boldsymbol{R}^{(1)} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{R}_2 \\ \vdots \\ \boldsymbol{R}_m \end{bmatrix} = \boldsymbol{Q}^{(2)}\boldsymbol{R}^{(2)} \text{ where } \boldsymbol{Q}^{(2)} = \begin{bmatrix} \boldsymbol{Q}_1^{(2)} \\ \boldsymbol{Q}_2^{(2)} \\ \vdots \\ \boldsymbol{Q}_m^{(2)} \end{bmatrix}$$

define - step 2 of reduce part

$$\boldsymbol{Q}_j^{(3)} = \boldsymbol{Q}_j^{(2)}\boldsymbol{Q}_j^{(1)} \text{ and } \boldsymbol{V}_j = \boldsymbol{Q}_j^{(3)\mathsf{T}}\boldsymbol{y}_j$$

and finally set - step 3 of reduce part

$$\widehat{\boldsymbol{\beta}} = [\boldsymbol{R}^{(2)}]^{-1}\sum_{j=1}^{m}\boldsymbol{V}_j$$

Econometrics Courses in a Nutshell - Updating Formulas

• Update with a new observation, as Ridell (1975, Recursive Estimation Algorithms for Economic Research)

Let $\boldsymbol{X}_{1:n}$ denote the matrix of covariates, with n observations (rows), and $\boldsymbol{x}_{n+1}$ denote a new one. Recall that

$$\widehat{\boldsymbol{\beta}}_n = [\boldsymbol{X}_{1:n}^\mathsf{T}\boldsymbol{X}_{1:n}]^{-1}\boldsymbol{X}_{1:n}^\mathsf{T}\boldsymbol{y}_{1:n}$$

Let $d = 1 + \boldsymbol{x}_{n+1}[\boldsymbol{X}_{1:n}^\mathsf{T}\boldsymbol{X}_{1:n}]^{-1}\boldsymbol{x}_{n+1}^\mathsf{T}$, then

$$\widehat{\boldsymbol{\beta}}_{n+1} = \widehat{\boldsymbol{\beta}}_n + \frac{[\boldsymbol{X}_{1:n}^\mathsf{T}\boldsymbol{X}_{1:n}]^{-1}\boldsymbol{x}_{n+1}^\mathsf{T}}{d}[y_{n+1} - \boldsymbol{x}_{n+1}\widehat{\boldsymbol{\beta}}_n^\mathsf{T}]$$

This updating formation is also called a differential correction, since it is proportional to the prediction error.

Econometrics Courses in a Nutshell - Updating Formulas

Note that the residual sum of squares can also be updated, with

$$S_{n+1} = S_n + \frac{1}{d}[y_{n+1} - \boldsymbol{x}_{n+1}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_n]^2$$

See online learning : we have a sample $\{y_1, \cdots, y_t\}$, consider a model $\widehat{m}_t$. What if we have a new observation, y_{t+1} ? Can we easily retrive $\widehat{m}_{t+1}$?

Econometrics Courses in a Nutshell - Updating Formulas

- Update with a new variable

Let $\boldsymbol{X}_{1:k}$ denote the matrix of covariates, with k explanatory variables (columns), and $\boldsymbol{x}_{k+1}$ denote a new one. Recall that

$$\widehat{\boldsymbol{\beta}}_k = [\boldsymbol{X}_{1:k}^{\mathsf{T}}\boldsymbol{X}_{1:k}]^{-1}\boldsymbol{X}_{1:k}^{\mathsf{T}}\boldsymbol{y}$$

Then $\widehat{\boldsymbol{\beta}}_{k+1} = (\widehat{\boldsymbol{\beta}}_k^{\star}\ ,\ \widehat{\beta}_{k+1}^{\star})^{\mathsf{T}}$ where

$$\widehat{\boldsymbol{\beta}}_k^{\star} = \widehat{\boldsymbol{\beta}}_k - \frac{[\boldsymbol{X}_{1:k}^{\mathsf{T}}\boldsymbol{X}_{1:k}]^{-1}\boldsymbol{X}_{1:k}^{\mathsf{T}}\boldsymbol{x}_{k+1}\boldsymbol{x}_{k+1}P_k^{\perp}\boldsymbol{y}}{\boldsymbol{x}_{k+1}^{\mathsf{T}}P_k^{\perp}\boldsymbol{x}_{k+1}}$$

with $P_k^{\perp} = \mathbb{I} - \boldsymbol{X}_{1:k}(\boldsymbol{X}_{1:k}^{\mathsf{T}}\boldsymbol{X}_{1:k})^{-1}\boldsymbol{X}_{1:k}^{\mathsf{T}}$, while

$$\widehat{\beta}_{k+1}^{\star} = \frac{\boldsymbol{x}_{k+1}^{\mathsf{T}}P_k^{\perp}\boldsymbol{y}}{\boldsymbol{x}_{k+1}^{\mathsf{T}}P_k^{\perp}\boldsymbol{x}_{k+1}}$$

If $\boldsymbol{x}_{k+1}$ is orthogonal to previous variables - $\boldsymbol{X}_{1:k}^{\mathsf{T}}\boldsymbol{x}_{k+1} = \boldsymbol{0}$, then $\widehat{\boldsymbol{\beta}}_k^{\star} = \widehat{\boldsymbol{\beta}}_k$.

Observe that $P_k^{\perp}\boldsymbol{y} = \varepsilon_k$.

Optimization Courses in a Nutshell - Convex Problems

Problem : $\boldsymbol{x}^\star \in \text{argmin}\{f(\boldsymbol{x});\ \boldsymbol{x} \in \mathbb{R}^d\}$

Gradient descent : $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)$ starting from some $\boldsymbol{x}_0$

Problem : $\boldsymbol{x}^\star \in \text{argmin}\{f(\boldsymbol{x});\ \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d\}$

Projected descent : $\boldsymbol{x}_{k+1} = \Pi_{\mathcal{X}}\big(\boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)\big)$ starting from some $\boldsymbol{x}_0$

A constrained problem is said to be convex if

$$\begin{cases} \min\{f(\boldsymbol{x})\} & \text{with } f \text{ convex} \\ \text{s.t. } g_i(\boldsymbol{x}) = 0,\ \forall i = 1, \cdots, n & \text{with } g_i \text{ linear} \\ \quad h_i(\boldsymbol{x}) \leq 0,\ \forall i = 1, \cdots, m & \text{with } h_i \text{ convex} \end{cases}$$

Lagrangian : $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{n} \lambda_i g_i(\boldsymbol{x}) + \sum_{i=1}^{m} \mu_i h_i(\boldsymbol{x})$ where $\boldsymbol{x}$ are primal variables and $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ are dual variables.

Remark $\mathcal{L}$ is an affine function in $(\boldsymbol{\lambda}, \boldsymbol{\mu})$

Optimization Courses in a Nutshell - Convex Problems

Karush-Kuhn-Tucker conditions : a convex problem has a solution $\boldsymbol{x}^\star$ if and only if there are $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ such that the following condition hold

- stationarity : $\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}$ at $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$
- primal admissibility : $g_i(\boldsymbol{x}^\star) = 0$ and $h_i(\boldsymbol{x}^\star) \leq 0$, $\forall i$
- dual admissibility : $\boldsymbol{\mu}^\star \geq 0$

Let L denote the associated dual function $L(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\boldsymbol{x}}\{\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})\}$

L is a convex function in $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ and the dual problem is $\max_{\boldsymbol{\lambda}, \boldsymbol{\mu}}\{L(\boldsymbol{\lambda}, \boldsymbol{\mu})\}$.

In the case of linear problems, specific algorithms can be used...

Machine Learning Courses in a Nutshell

Data Mining and Statistics: What's the Connection?

Jerome H. Friedman
Department of Statistics and
Stanford Linear Accelerator Center
Stanford University
Stanford, CA 94305
jhf@stat.stanford.edu

Abstract

Data Mining is used to discover patterns and relationships in data, with an emphasis on large observational data bases. It sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization. From a statistical perspective it can be viewed as computer automated exploratory data analysis of (usually) large complex data sets. In spite of (or perhaps because of) the somewhat exaggerated hype, this field is having a major impact in business, industry, and science. It also affords enormous research opportu-

Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. - Zekulin.

Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. - Ferruzza.

Data mining is a decision support process where we look in large data bases for unknown and unexpected patterns of information. - Parsaye

Data Mining is ...

- Decision Trees
- Neural Networks
- Rule Induction
- Nearest Neighbors
- Genetic Algorithms

- Mehta

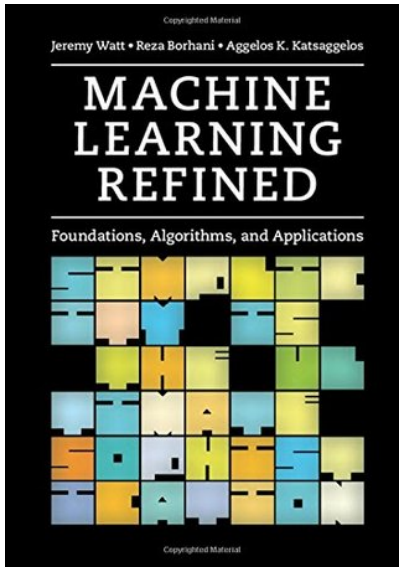
Friedman (1997, Datamining & Statistics, what's the connection)

Machine Learning Courses in a Nutshell

Machine learning (initially) did not need any probabilistic framework.

Consider observations $(y_i, \boldsymbol{x}_i)$, and loss function ℓ and some class of models $\mathcal{M}$. The goal is to solve

$$m^\star \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m(\boldsymbol{x}_i)) \right\}$$



4.2 The logistic regression perspective on the softmax cost

This section describes a common way of both deriving and thinking about the softmax cost function first introduced in Section 4.1.2. Here we will see how the softmax cost naturally arises as a direct approximation of the fundamental counting cost discussed in Section 4.1.5. However the major benefit of this new perspective is in adding a useful geometric viewpoint,[11] that of regression/surface-fitting, to the classification framework in general, and the softmax cost in particular.

[10] We will also see in Section 4.2 how the softmax cost can be thought of as a direct approximation of the counting cost.

[11] Logistic regression can also be interpreted from a *probabilistic* perspective (see Exercise 4.12).

Watt *et al.* (2016, Machine learning refined foundations algorithms and applications)

Probabilistic Foundations and "Statistical Learning"

Between 1960 and 1980 a revolution in statistics occurred: Fisher's paradigm, introduced in the 1920s and 1930s was replaced by a new one. This paradigm reflects a new answer to the fundamental question:

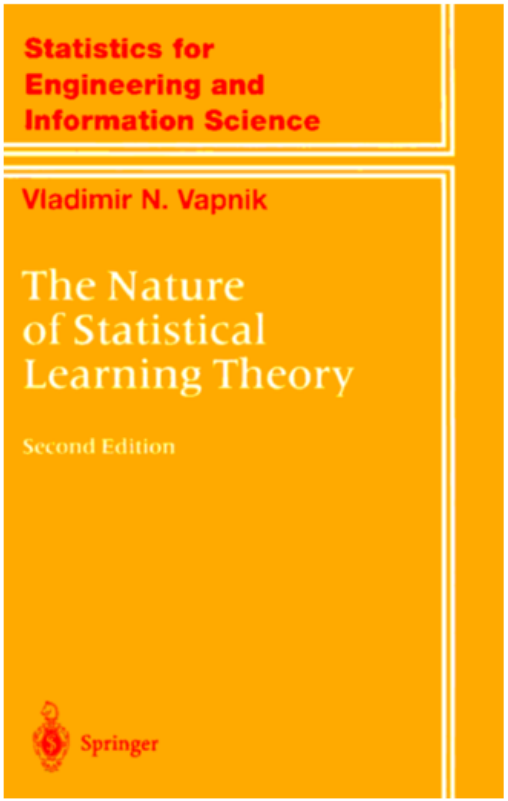
What must one know a priori about an unknown functional dependency in order to estimate it on the basis of observations?

In writing this book I had one more goal in mind: I wanted to stress the practical power of abstract reasoning. The point is that during the last few years at different computer science conferences, I heard reiteration of the following claim:

Complex theories do not work, simple algorithms do.

One of the goals of this book is to show that, at least in the problems of statistical inference, this is not true. I would like to demonstrate that in this area of science a good old principle is valid:

Nothing is more practical than a good theory.



Vapnik (2000, The Nature of Statistical Learning Theory)

Probabilistic Foundations and "Statistical Learning"

RESEARCH CONTRIBUTIONS

Artificial Intelligence and Language Processing

David Waltz Editor

A Theory of the Learnable

L. G. VALIANT

The main contribution of this paper is that it shows that it is possible to design *learning machines* that have all three of the following properties:

1. The machines can provably learn whole classes of concepts. Furthermore, these classes can be characterized.
2. The classes of concepts are appropriate and nontrivial for general-purpose knowledge.
3. The computational process by which the machines deduce the desired programs requires a feasible (i.e., polynomial) number of steps.

More precisely we say that a class X of programs is *learnable* with respect to a given learning protocol if and only if there exists an algorithm A (the deduction procedure) invoking the protocol with the following properties:

1. The algorithm runs in time polynomial in an adjustable parameter h, in the various parameters that quantify the size of the program to be learned, and in the number of variables t.
2. For all programs $f \in X$ and all distributions D over vectors v on which f outputs 1, the algorithm will deduce with probability at least $(1 - h^{-1})$ a program $g \in X$ that never outputs one when it should not but outputs one almost always when it should. In particular, (i) for all vectors v, $g(v) = 1$ implies $f(v) = 1$, and (ii) the sum of $D(v)$ over all v such that $f(v) = 1$, but $g(v) \neq 1$ is at most h^{-1}.

Vailiant (1984, A Theory of Learnable)

Probabilistic Foundations and "Statistical Learning"

Consider a classification problem, $y \in \{-1, +1\}$. The "true" model is m_0 (target), and consider some model $m/$

Let $\mathbb{P}$ denote the (unknown) distribution of $\boldsymbol{X}$'s. The error of m can be written

$$\mathcal{R}_{\mathbb{P},m_0}(\widehat{m}) = \mathbb{P}\big[\widehat{m}(\boldsymbol{X}) \neq m_0(\boldsymbol{X})\big] = \mathbb{P}\big[\{\boldsymbol{x} : \widehat{m}(\boldsymbol{x}) \neq m_0(\boldsymbol{x})\}\big] \text{ where } \boldsymbol{X} \sim \mathbb{P}.$$

Naturally, we can assume that observations $\boldsymbol{x}_i$'s were drawn from $\mathbb{P}$. Empirical risk is

$$\widehat{\mathcal{R}}(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\widehat{m}(\boldsymbol{x}_i) \neq y_i).$$

It is not possible to get a perfect model, we should seek a Probably Almost Correct model. Given ϵ, we want to find m such that $\mathcal{R}_{\mathbb{P},m_0}(\widehat{m}) \leq \epsilon$ with probability $1 - \delta$.

More precisely, we want an algorithm that might lead us to a candidate $\widehat{m}$.

Probabilistic Foundations and "Statistical Learning"

Suppose that $\mathcal{M}$ contains a finite number of models. For any ϵ, δ, $\mathbb{P}$ and m_0, if we have enough observations ($n \geq \epsilon^{-1}\log[\delta^{-1}\|\mathcal{M}\|]$), if

$$m^\star \in \underset{m\in\mathcal{M}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(m(\boldsymbol{x}_i) \neq y_i) \right\}$$

then with probability higher than $1-\delta$, $\mathcal{R}_{\mathbb{P},f}(m^\star) \leq \epsilon$.

Here $n_{\mathcal{M}}(\epsilon,\delta) = \epsilon^{-1}\log[\delta^{-1}\|\mathcal{M}\|]$ is called complexity, and $\mathcal{M}$ is PAC-learnable.

If $\mathcal{M}$ is not finite, the problem is more complicated, it is necessary to define a dimension d - so called VC-dimension - of $\mathcal{M}$, that will be a substitute to $\|\mathcal{M}\|$.