Data Science & Big Data for Actuaries

Arthur Charpentier (Université de Rennes 1 & UQàM)

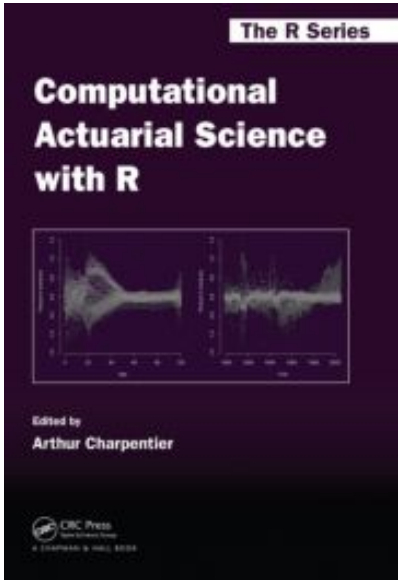
Universitat de Barcelona, April 2016.

http://freakonometrics.hypotheses.org

Data Science & Big Data for Actuaries

Arthur Charpentier (Université de Rennes 1 & UQàM)

Professor, Economics Department, Univ. Rennes 1
In charge of Data Science for Actuaries program, IA
Research Chair *actinfo* (Institut Louis Bachelier)
(previously Actuarial Sciences at UQàM & ENSAE Paristech
actuary in Hong Kong, IT & Stats FFSA)
PhD in Statistics (KU Leuven), Fellow Institute of Actuaries
MSc in Financial Mathematics (Paris Dauphine) & ENSAE
Editor of the freakonometrics.hypotheses.org's blog
Editor of Computational Actuarial Science, CRC



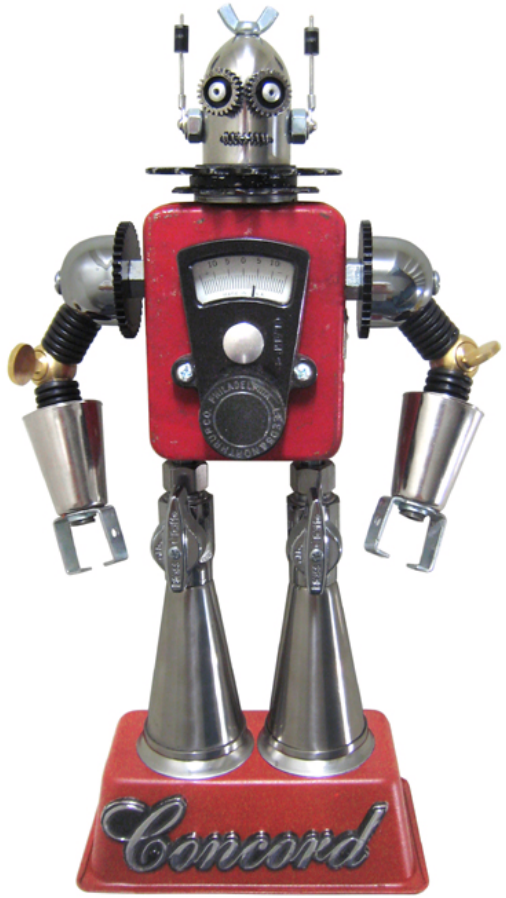
Data

"People use statistics as the drunken man uses lamp posts - for support rather than illumination", Andrew Lang or not

see also Chris Anderson The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, 2008

1. **An Overview on (Big) Data**
2. **Big Data & Statistical/Machine Learning**
3. **Classification Models**
4. **Small Data & Bayesian Philosophy**
5. **Data, Models & Actuarial Science**

Part 1.
An Overview on (Big) Data

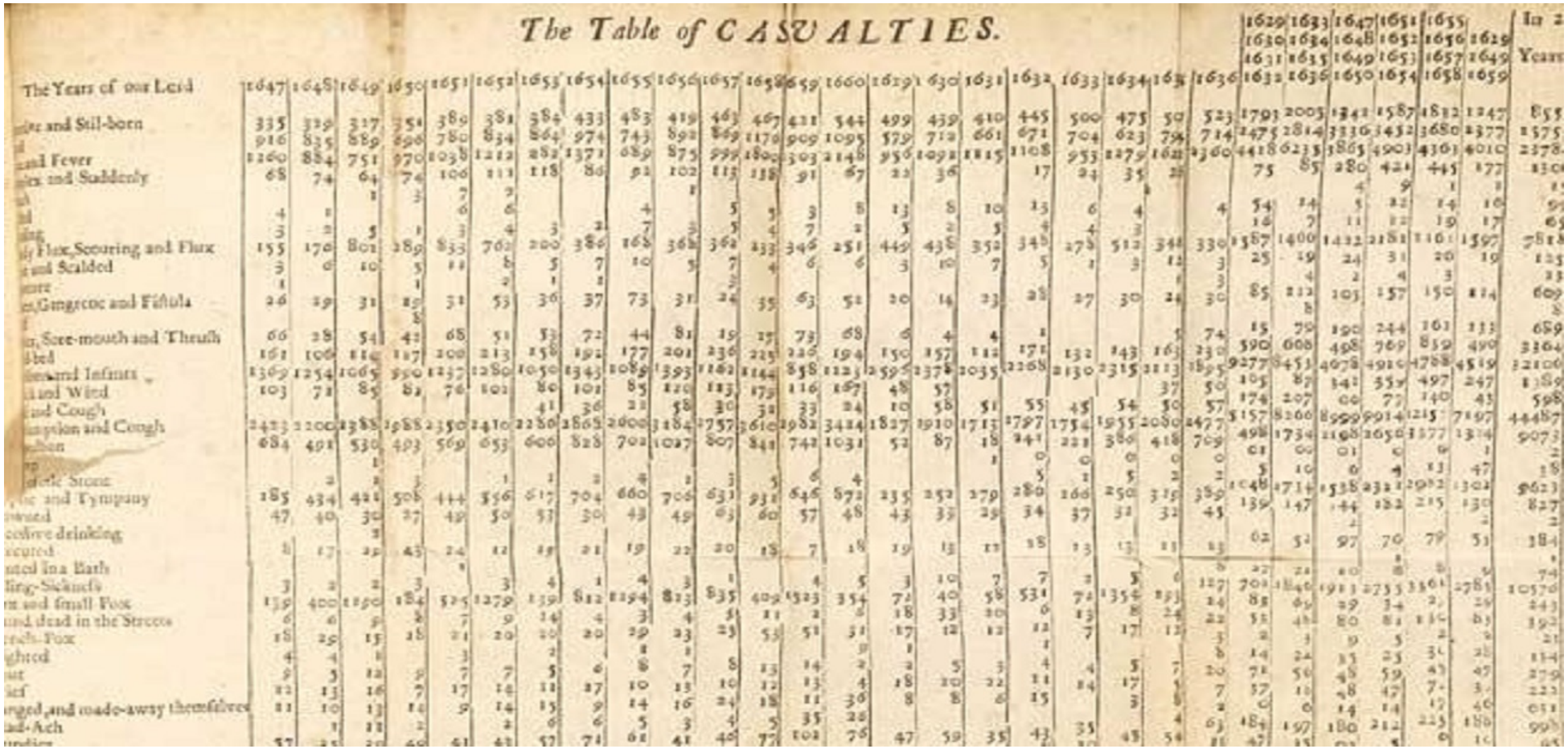


Historical Aspects of Data

Storing Data: Tally sticks, used starting in the Paleolithic area

A tally (or tally stick) was an ancient memory aid device used to record and document numbers, quantities, or even messages.

Historical Aspects of Data



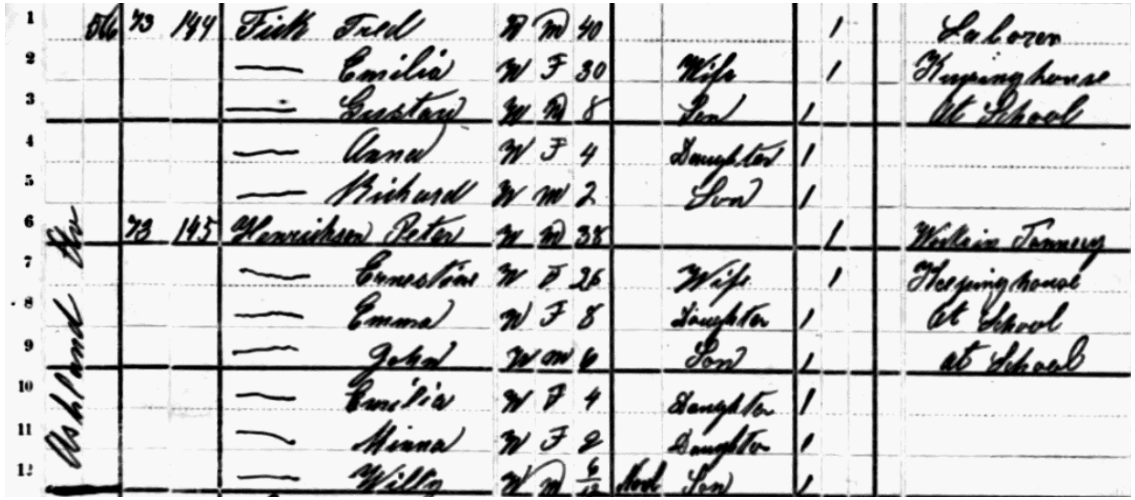
The Table of CASUALTIES.

The Years of our Lord	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1629	1630	1631	1632	1633	1634	1635	1636	1629 1630 1631 1632	1633 1634 1635 1636	1637 1638 1639 1640	1641 1642 1643 1644	1645 1646 1647 1648	1649 1650 1651 1652	1653 1654 1655 1656	1657 1658 1659 1660	In 2 Years
and Still-born	335	339	317	354	389	381	384	433	483	419	463	467	421	544	499	439	410	445	500	475	50	523	1793	2005	1341	1587	1832	1247	855		
	916	835	889	696	780	834	864	974	743	892	869	1170	909	1095	579	712	661	671	704	623	794	714	2475	2814	3330	3432	3680	2377	1575		
and Fever	1260	884	751	970	1038	1212	982	1371	689	875	998	1800	1303	2148	956	1092	1815	1108	953	1279	1622	1360	4418	6233	3863	4903	4301	4010	2378		
and Suddenly	68	74	64	74	100	113	118	86	92	102	113	138	91	87	23	36		17	24	35	25		75	85	280	421	445	177	130		
	4	1		3	6	6			4		5	5	3	8	13	8	10	13	6	4		4	54	14	5	22	14	16	9		
	3	2	5	1	3	4	3	2	7	3	5	4	7	2	3	2	5	4	4	3			16	7	11	12	19	17	6		
Flux, Scouring and Flux	155	170	801	289	833	702	200	386	168	368	362	233	346	251	449	438	352	345	278	512	341	330	1587	1400	1432	2181	1161	1597	781		
and Scalded	3	6	10	5	11	8	5	7	10	5	7	4	6	6	3	10	7	5	1	3	12	3	25	19	24	31	20	19	125		
	1			1		2	1	1			3											3		4	2	4	3		33		
Gangrene and Fistula	26	29	31	29	31	53	36	37	73	31	24	35	63	51	20	14	23	28	27	30	24	30	85	212	103	157	150	114	609		
				8																				8					8		
Sore-mouth and Thrush	66	28	54	42	68	51	53	72	44	81	19	25	73	68	6	4	4	1			3	74	15	79	190	244	161	133	689		
Bed	161	100	114	127	200	213	158	192	177	201	236	225	220	194	150	157	112	171	132	143	163	230	590	608	498	769	859	490	3364		
and Infants	1369	1254	1065	950	1237	1280	1050	1343	1089	1393	1162	1144	858	1123	1595	1378	2035	2268	2130	2315	2213	1895	9277	9453	4078	4916	4788	4519	32100		
and Wind	103	72	85	82	76	102	80	102	85	120	123	179	116	167	48	57					37	50	174	207	141	353	497	247	1389		
and Cough							41	36	22	58	30	32	33	24	10	58	51	55	45	54	30	57	3157	8200	8999	9914	1215	7197	44487		
and Cough	2423	2200	2388	2588	2350	2410	2280	2865	2000	3184	2757	3610	2982	3424	1827	1910	1715	1797	1754	1955	2030	2477	4498	1734	2168	2650	2377	1314	9073		
	684	491	530	493	569	653	600	828	702	1027	807	841	742	1031	52	87	18	241	221	380	418	709	61	60	61	6	9	1	2		
			1														1	0	0	0	0	0	5	10	6	4	13	47	18		
	2	2	2	3	1	2	2	8	2	3	5	6	4					5	1	5	2	2	1048	1714	1538	2321	2902	1302	9623		
and Tympany	185	434	424	508	444	356	617	704	660	706	631	931	646	572	235	252	279	289	266	250	335	389	139	147	144	182	215	130	827		
	47	40	30	27	48	50	53	30	43	49	63	60	57	48	43	33	25	34	37	32	32	43			2				2		
			3																												
	8	17	22	43	24	12	18	21	19	22	20	18	7	18	19	13	11	18	13	13	11	13	62	52	97	70	79	51	384		
																									1				1		
	3	2	2	3		3	4	1	4	3	1		4	5	3	10	7	7	2	5	6	8	22	22	40	8	8	5	74		
	139	400	1250	184	525	1278	139	812	2194	823	835	409	1523	354	72	40	58	534	72	1354	253	127	703	1840	1913	2733	3361	2781	1057		
	6	6	5	8	7	9	14	4	3	4	5	11	2	6	18	33	20	6	13	8	24	24	88	69	29	34	2	29	243		
	18	25	15	28	21	20	20	20	29	23	23	53	51	31	17	12	12	12	7	17	12	22	55	48	80	84	116	63	392		
									1													3	2	3	9	5	2	2	21		
	4	4	8		3		2		8	1					1			1				6	14	24	33	23	34	28	134		
	9	3	12	8	7	7	5	6	8	7	8	13	14	2	2	5	3	4	4	5	7	20	72	56	48	59	43	47	279		
	22	13	16	7	17	18	11	27	10	13	10	12	13	4	18	20	22	11	14	17	5	7	37	11	48	47	74	31	223		
ged, and made away themselves	22	10	13	14	9	14	15	9	14	16	24	18	11	36	8	8	6	15		3	8	2	0	0	14	14	13	46	611		
-Ach		1	11	2		2	6	6	5	3	4	5	35	28								63	184	197	180	212	225	186	998		
	57	35	20	60	61	43	57	74	61	41	40	72	102	76	47	59	35	43	35	43	54	81	47	35	91	5	9	16	95		

Collecting Data: John Graunt conducted a statistical analysis to curb the spread of the plage, in Europe, in 1663

Historical Aspects of Data

Data Manipulation: Herman Hollerith created a Tabulating Machine that uses punch carts to reduce the workload of US Census, in 1881, see 1880 Census, n =50 million Americans.



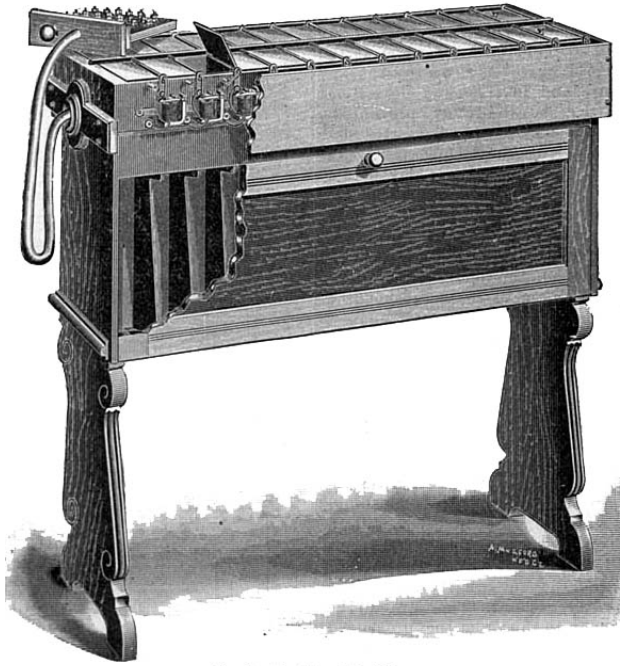


Fig. 3.—Sorting Machine.

Hollerith's Electric Sorting and Tabulating Machine.

Historical Aspects of Data

Survey and Polls: 1936 US elections

Literary Digest Poll based on 2.4 million readers

A. Landon: 57% vs. F.D. Roosevelt: 43%


Today | AMERICA SPEAKS | Next Sunday

THE NATIONAL WEEKLY POLL of PUBLIC OPINION

Institute Forecasts the Re-election of Franklin D. Roosevelt, Gives Him 54% of Popular Vote, Minimum of 315 Electors

Major Party Percent Is 55.7; New York in F.D.R. 'Sure' Column

Election Forecast

Election Will Test Clashing Poll Methods

George Gallup sample of about 50,000 people

A. Landon: 44% vs. F.D. Roosevelt: 56%

Actual results

A. Landon: 38% vs. F.D. Roosevelt: 62%

Sampling techniques, polls, predictions based on small samples

Historical Aspects of Data

Allen-Scott Report

Data Center Plan Called Privacy Invasion

By ROBERT S. ALLEN
and PAUL SCOTT

WASHINGTON — A special White House task force is recommending the creation of a federal data center which eventually could have a comprehensive file on every man, woman and child in the country.

Now under study in inner administration circles, the still-secret report advocates the gradual transfer of all governmental records and statistics to magnetic computer tape, which would be turned over to a newly-created agency that would function as a general data center.

The computerized information would be available, at the push of a button, to a wide range of government authorities.

Estimated cost of the pro-

cial Security, census data, medical, credit and criminal reports.

"Comprehensive information of this kind, centralized in one agency," says Gallagher, "could constitute a highly dangerous dossier bank. Such an agency would be a distinct departure from our American tradition."

Subcommittee investigators have ascertained that the task force's report states that a vast accumulation of government records already is on computer tape and could be turned over to the proposed general data center immediately. Listed as among these available files are:

Internal Revenue Service — 742 million personal and corporate tax returns.

Defense Department — 14

the most intimate information, the investigators learned, are freely passed around among agencies. Graphically illustrative of this practice and its harsh consequences are the following two instances:

A teenager visiting Washington stayed with an uncle, at his mother's suggestion. During the night the boy was sexually assaulted by the uncle. Years later, as a Phi Beta Kappa graduate from a leading Eastern university, the boy applied for a job with the National Security Agency. During a required lie detector test he told about the assault. His frank admission cost him the desired job.

But that wasn't all. This affair, in which he was an inno-

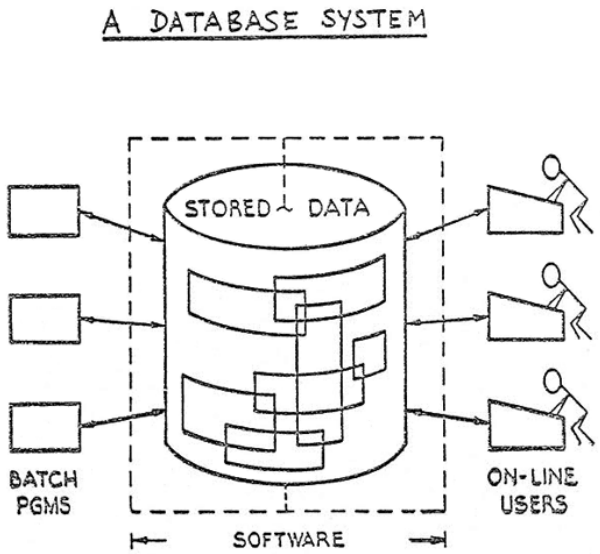
Data Center: The US Government plans the world's first data center to store 742 million tax returns and 175 million sets of fingerprints, in 1965.

Historical Aspects of Data

Historical Aspects of Data

Data Manipulation: Relational Database model developed by Edgar F. Codd

See Relational Model of Data for Large Shared Data Banks, Codd (1970)



Considered as a major breakthrough for users and machine designers

Data or tables are thought as a matrix composted of intersecting rows and columns, each columns being attributes.

Tables are related to each other through a common attribute.

Concept of relational diagrams

The Two Cultures

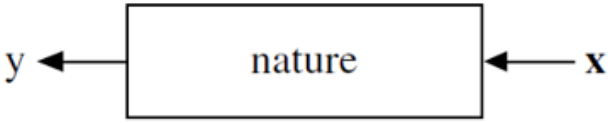
'The Two Cultures', see Breiman (2001)

- Data Modeling (statistics, econometrics)
- Algorithmic Modeling (computational & algorithmics)

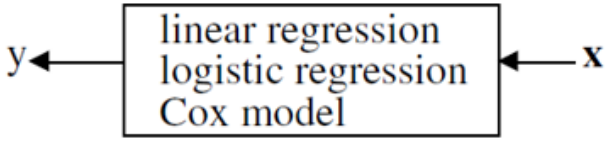
'**Big Data** Dynamic Factor Models for Macroeconomic Measurementand Forecasting', Diebold (2000)

Statistical Science
2001, Vol. 16, No. 3, 199–231

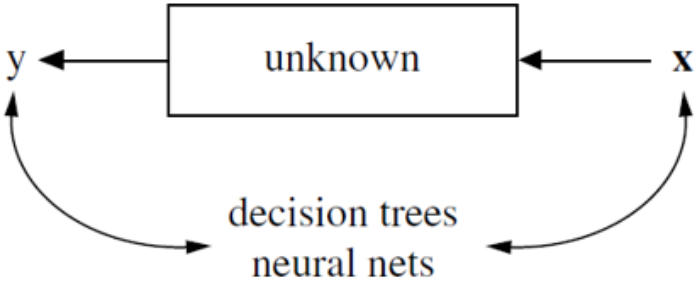
Statistical Modeling: The Two Cultures



The Data Modeling Culture



The Algorithmic Modeling Culture

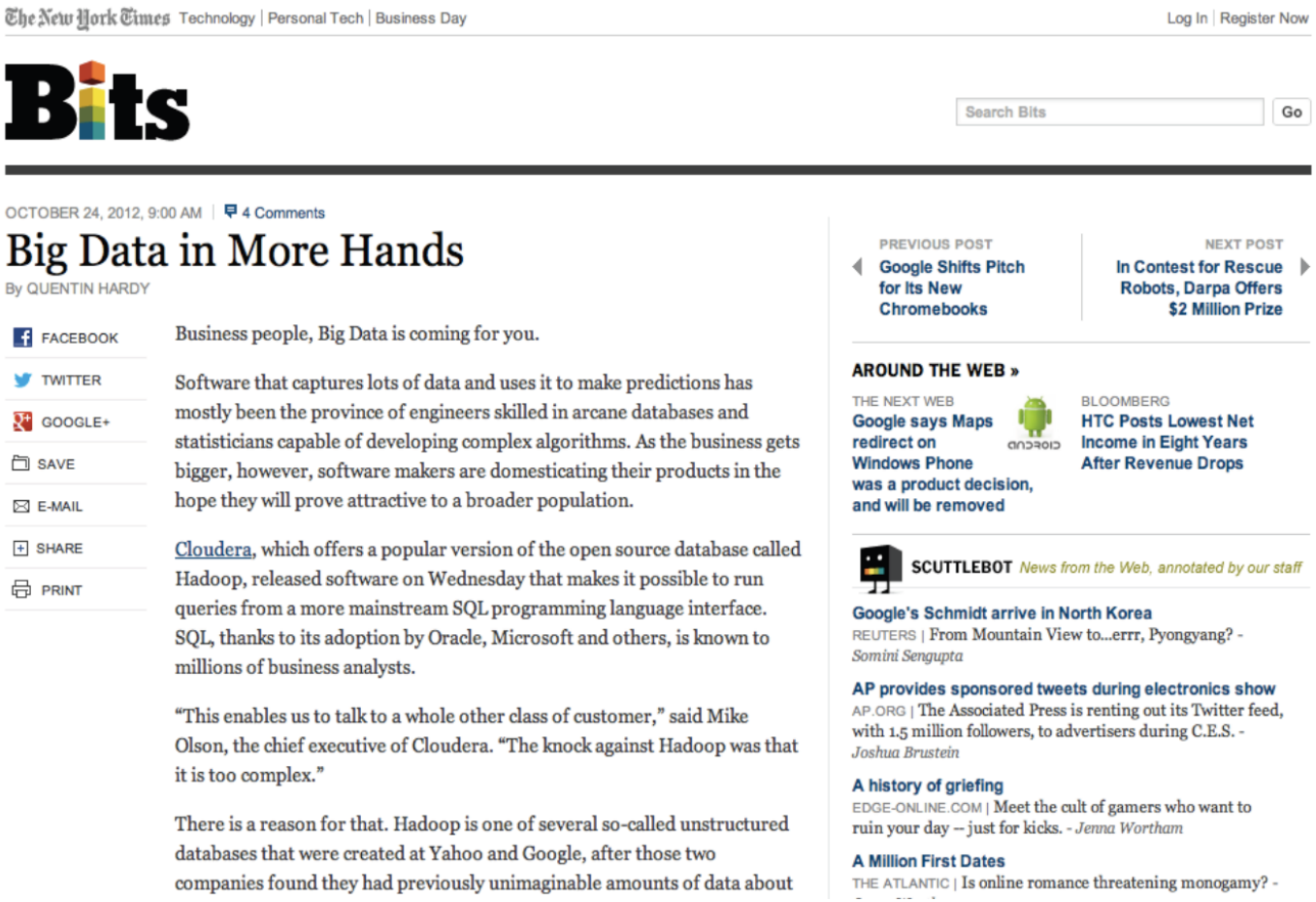


And the XIXth Century...

Nature's special issue on Big Data, Nature (2008) and many of business journals

And the XIXth Century...

Techology changed, HDFS (Hadoop Distribution File System), MapReduce



The New York Times Technology | Personal Tech | Business Day

Log In | Register Now

Bits

Search Bits Go

OCTOBER 24, 2012, 9:00 AM | 4 Comments

Big Data in More Hands

By QUENTIN HARDY

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT

Business people, Big Data is coming for you.

Software that captures lots of data and uses it to make predictions has mostly been the province of engineers skilled in arcane databases and statisticians capable of developing complex algorithms. As the business gets bigger, however, software makers are domesticating their products in the hope they will prove attractive to a broader population.

Cloudera, which offers a popular version of the open source database called Hadoop, released software on Wednesday that makes it possible to run queries from a more mainstream SQL programming language interface. SQL, thanks to its adoption by Oracle, Microsoft and others, is known to millions of business analysts.

"This enables us to talk to a whole other class of customer," said Mike Olson, the chief executive of Cloudera. "The knock against Hadoop was that it is too complex."

There is a reason for that. Hadoop is one of several so-called unstructured databases that were created at Yahoo and Google, after those two companies found they had previously unimaginable amounts of data about

PREVIOUS POST
Google Shifts Pitch for Its New Chromebooks

NEXT POST
In Contest for Rescue Robots, Darpa Offers $2 Million Prize

AROUND THE WEB »

THE NEXT WEB
Google says Maps redirect on Windows Phone was a product decision, and will be removed

BLOOMBERG
HTC Posts Lowest Net Income in Eight Years After Revenue Drops

SCUTTLEBOT *News from the Web, annotated by our staff*

Google's Schmidt arrive in North Korea
REUTERS | From Mountain View to...errr, Pyongyang? - *Somini Sengupta*

AP provides sponsored tweets during electronics show
AP.ORG | The Associated Press is renting out its Twitter feed, with 1.5 million followers, to advertisers during C.E.S. - *Joshua Brustein*

A history of griefing
EDGE-ONLINE.COM | Meet the cult of gamers who want to ruin your day -- just for kicks. - *Jenna Wortham*

A Million First Dates
THE ATLANTIC | Is online romance threatening monogamy? -

And the XIXth Century...

Data changed, because of the digital/numeric revolution, see Gartner's 3V (Volume, Variety, Velocity), see Gartner.

The New York Times
SundayReview | The Opinion Pages

Search All NYTimes.com Go

Capital One

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION ARTS STYLE TRAVEL JOBS REAL ESTATE AUTOS

NEWS ANALYSIS

The Age of Big Data

By STEVE LOHR
Published: February 11, 2012 | 82 Comments

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

Enlarge This Image

Chad Hagen

Multimedia

Siera = 6.145−16.986×(S
−1.858×((GB-FB-PU)÷PA
×(((GB-FB-PU)÷PA)^2)+1
÷PA)−5.195×(BB÷PA)×((

Graphic

Play (Data-Driven) Ball!

Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers," says Ms. Zhou, whose job as a data analyst suits her skills.

To exploit the data flood, America will need many more like her. A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired.

The impact of data abundance extends well beyond business. Justin Grimmer, for example, is one of the new breed of political scientists. A 28-year-old assistant

RECOMMEND
TWITTER
LINKEDIN
COMMENTS (82)
E-MAIL
PRINT
REPRINTS
SHARE

12 YEARS A SLAVE
WATCH THE TRAILER

Log in to see what your friends are sharing on nytimes.com. Privacy Policy | What's This?
Log In With Facebook

What's Popular Now

Felony Counts for 2 in Suicide of Bullied 12-Year-Old

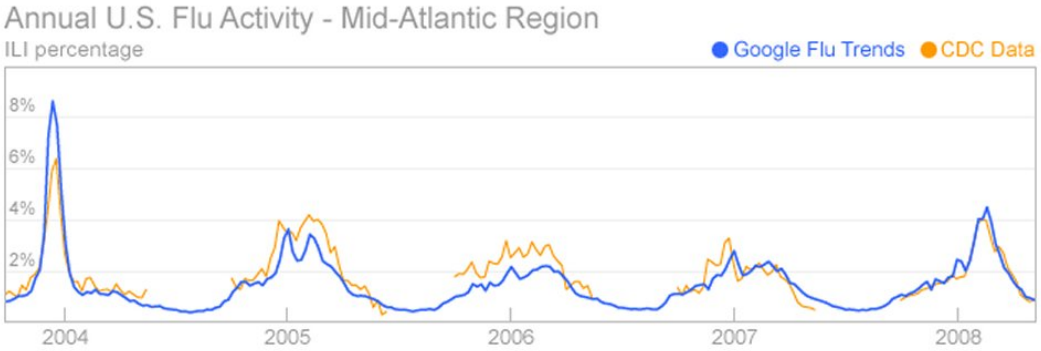
Senate Women Lead in Effort to Find Accord

And the XIXth Century...

Business Intelligence, transversal approach

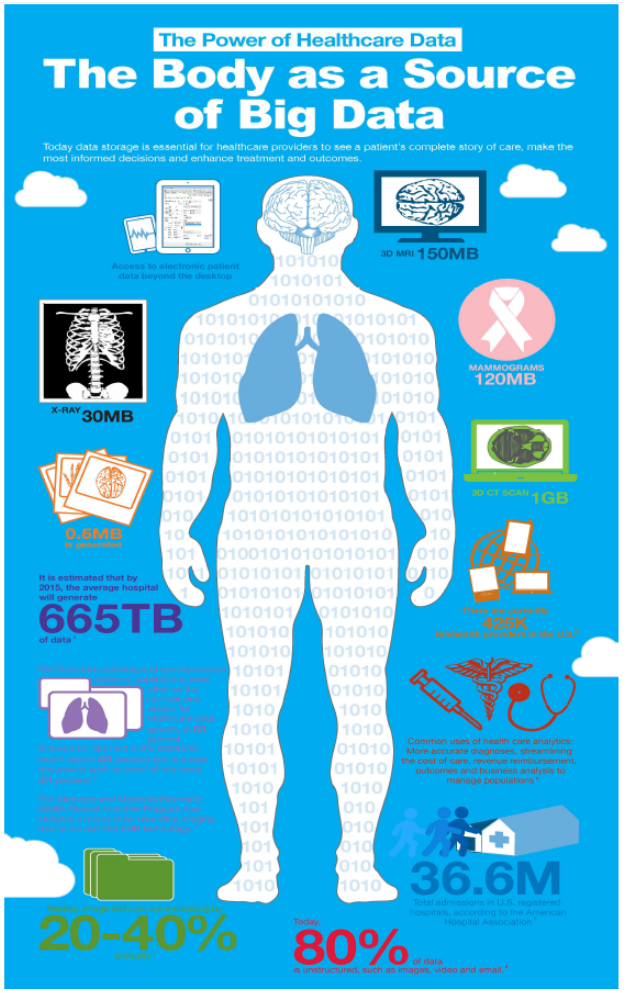
Big Data & (Health) Insurance

Example: popular application, Google Flu Trend



See also Lazer *et al.* (2014)

But much more can be done on an individual level.



Big Data & Computational Issues

parallel computing is a necessity*

CPU Central Processing Unit, the heart of the computer

RAM Random Access Memory non-persistent memory

HD Hard Drive persistent memory

Practical issues: CPU can be fast, but finite speed;

RAM is non persistent, fast but slow vs. HD is persistent, slow but big

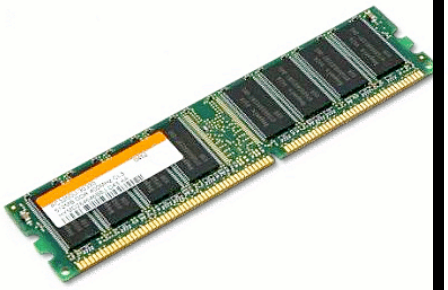
How could we measure speed: Latency and performance

Latency is a time interval between the stimulation and response (e.g. 10ms to read the first bit)

Performance is the number of operations per second (e.g. 100Mb/sec)

Example Read one file of 100Mb $\sim$ 1.01sec.

Example Read 150 files of 1b $\sim$ 0.9sec.



* thanks to David Sibaï for this section.

Big Data & Computational Issues

Standard PC :

CPU : 4 core, 1ns latenty

RAM : 32 or 64 Gb, 100ns latency, 20Gb/sec

HD : 1 Tb, 10ms latency, 100Mo/sec

How long does it take ?

e.g. count *spaces* in a 2Tb text file

about 2.10^{12} operations (comparaison)

File on the HD, 100Mb/sec $\sim 2.10^4$ sec $\sim$ 6 hours

Big Data & Computational Issues

Why not parallelize ? between machines

Spread data on 10 blocks of 200Gb, each machine count spaces, then sum the 10 totals... should be 10 times faster.

Many machines connected, in a datacenter

Alternative: use more cores in the CPU (2, 4, 16 cores, e.g.)

A CPU is multi-tasks, and it could be possible to vectorize. E.g. summing n numbers takes $O(n)$ operations,

Example $a_1 + b_1$, $a_2 + b_2$, $\cdots$, $a_n + b_n$ takes n nsec.

But it is possible to use SIMD (single instruction multiple data)

Example $\boldsymbol{a} + \boldsymbol{b} = (a_1, \cdots, a_n) + (b_1, \cdots, b_n)$ take 1 nsec.

Big Data & Computational Issues

Alternatives to standard PC material

Games from the 90s, more and more 3d viz, based on more and more computations

GPU Graphical Processing Unit that became GPGPU General Purpose GPU

Hundreds of small processors, slow, high specialized (and dedicated to simple computations)

Difficult to use (needs of computational skills) but more and more libraries

Complex and slow communication CPU - RAM - GPU

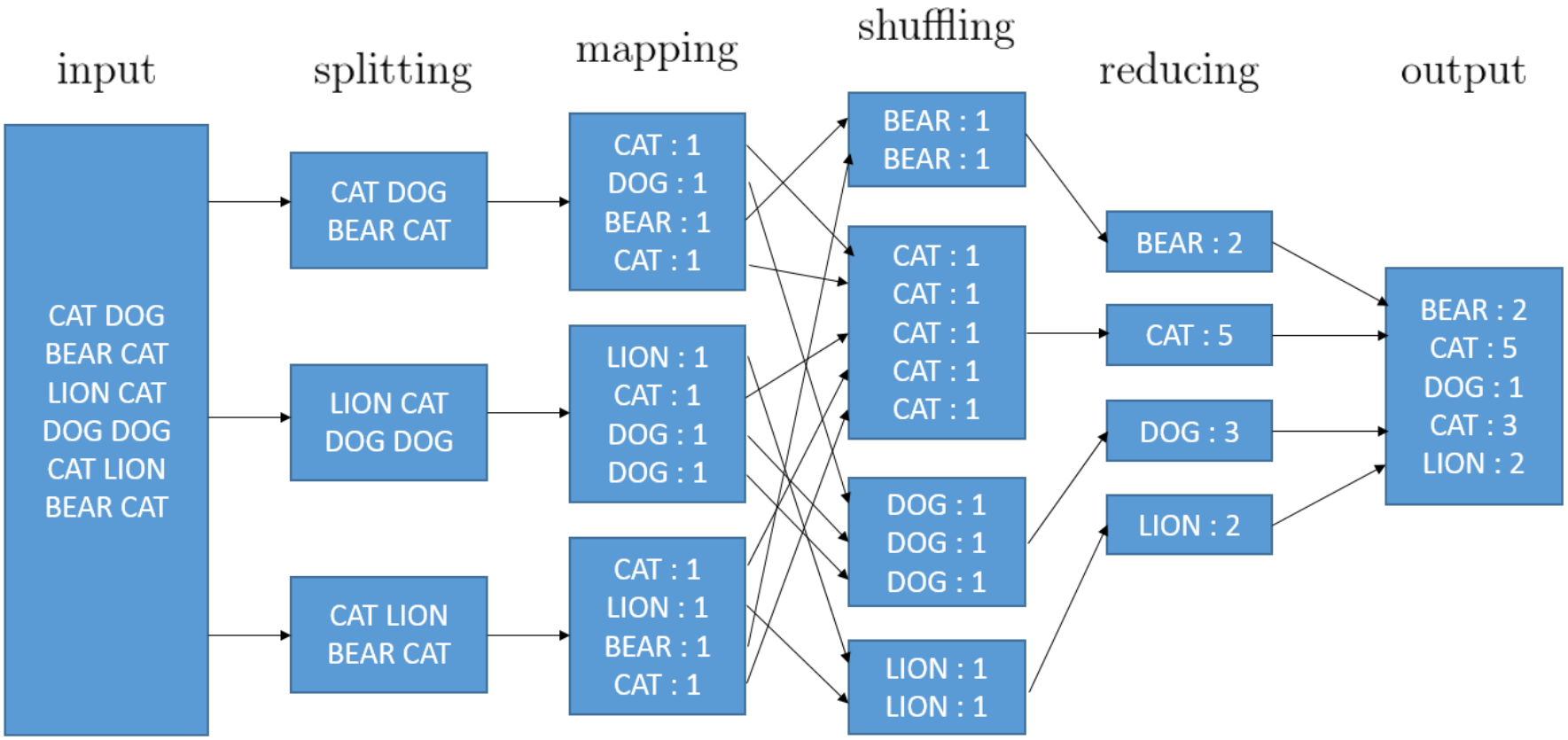
Sequential code is extremely slow, but highly parallelized

Interesting for Monte Carlo computations

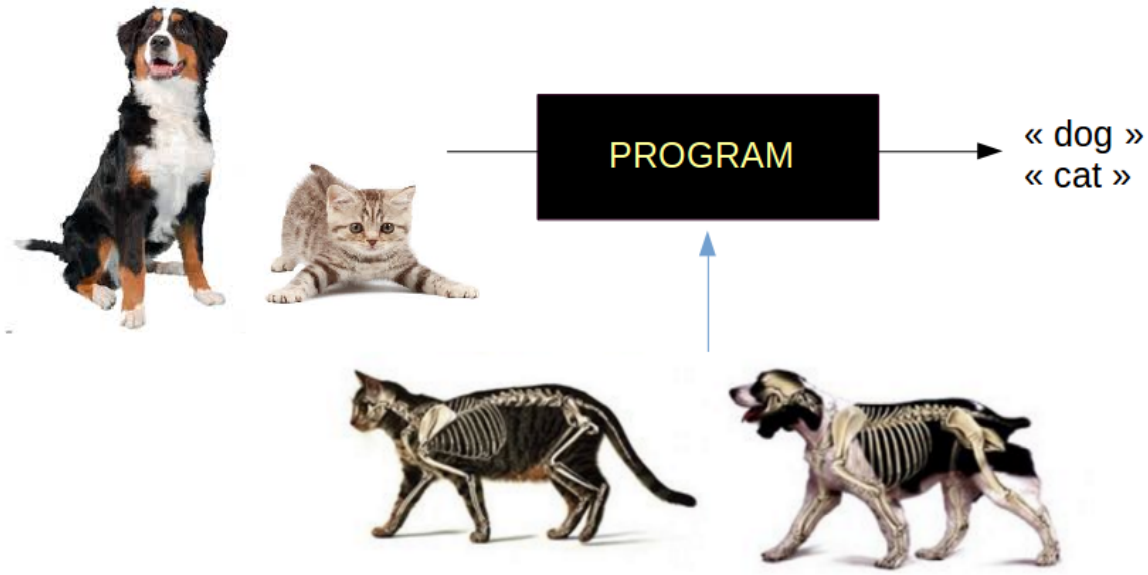
E.g. pricing of Variable Annuities

Big Data & Computational Issues

A parallel algorithm is a computational strategy which divide a target computation into independent part, and assemble them so as to obtain the target computation. E.g. Couting words with MapReduce



Data, (deep) Learning & AI



What can we do with those data?

Part 2.
Big Data and Statistical/Machine Learning

Statistical Learning and Philosophical Issues

From Machine Learning and Econometrics, by Hal Varian :

"**Machine learning** use data to predict some variable as a function of other covariables,

- may, or may not, care about insight, importance, patterns
- may, or may not, care about inference (how y changes as some $\boldsymbol{x}$ change)

Econometrics use statistical methodes for prediction, inference and causal modeling of economic relationships

- hope for some sort of insight (inference is a goal)
- in particular, causal inference is goal for decision making."

$\rightarrow$ machine learning, 'new tricks for econometrics'

Statistical Learning and Philosophical Issues

Remark machine learning can also learn from econometrics, especially with non i.i.d. data (time series and panel data)

Remark machine learning can help to get better predictive models, given good datasets. No use on several data science issues (e.g. selection bias).

non-supervised vs. supervised techniques

Non-Supervised and Supervised Techniques

Just $\boldsymbol{x}_i$'s, here, no y_i: unsupervised.

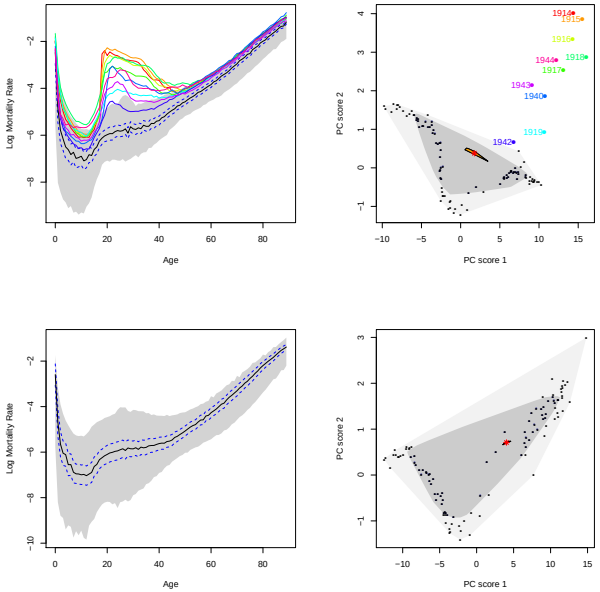
Use principal components to reduce dimension: we want d vectors $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_d$ such that

$$\boldsymbol{x}_i \sim \sum_{j=1}^{d} \omega_{i,j} \boldsymbol{z}_j \text{ or } \boldsymbol{X} \sim \boldsymbol{Z}\boldsymbol{\Omega}^\intercal$$

where $\boldsymbol{\Omega}$ is a $k \times d$ matrix, with $d < k$.

First Compoment is $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{\omega}_1$ where

$$\boldsymbol{\omega}_1 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ \|\boldsymbol{X} \cdot \boldsymbol{\omega}\|^2 \right\} = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ \boldsymbol{\omega}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{\omega} \right\}$$



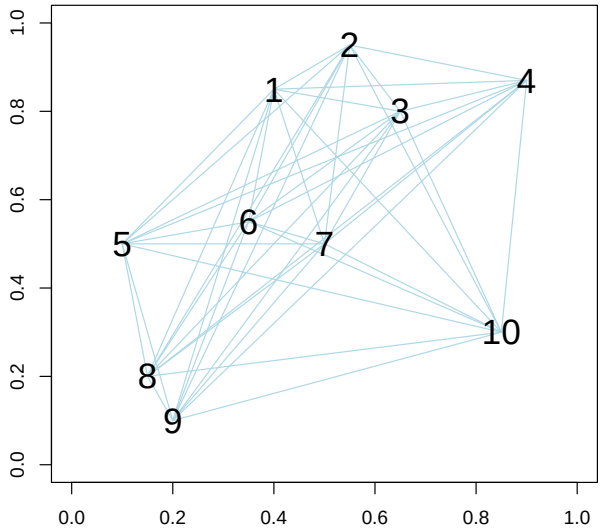
Second Compoment is $\boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{\omega}_2$ where

$$\boldsymbol{\omega}_2 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ \|\widetilde{\boldsymbol{X}}^{(1)} \cdot \boldsymbol{\omega}\|^2 \right\} \text{ where } \widetilde{\boldsymbol{X}}^{(1)} = \boldsymbol{X} - \underbrace{\boldsymbol{X}\boldsymbol{\omega}_1}_{\boldsymbol{z}_1} \boldsymbol{\omega}_1^\intercal$$

Unsupervised Techniques: Cluster Analysis

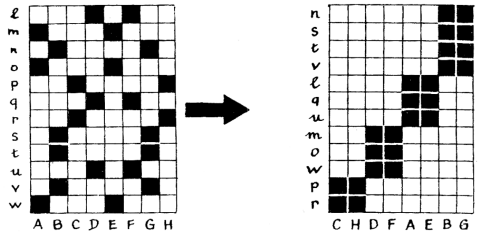
Data : $\{\boldsymbol{x}_i = (x_{1,i}, x_{2,i}), i = 1, \cdots, n\}$

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



the distance is between clusters, not (only) individuals,

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



for some (standard) distance d, e.g. Euclidean (ℓ_2), Manhattan (ℓ_1), Jaccard, etc.

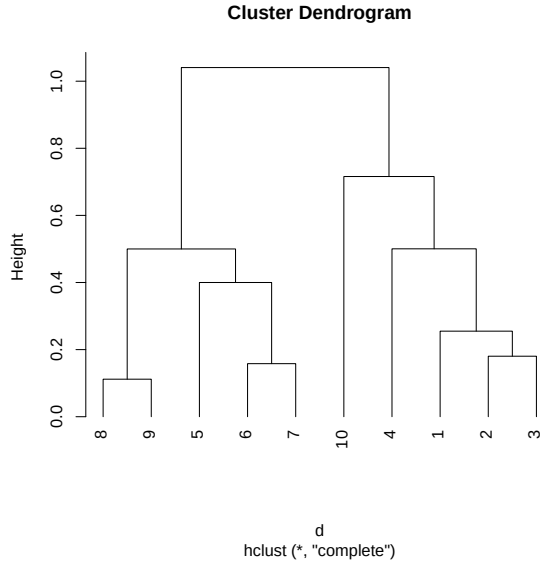
See also Bertin (1967).

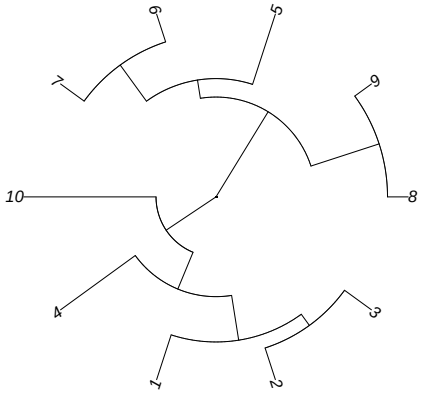
Unsupervised Techniques: Cluster Analysis

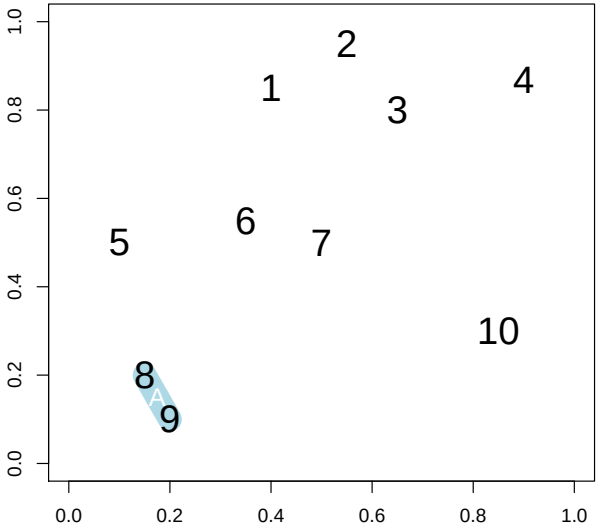
Data : $\{\boldsymbol{x}_i = (x_{1,i}, x_{2,i}), i = 1, \cdots, n\}$

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

The standard output is usually a dendrogram.





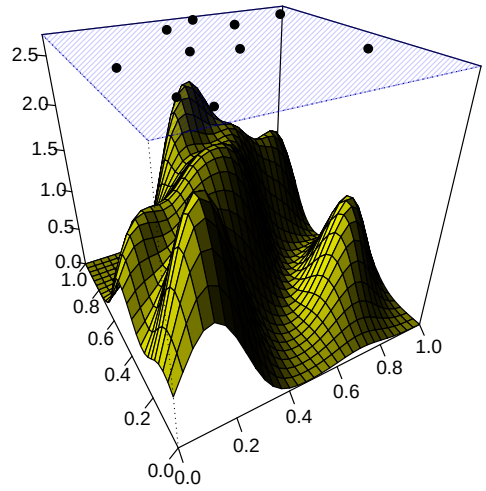


Unsupervised Techniques

Data : $\{\boldsymbol{x}_i = (x_{1,i}, x_{2,i}), i = 1, \cdots, n\}$

$\boldsymbol{x}_i$'s are observations from i.i.d random variables $\boldsymbol{X}_i$ with distribution $F_{\boldsymbol{p},\boldsymbol{\theta}}$,

$$F_{\boldsymbol{p},\boldsymbol{\theta}}(\boldsymbol{x}) = \underbrace{p_1 \cdot F_{\boldsymbol{\theta}_1}(\boldsymbol{x})}_{\text{Cluster 1}} + \underbrace{p_2 \cdot F_{\boldsymbol{\theta}_2}(\boldsymbol{x})}_{\text{Cluster 2}} + \cdots$$



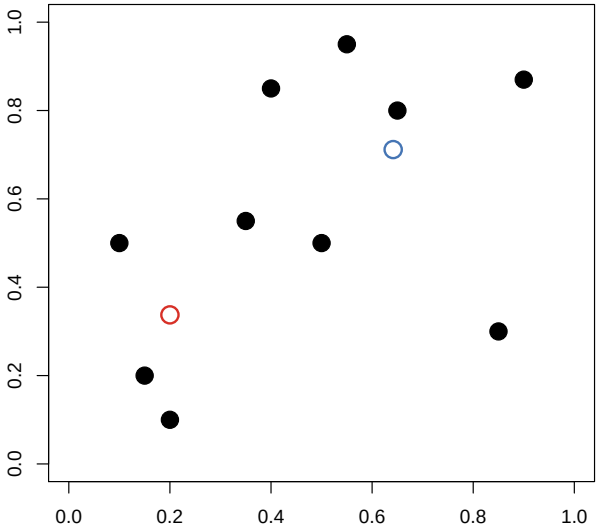
E.g. $F_{\boldsymbol{\theta}_k}$ is the c.d.f. of a $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ distribution.

Unsupervised Techniques

Data : $\{\boldsymbol{x}_i = (x_{1,i}, x_{2,i}), i = 1, \cdots, n\}$

iterative procedure:

1. start with k points $\boldsymbol{z}_1, \cdots \boldsymbol{z}_k$
2. cluster c_j are $\{d(\boldsymbol{x}_i, \boldsymbol{z}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{z}_{j'}), j' \neq j\}$
3. $\boldsymbol{z}_j = \overline{\boldsymbol{x}}_{\mathrm{c}_j}$



See Steinhaus (1957)) or Lloyd (1957))
But curse of dimensionality, unhelpful in high dimension

Datamining, Explantory Analysis, Regression, Statistical Learning, Predictive Modeling, etc

In statistical learning, data are approched with little priori information.

In regression analysis, see Cook & Weisberg (1999)

> The primary goal in a regression analysis is to understand, as far as possible with the available data, how the conditional distribution of the response y varies across subpopulations determined by the possible values of the predictor or predictors. Since this is the central idea, it will be helpful to have a conve-

i.e. we would like to get the distribution of the response variable Y conditioning on one (or more) predictors $\boldsymbol{X}$.

Consider a regression model, $y_i = m(\boldsymbol{x_i}) + \varepsilon_i$, where ε_i 's are i.i.d. $\mathcal{N}(0, \sigma^2)$, possibly linear $y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i$, where ε_i's are (somehow) unpredictible.

Machine Learning and 'Statistics'

Machine learning and statistics seem to be very similar, they share the same goals—they both focus on data modeling—but their methods are affected by their cultural differences.

"The goal for a statistician is to predict an interaction between variables with some degree of certainty (we are never 100% certain about anything). Machine learners, on the other hand, want to build algorithms that predict, classify, and cluster with the most accuracy, see Why a Mathematician, Statistician & Machine Learner Solve the Same Problem Differently

Machine learning methods are about algorithms, more than about asymptotic statistical properties.

Validation is not based on mathematical properties, but on properties out of sample: we must use a training sample to train (estimate) model, and a testing sample to compare algorithms (hold out technique).

Goldilock Principle: the Mean-Variance Tradeoff

In statistics and in machine learning, there will be parameters and meta-parameters (or tunning parameters. The first ones are estimated, the second ones should be chosen.

See Hill estimator in extreme value theory. X has a Pareto distribution - with index ξ - above some threshold u if

$$\mathbb{P}[X > x | X > u] = \left(\frac{u}{x}\right)^{\frac{1}{\xi}} \text{ for } x > u.$$

Given a sample $\boldsymbol{x}$, consider the Pareto-QQ plot, i.e. the scatterplot

$$\left\{ -\log\left(1 - \frac{i}{n+1}\right), \log x_{i:n} \right\}_{i=n-k,\cdots,n}$$

for points exceeding $X_{n-k:n}$. The slope is ξ, i.e.

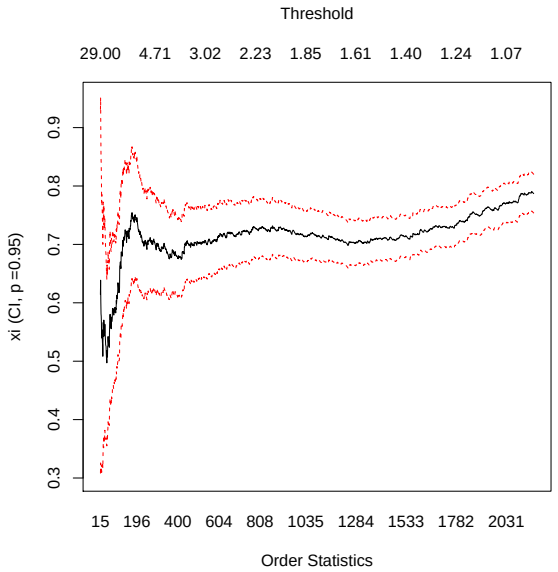
$$\log X_{n-i+1:n} \approx \log X_{n-k:n} + \xi \left(-\log \frac{i}{n+1} - \log \frac{n+1}{k+1} \right)$$

Goldilock Principle: the Mean-Variance Tradeoff

Hence, consider estimator

$$\widehat{\xi}_k = \frac{1}{k}\sum_{i=0}^{k-1} \log x_{n-i:n} - \log x_{n-k:n}.$$

k is the number of large observations, in the upper tail.



Standard mean-variance tradeoff,

- k large: bias too large, variance too small
- k small: variance too large, bias too small

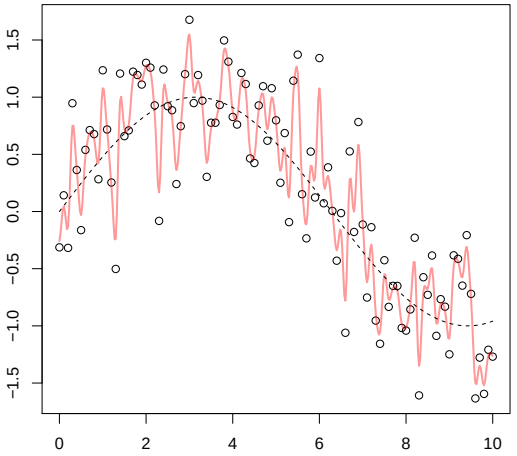
Goldilock Principle: the Mean-Variance Tradeoff

Same holds in kernel regression, with bandwidth h (length of neighborhood)

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^n K_h(x - x_i) y_i}{\sum_{i=1}^n K_h(x - x_i)}$$

since

$$\mathbb{E}(Y|X = x) = \int \frac{f(x, y) \cdot y}{f(x)} dy$$



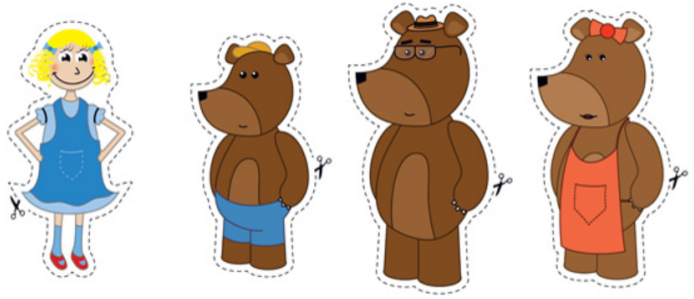
Standard mean-variance tradeoff,

- h large: bias too large, variance too small
- h small: variance too large, bias too small

Goldilock Principle: the Mean-Variance Tradeoff

More generally, we estimate $\widehat{\boldsymbol{\theta}}_h$ or $\widehat{m}_h(\cdot)$

Use the mean squared error for $\widehat{\boldsymbol{\theta}}_h$



$$\mathbb{E}\left[\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_h\right)^2\right]$$

or mean integrated squared error $\widehat{m}_h(\cdot)$,

$$\mathbb{E}\left[\int\left(m(\boldsymbol{x})-\widehat{m}_h(\boldsymbol{x})\right)^2 d\boldsymbol{x}\right]$$

In statistics, derive an asymptotic expression for these quantities, and find $h^\star$ that minimizes those.

Goldilock Principle: the Mean-Variance Tradeoff

For kernel regression, the MISE can be approximated by

$$\frac{h^4}{4}\left(\int \boldsymbol{x}^\intercal \boldsymbol{x} K(\boldsymbol{x})d\boldsymbol{x}\right)^2 \int \left(m''(\boldsymbol{x}) + 2m'(\boldsymbol{x})\frac{f'(\boldsymbol{x})}{f(\boldsymbol{x})}\right) d\boldsymbol{x} + \frac{1}{nh}\sigma^2 \int K^2(\boldsymbol{x})d\boldsymbol{x} \int \frac{d\boldsymbol{x}}{f(\boldsymbol{x})}$$

where f is the density of $\boldsymbol{x}$'s. Thus the optimal h is

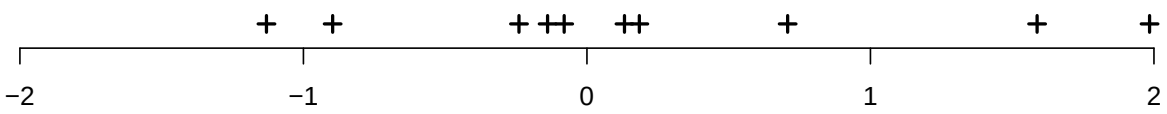
$$h^\star = n^{-\frac{1}{5}} \left(\frac{\sigma^2 \int K^2(\boldsymbol{x})d\boldsymbol{x} \int \frac{d\boldsymbol{x}}{f(\boldsymbol{x})}}{\left(\int \boldsymbol{x}^\intercal \boldsymbol{x} K(\boldsymbol{x})d\boldsymbol{x}\right)^2 \int \left(\int m''(\boldsymbol{x}) + 2m'(\boldsymbol{x})\frac{f'(\boldsymbol{x})}{f(\boldsymbol{x})}\right)^2 d\boldsymbol{x}} \right)^{\frac{1}{5}}$$

(hard to get a simple rule of thumb... up to a constant, $h^\star \sim n^{-\frac{1}{5}}$)

Use bootstrap, or cross-validation to get an optimal h

Randomization is too important to be left to chance!

Bootstrap (resampling) algorithm is very important (nonparametric monte carlo)



→ data (and not model) driven algorithm

Randomization is too important to be left to chance!

Consider some sample $\boldsymbol{x} = (x_1, \cdots, x_n)$ and some statistics $\widehat{\theta}$. Set $\widehat{\theta}_n = \widehat{\theta}(\boldsymbol{x})$

Jackknife used to reduce bias: set $\widehat{\theta}_{(-i)} = \widehat{\theta}(\boldsymbol{x}_{(-i)})$, and $\tilde{\theta} = \dfrac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_{(-i)}$

$$\text{If } \mathbb{E}(\widehat{\theta}_n) = \theta + O(n^{-1}) \text{ then } \mathbb{E}(\tilde{\theta}_n) = \theta + O(n^{-2}).$$

See also leave-one-out cross validation, for $\widehat{m}(\cdot)$

$$\text{mse} = \frac{1}{n}\sum_{i=1}^{n}[y_i - \widehat{m}_{(-i)}(x_i)]^2$$

Boostrap estimate is based on bootstrap samples: set $\widehat{\theta}_{(b)} = \widehat{\theta}(\boldsymbol{x}_{(b)})$, and $\tilde{\theta} = \dfrac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_{(b)}$, where $\boldsymbol{x}_{(b)}$ is a vector of size n, where values are drawn from $\{x_1, \cdots, x_n\}$, with replacement. And then use the law of large numbers...

See Efron (1979).

Hold-Out, Cross Validation, Bootstrap

Hold-out: Split $\{1, \cdots, n\}$ into T (training) and V (validation)

Train the model on $\{(y_i, \boldsymbol{x}_i), i \in T\}$ and compute

$$\widehat{R} = \frac{1}{\#(V)} \sum_{i \in V} \ell(y_i, \widehat{m}(\boldsymbol{x}_i)$$

k-fold cross validation: Split $\{1, \cdots, n\}$ into $I_1, \cdots, I_k$. Set $I_{\overline{j}} = \{1, \cdots, n\} \backslash I_j$

Train model on $I_{\overline{j}}$ and compute

$$\widehat{R} = \frac{1}{k} \sum_j R_j \text{ where } R_j = \frac{k}{n} \sum_{i \in I_j} \ell(y_i, \widehat{m}_{\overline{j}}(\boldsymbol{x}_i))$$

Hold-Out, Cross Validation, Bootstrap

Leave-one-out bootstrap: generate $I_1, \cdots, I_B$ bootstrapped samples from $\{1, \cdots, n\}$

set $n_i = \mathbf{1}_{i \notin I_1} + \cdots + \mathbf{1}_{i \notin I_B}$

$$\widehat{R} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{b: i \notin I_b} \ell(y_i, \widehat{m}_b(\boldsymbol{x}_i)$$

Remark Probability that ith raw is not selection $(1 - n^{-1})^n \to e^{-1} \sim 36.8\%$,

cf training / validation samples (2/3-1/3)

Statistical Learning and Philosophical Issues

From $(y_i, \boldsymbol{x}_i)$, there are different stories behind, see Freedman (2005)

- the causal story : $x_{j,i}$ is usually considered as independent of the other covariates $x_{k,i}$. For all possible $\boldsymbol{x}$, that value is mapped to $m(\boldsymbol{x})$ and a noise is atatched, ε. The goal is to recover $m(\cdot)$, and the residuals are just the difference between the response value and $m(\boldsymbol{x})$.
- the conditional distribution story : for a linear model, we usually say that Y given $\boldsymbol{X} = \boldsymbol{x}$ is a $\mathcal{N}(m(\boldsymbol{x}), \sigma^2)$ distribution. $m(\boldsymbol{x})$ is then the conditional mean. Here $m(\cdot)$ is assumed to really exist, but no causal assumption is made, only a conditional one.
- the explanatory data story : there is no model, just data. We simply want to summarize information contained in $\boldsymbol{x}$'s to get an accurate summary, close to the response (i.e. $\min\{\ell(\boldsymbol{y}_i, m(\boldsymbol{x}_i))\}$) for some loss function ℓ.

Machine Learning vs. Statistical Modeling

In **machine learning**, given some dataset $(\boldsymbol{x}_i, y_i)$, solve

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

for some loss functions $\ell(\cdot,\cdot)$.

In **statistical modeling**, given some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, assume that y_i are realization of i.i.d. variables Y_i (given $\boldsymbol{X}_i = \boldsymbol{x}_i$) with distribution F_i. Then solve

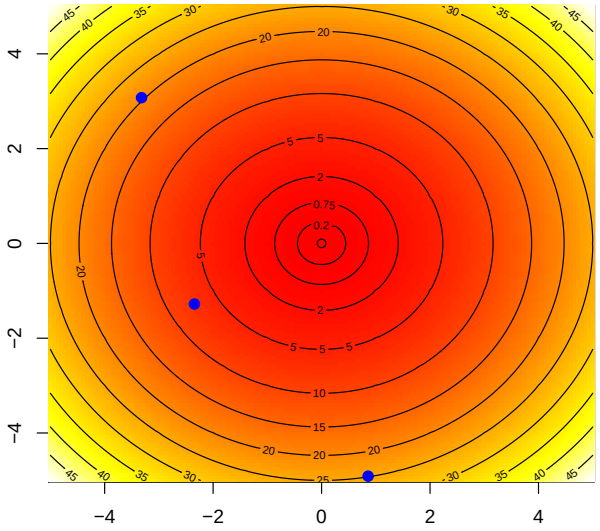
$$\widehat{m}(\cdot) = \underset{m(\cdot)\in\mathcal{F}}{\operatorname{argmax}} \left\{ \log \mathcal{L}(m(\boldsymbol{x}); \boldsymbol{y}) \right\} = \underset{m(\cdot)\in\mathcal{F}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \log f(y_i; m(\boldsymbol{x}_i)) \right\}$$

where $\log \mathcal{L}$ denotes the log-likelihood.

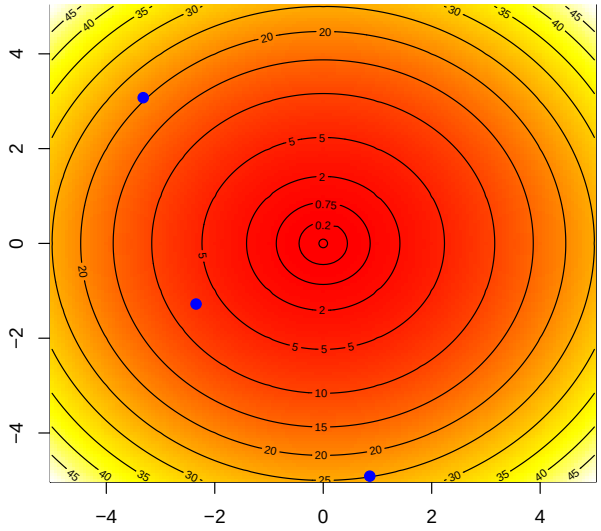
Computational Aspects: Optimization

Econometrics, Statistics and Machine Learning rely on the same object: optimization routines.

A gradient descent/ascent algorithm



A stochastic algorithm



Loss Functions

Fitting criteria are based on loss functions (also called cost functions). For a quantitative response, a popular one is the quadratic loss, $\ell(y, m(\boldsymbol{x})) = [y - m(\boldsymbol{x})]^2$.

Recall that

$$\begin{cases} \mathbb{E}(Y) = \underset{m\in\mathbb{R}}{\text{argmin}}\{\|Y-m\|_{\ell_2}^2\} = \underset{m\in\mathbb{R}}{\text{argmin}}\{\mathbb{E}\left([Y-m]^2\right)\} \\ \text{Var}(Y) = \underset{m\in\mathbb{R}}{\min}\{\mathbb{E}\left([Y-m]^2\right)\} = \mathbb{E}\left([Y-\mathbb{E}(Y)]^2\right) \end{cases}$$

The empirical version is

$$\begin{cases} \overline{y} = \underset{m\in\mathbb{R}}{\text{argmin}}\{\sum_{i=1}^{n}\frac{1}{n}[y_i-m]^2\} \\ s^2 = \underset{m\in\mathbb{R}}{\min}\{\sum_{i=1}^{n}\frac{1}{n}[y_i-m]^2\} = \sum_{i=1}^{n}\frac{1}{n}[y_i-\overline{y}]^2 \end{cases}$$

Loss Functions

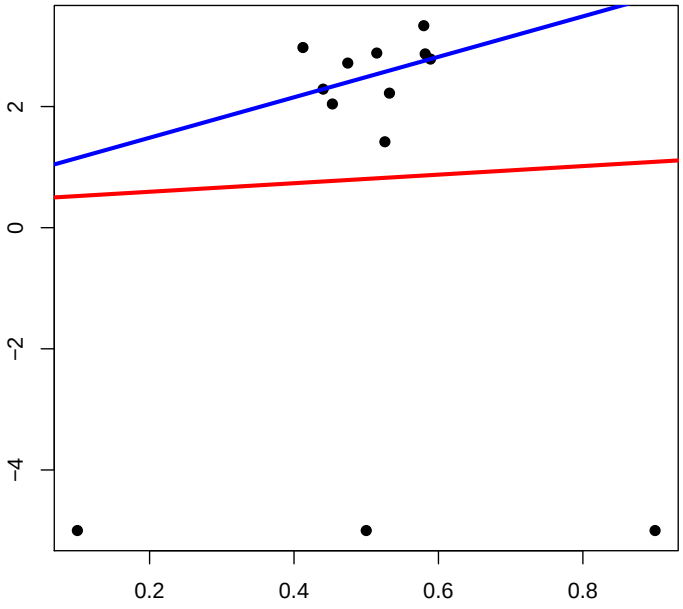
Remark $\text{median}(\boldsymbol{y}) = \underset{m\in\mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \frac{1}{n} |y_i - m| \right\}$

Quadratic loss function $\ell(a,b)^2 = (a-b)^2$,

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta})^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2$$

Absolute loss function $\ell(a,b) = |a-b|$

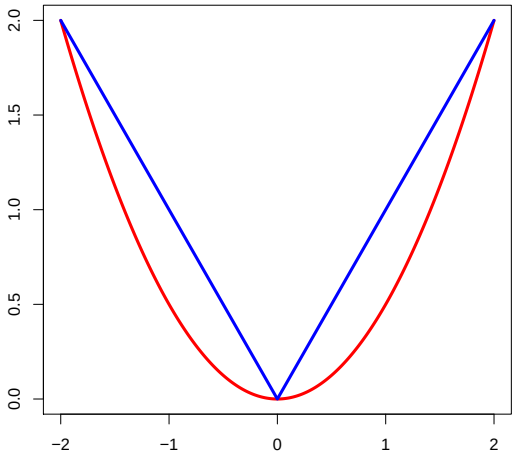
$$\sum_{i=1}^{n} |y_i - \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}| = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\ell_1}$$



Loss Functions

Quadratic loss function $\ell_2(x,y)^2 = (x-y)^2$,
Absolute loss function $\ell_1(x,y) = |x-y|$
Quantile loss function $\ell_\tau(x,y) = |(x-y)(\tau - \mathbf{1}_{x\leq y})|$
Huber loss function

$$\ell_\tau(x,y) = \begin{cases} \frac{1}{2}(x-x)^2 & \text{for } |x-y| \leq \tau, \\ \tau\,|x-y| - \frac{1}{2}\tau^2 & \text{otherwise.} \end{cases}$$



i.e. quadratic when $|x-y| \leq \tau$ and linear otherwise.

Loss Functions

For classification: misclassification loss function

$$\ell(x,y)=\mathbf{1}_{x\neq y} \text{ or } \ell(x,y)=\mathbf{1}_{\text{sign}(x)\neq\text{sign}(y)}$$

$$\ell_\tau(x,y)=\tau\mathbf{1}_{\text{sign}(x)<0,\text{sign}(y)>0}+[1-\tau]\mathbf{1}_{\text{sign}(x)>0,\text{sign}(y)<0}$$

For $\{-1,+1\}$ classes,

Hinge loss ('maximum-margin' classification) $\ell(x,y)=(1-xy)_+$

Logistic/log loss $\ell(x,y)=\log[1+e^{-xy}]$

Squared Loss $\ell(x,y)=[x-y]^2=[1-xy]^2$

Linear Predictors

In the linear model, least square estimator yields

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\boldsymbol{X}^\mathsf{T}}_{\boldsymbol{H}}\boldsymbol{Y}$$

We have a linear predictor if the fitted value $\widehat{y}$ at point $\boldsymbol{x}$ can be written

$$\widehat{y} = \widehat{m}(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{x},i} y_i = \boldsymbol{S}_{\boldsymbol{x}}^\mathsf{T}\boldsymbol{y}$$

where $\boldsymbol{S}_{\boldsymbol{x}}$ is some vector of weights (called smoother vector), related to a $n \times n$ smoother matrix,

$$\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$$

where prediction is done at points $\boldsymbol{x}_i$'s.

Degrees of Freedom and Model Complexity

E.g.

$$\boldsymbol{S_x} = \boldsymbol{X}[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\boldsymbol{x}$$

that is related to the hat matrix, $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$.

Note that

$$T = \frac{\|\boldsymbol{S}\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}\|}{\text{trace}([\boldsymbol{S} - \boldsymbol{H}]^\mathsf{T}[\boldsymbol{S} - \boldsymbol{H}])}$$

can be used to test a linear assumtion: if the model is linear, then T has a Fisher distribution.

In the context of linear predictors, trace($\boldsymbol{S}$) is usually called equivalent number of parameters and is related to $n-$ effective degrees of freedom (as in Ruppert *et al.* (2003)).

Model Evaluation

In linear models, the R^2 is defined as the proportion of the variance of the the response y that can be obtained using the predictors.

But maximizing the R^2 usually yields overfit (or unjustified optimism in Berk (2008)).

In linear models, consider the adjusted R^2,

$$\overline{R}^2 = 1 - [1 - R^2]\frac{n-1}{n-p-1}$$

where p is the number of parameters (or more generally trace($\boldsymbol{S}$)).

Model Evaluation

Alternatives are based on the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), based on a penalty imposed on some criteria (the logarithm of the variance of the residuals),

$$AIC = \log\left(\frac{1}{n}\sum_{i=1}^{n}[y_i - \widehat{y}_i]^2\right) + \frac{2p}{n}$$

$$BIC = \log\left(\frac{1}{n}\sum_{i=1}^{n}[y_i - \widehat{y}_i]^2\right) + \frac{\log(n)p}{n}$$

In a more general context, replace p by $\text{trace}(\boldsymbol{S})$

Goodhart's Law

'when a measure becomes a target, it ceases to be a good measure', by Charles Goodhart

Occam's Razor

Between two models which explain as well the data, choose the simplest one

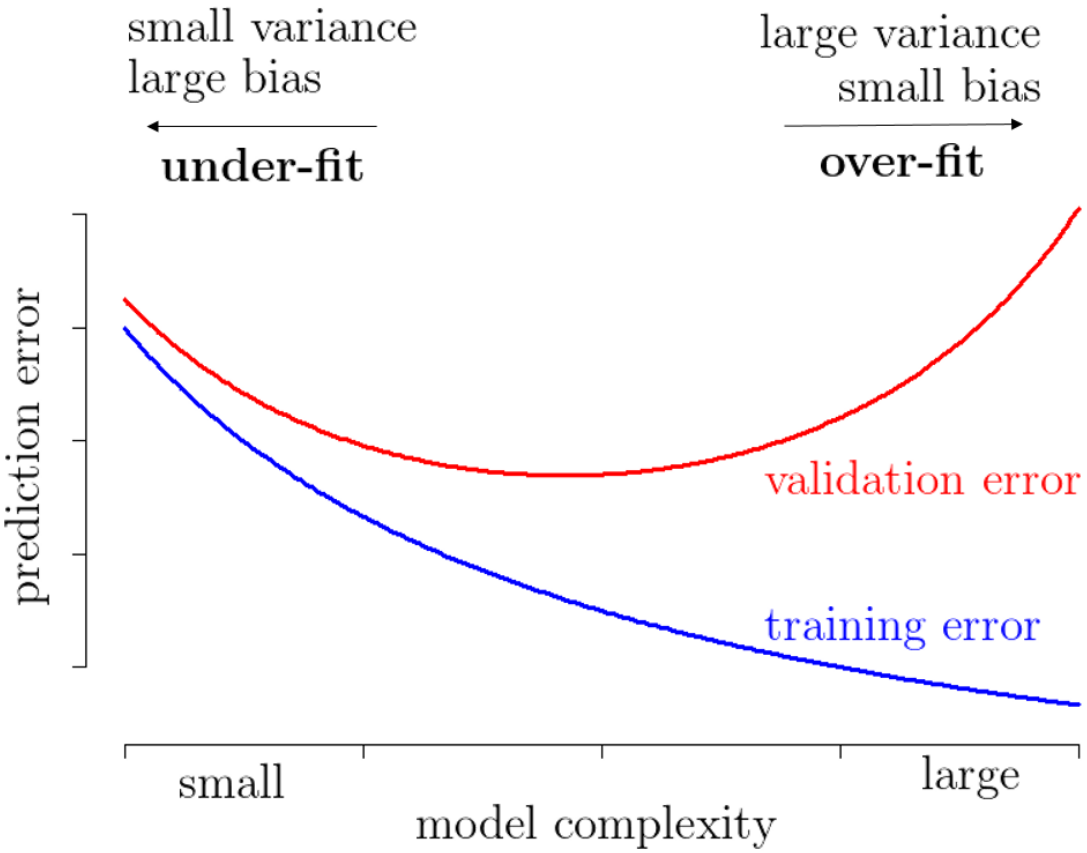
Machine Learning: usually need to tradeoff between the training error and model complexity

$$\widehat{m} = \underset{m}{\operatorname{argmin}} \{\ell(Y, m(\boldsymbol{X})) + \Omega(m)\}$$

Occam's Razor

$$\widehat{m} = \underset{m}{\operatorname{argmin}} \{\ell(Y, m(\boldsymbol{X}) + \Omega(m)\}$$

where $\Omega(m)$ is a regularizer, that characterizes the complexity of the model.



Model Evaluation

One can also consider the expected prediction error (with a probabilistic model)

$$\mathbb{E}[\ell(Y, \widehat{m}(\boldsymbol{X})]$$

We cannot claim (using the law of large number) that

$$\frac{1}{n}\sum_{i=1}^{n}\ell(y_i, \widehat{m}(\boldsymbol{x}_i)) \overset{a.s.}{\nrightarrow} \mathbb{E}[\ell(Y, m(\boldsymbol{X})]$$

since $\widehat{m}$ depends on $(y_i, \boldsymbol{x}_i)$'s.

Natural option : use two (random) samples, a training one and a validation one.

Alternative options, use cross-validation, leave-one-out or k-fold.

Underfit / Overfit and Variance - Mean Tradeoff

Goal in predictive modeling: reduce uncertainty in our predictions.

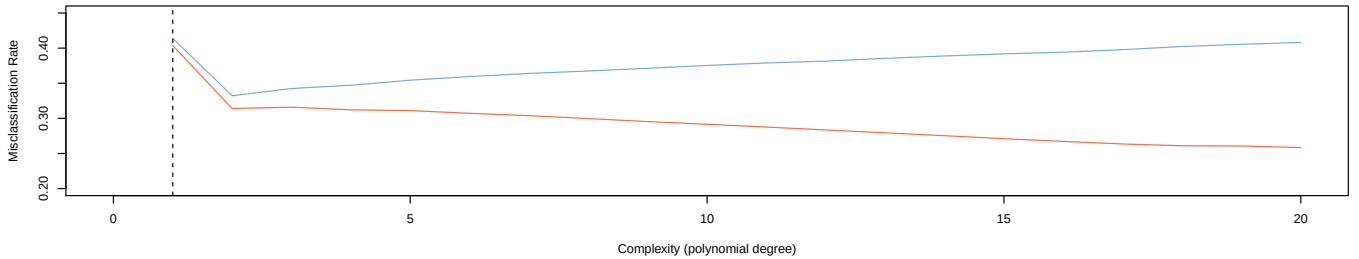
Need more data to get a better knowledge.

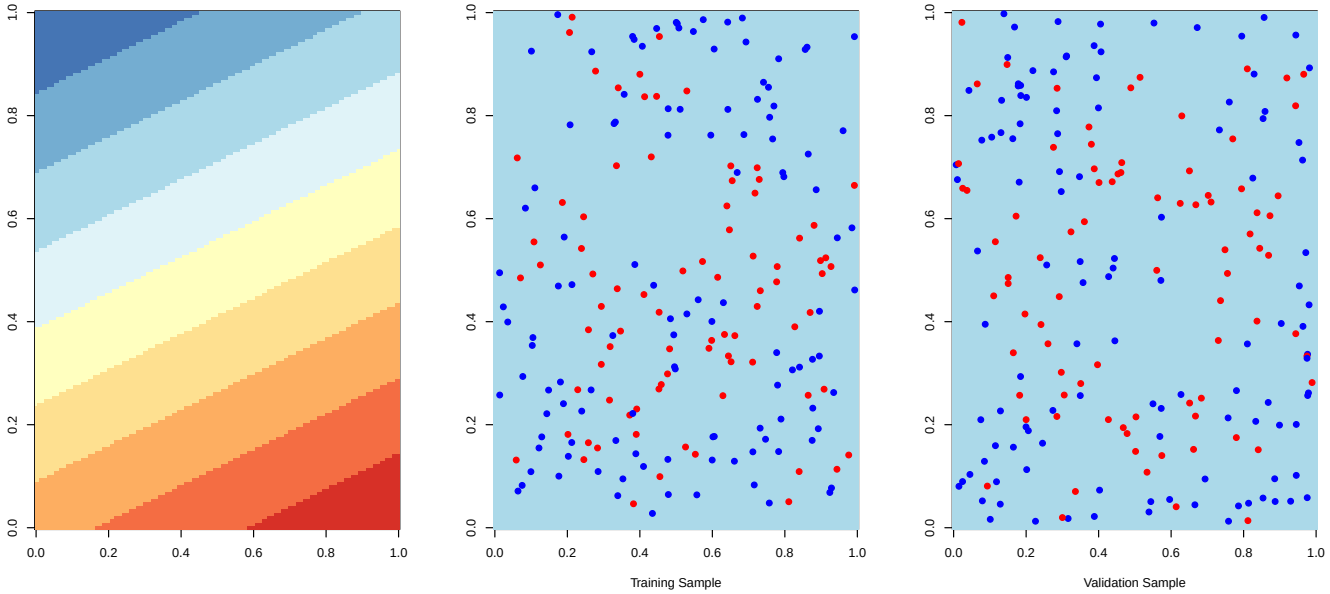
Unfortunately, reducing the error of the prediction on a dataset does not generally give a good generalization performance

$\longrightarrow$ need a training and a validation dataset

Overfit, Training vs. Validation and Complexity

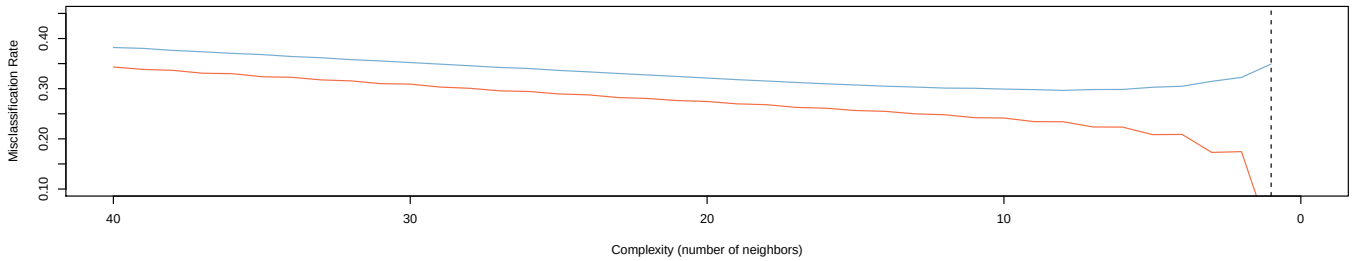
complexity $\longleftrightarrow$ polynomial degree

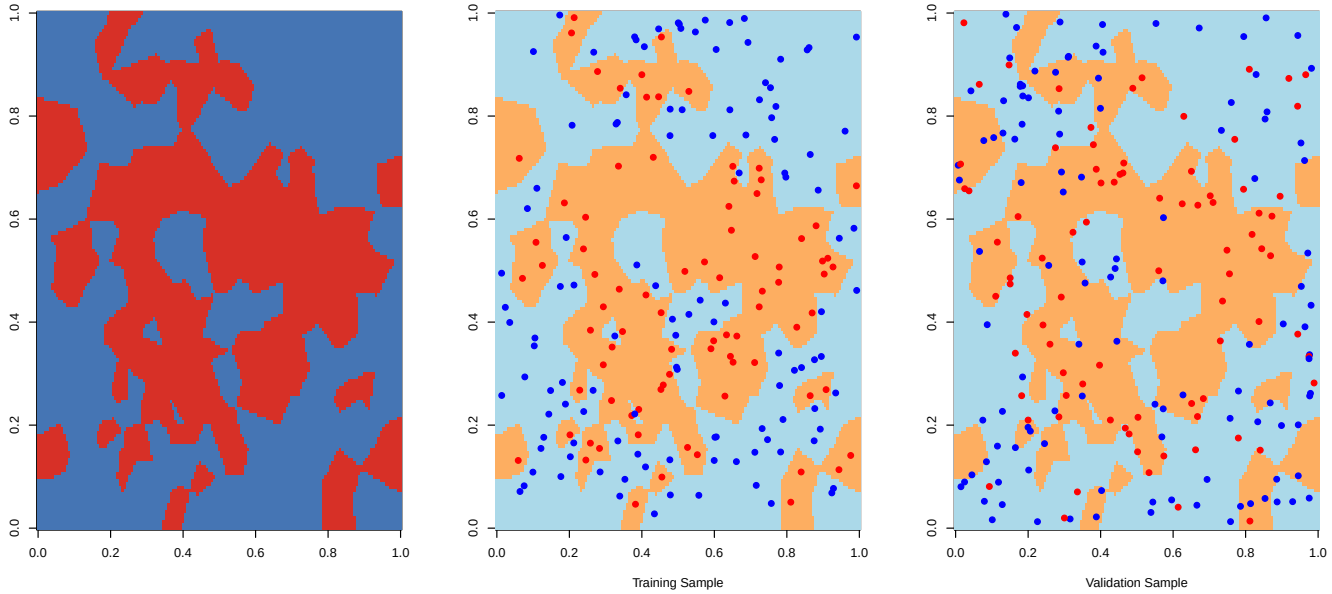




Overfit, Training vs. Validation and Complexity

complexity $\longleftrightarrow$ number of neighbors (k)





Themes in Data Science

Predictive Capability we want here to have a model that predict well for new observations

Bias-Variance Tradeoff A very smooth prediction has less variance, but a large bias. We need to find a good balance between the bias and the variance

Loss Functions In machine learning, goodness of fit is discussed based on disparities between predicted values, and observed one, based on some loss function

Tuning or Meta Parameters Choice will be made in terms of tuning parameters

Interpretability Does it matter to have a good model if we cannot interpret it ?

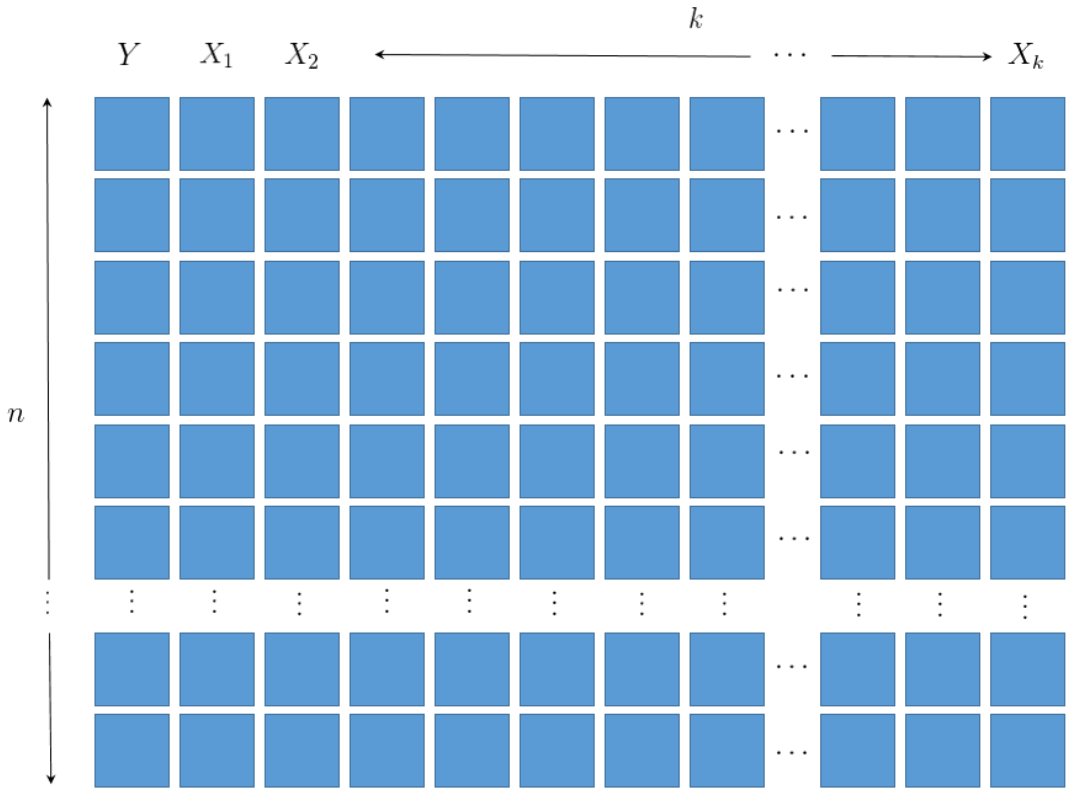
Coding Issues Most of the time, there are no analytical expression, just an alogrithm that should converge to some (possibly) optimal value

Data Data collection is a crucial issue (but will not be discussed here)

Scalability Issues

Dealing with big (or massive) datasets, large number of observations (n) and/or large number of predictors (features or covariates, k).

Ability to parallelize algorithms might be important (map-reduce).



n can be large, but limited (portfolio size)
large variety k
large volume nk

$\rightarrow$ Feature Engineering

Part 3.
Application to Classification

Econometric Based Models in Actuarial Science

Consider an i.i.d. sample $\{y_1, \cdots, y_n\}$ with $y_i \in \{0, 1\}$,

$$\mathbb{P}(Y_i = y_i) = \pi^{y_i}[1-\pi]^{1-y_i}, \text{ with } y_i \in \{0, 1\}.$$

where $\pi \in [0, 1]$, so that $\mathbb{P}(Y_i = 1) = \pi$ and $\mathbb{P}(Y_i = 0) = 1 - \pi$.

The likelihood is

$$\mathcal{L}(\pi; \boldsymbol{y}) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \pi^{y_i}[1-\pi]^{1-y_i}$$

and the log-likelihood is

$$\log \mathcal{L}(\pi; \boldsymbol{y}) = \sum_{i=1}^{n} y_i \log[\pi] + (1 - y_i) \log[1 - \pi]$$

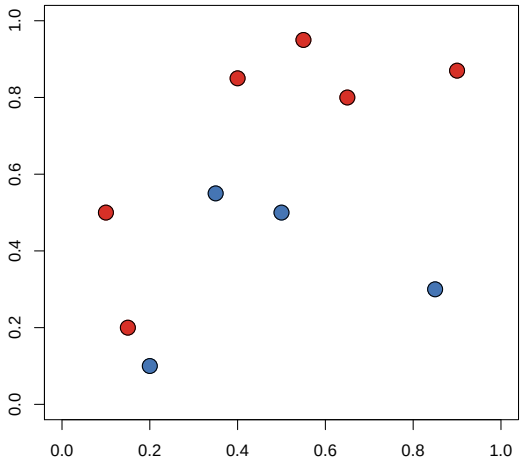
The first order condition is

$$\frac{\partial \log \mathcal{L}(\pi; \boldsymbol{y})}{\partial \pi} = \sum_{i=1}^{n} \frac{y_i}{\pi} - \frac{1 - y_i}{1 - \pi} = 0, \text{ i.e. } \pi^\star = \overline{y}.$$

Econometric Based Models in Actuarial Science

Assume that $\mathbb{P}(Y_i = 1) = \pi_i$,

$$\text{logit}(\pi_i) = \boldsymbol{X}_i'\boldsymbol{\beta}, \text{ where } \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right),$$



or

$$\pi_i = \text{logit}^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta}) = \frac{\exp[\boldsymbol{X}_i'\boldsymbol{\beta}]}{1+\exp[\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}]}.$$

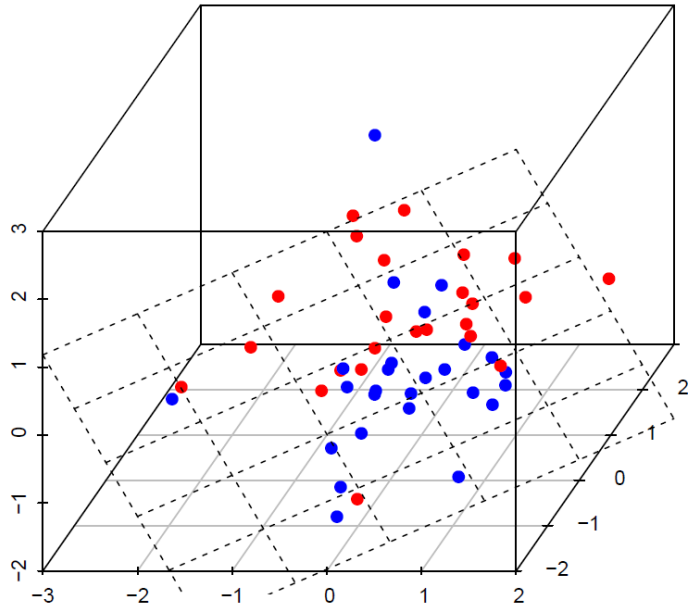
The log-likelihood is

$$\log\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i) = \sum_{i=1}^{n} y_i \log(\pi_i(\boldsymbol{\beta})) + (1-y_i)\log(1-\pi_i(\boldsymbol{\beta}))$$

and the first order conditions are solved numerically

$$\frac{\partial \log\mathcal{L}(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{n} X_{k,i}[y_i - \pi_i(\boldsymbol{\beta})] = 0.$$

Logistic Classification



It is a linear classifier since we have a linear separation between the •'s and the •'s.

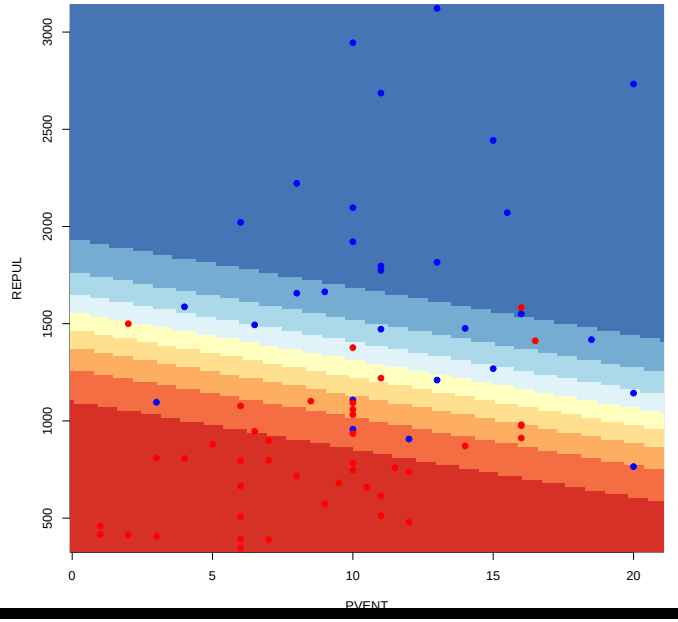
Let $m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$.

With a logistic regression, we can get a prediction

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

Is that the 'best' model we can get from the data?

What if n and/or k are very large?



Why a Logistic and not a Probit Regression?

Bliss (1934) suggested a model such that

$$\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})=H(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) \text{ where } H(\cdot)=\Phi(\cdot)$$

the c.d.f. of the $\mathcal{N}(0,1)$ distribution. This is the probit model. This yields a latent model, $y_i=\mathbf{1}(y_i^\star>0)$ where

$$y_i^\star=\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}+\varepsilon_i \text{ is a nonobservable score.}$$

In the logistic regression, we model the odds ratio,

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

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

which is the c.d.f. of the logistic variable, see Verhulst (1845)

Table 3.2 Transformation of percentages to probits

%	0	1	2	3	4	5	6	7	8	9
0	—	2.67	2.95	3.12	3.25	3.36	3.45	3.52	3.59	3.66
10	3.72	3.77	3.82	3.87	3.92	3.96	4.01	4.05	4.08	4.12
20	4.16	4.19	4.23	4.26	4.29	4.33	4.36	4.39	4.42	4.45
30	4.48	4.50	4.53	4.56	4.59	4.61	4.64	4.67	4.69	4.72
40	4.75	4.77	4.80	4.82	4.85	4.87	4.90	4.92	4.95	4.97
50	5.00	5.03	5.05	5.08	5.10	5.13	5.15	5.18	5.20	5.23
60	5.25	5.28	5.31	5.33	5.36	5.39	5.41	5.44	5.47	5.50
70	5.52	5.55	5.58	5.61	5.64	5.67	5.71	5.74	5.77	5.81
80	5.84	5.88	5.92	5.95	5.99	6.04	6.08	6.13	6.18	6.23
90	6.28	6.34	6.41	6.48	6.55	6.64	6.75	6.88	7.05	7.33
—	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
99	7.33	7.37	7.41	7.46	7.51	7.58	7.65	7.75	7.88	8.09

Soit p la population : représentons par dp l'accroissement infiniment petit qu'elle reçoit pendant un temps infiniment court dt. Si la population croissait en progression géométrique, nous aurions l'équation $\frac{dp}{dt}=mp$. Mais comme la vitesse d'accroissement de la population est retardée par l'augmentation même du nombre des habitans, nous devrons retrancher de mp une fonction inconnue de p; de manière que la formule à intégrer deviendra

$$\frac{dp}{dt}=mp-\varphi(p).$$

L'hypothèse la plus simple que l'on puisse faire sur la forme de la fonction φ, est de supposer $\varphi(p)=np^2$. On trouve alors pour intégrale de l'équation ci-dessus

$$t=\frac{1}{m}[\log. p-\log.(m-np)]+\text{constante},$$

et il suffira de trois observations pour déterminer les deux coefficiens constans m et n et la constante arbitraire.

En résolvant la dernière équation par rapport à p, il vient

$$p=\frac{mp'e^{mt}}{np'e^{mt}+m-np'} \quad \ldots\ldots (1)$$

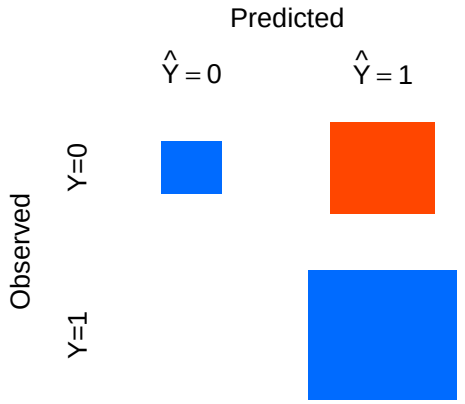
en désignant par p' la population qui répond à $t=o$, et par e la base des logarithmes népériens. Si l'on fait $t=\infty$, on voit que la valeur de p correspondante est $P=\frac{m}{n}$. Telle est donc *la limite supérieure de la population*.

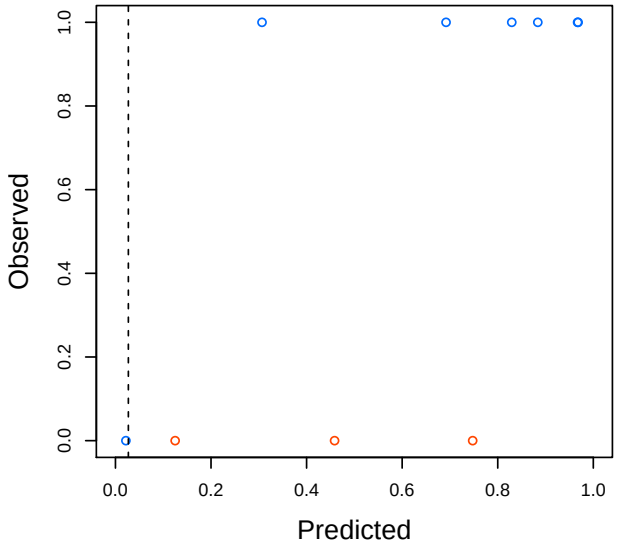
Predictive Classifier

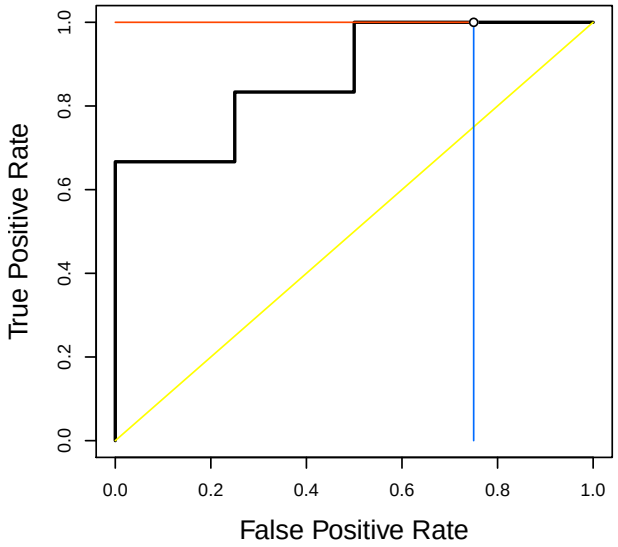
To go from a score to a class:

if $s(\boldsymbol{x}) > s$, then $\widehat{Y}(\boldsymbol{x}) = 1$ and $s(\boldsymbol{x}) \leq s$, then $\widehat{Y}(\boldsymbol{x}) = 0$

Plot $TP(s) = \mathbb{P}[\widehat{Y} = 1 | Y = 1]$ against $FP(s) = \mathbb{P}[\widehat{Y} = 1 | Y = 0]$







Comparing Classifiers: Accuracy and Kappa

Kappa statistic κ compares an Observed Accuracy with an Expected Accuracy (random chance), see Landis & Koch (1977).

	$Y=0$	$Y=1$	
$\widehat{Y}=0$	TN	FN	TN+FN
$\widehat{Y}=1$	FP	TP	FP+TP
	TN+FP	FN+TP	n

See also Observed and Random Confusion Tables

	$Y=0$	$Y=1$	
$\widehat{Y}=0$	25	3	28
$\widehat{Y}=1$	4	39	43
	29	42	71

	$Y=0$	$Y=1$	
$\widehat{Y}=0$	11.44	16.56	28
$\widehat{Y}=1$	17.56	25.44	43
	29	42	71

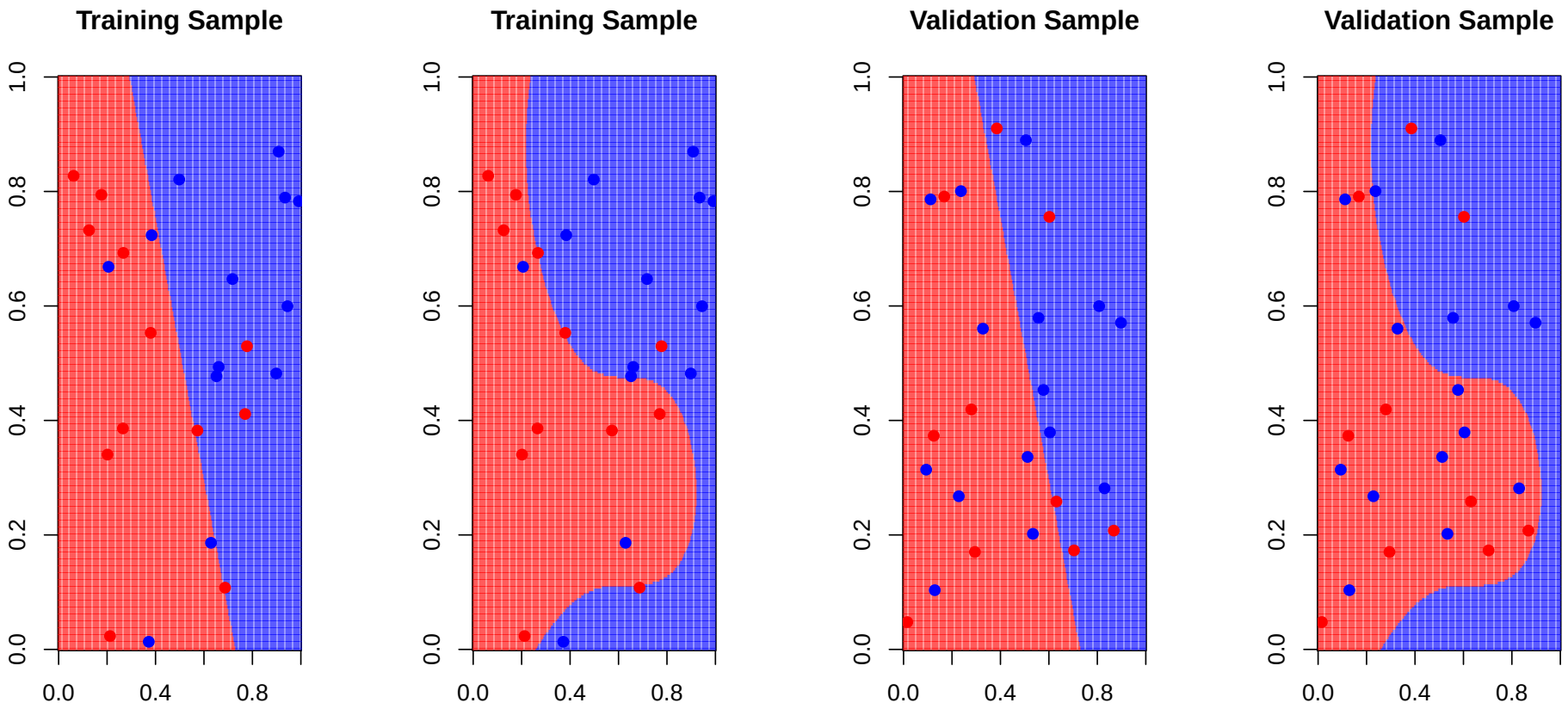
$$\text{total accuracy} = \frac{TP+TN}{n} \sim 90.14\%$$

$$\text{random accuracy} = \frac{[TN+FP]\cdot[TP+FN]+[TP+FP]\cdot[TN+FN]}{n^2} \sim 51.93\%$$

$$\kappa = \frac{\text{total accuracy} - \text{random accuracy}}{1 - \text{random accuracy}} \sim 79.48\%$$

On Model Selection

Consider predictions obtained from a linear model and a nonlinear model, either on the training sample, or on a validation sample,

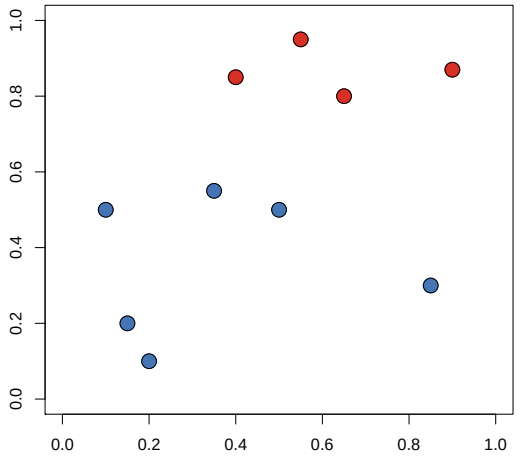


Penalization and Support Vector Machines

SVMs were developed in the 90's based on previous work, from Vapnik & Lerner (1963), see also Vailant (1984).

Assume that points are linearly separable, i.e. there is $\boldsymbol{\omega}$ and b such that

$$Y = \begin{cases} +1 \text{ if } \boldsymbol{\omega}^\mathsf{T}\boldsymbol{x} + b > 0 \\ -1 \text{ if } \boldsymbol{\omega}^\mathsf{T}\boldsymbol{x} + b < 0 \end{cases}$$



Problem: infinite number of solutions, need a good one, that separate the data, (somehow) far from the data.

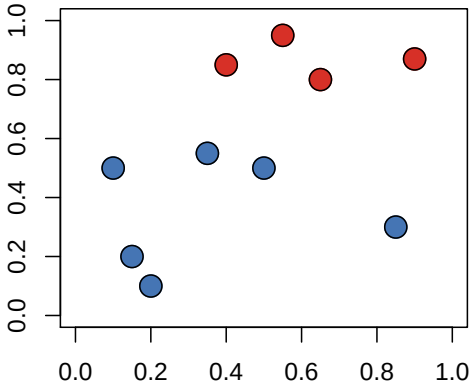
maximize the distance s.t. $H_{\boldsymbol{\omega},b}$ separates ± 1 points, i.e.

$$\min\left\{\frac{1}{2}\boldsymbol{\omega}^\mathsf{T}\boldsymbol{\omega}\right\} \text{ s.t. } Y_i(\boldsymbol{\omega}^\mathsf{T}\boldsymbol{x}_i + b) \geq 1, \ \forall i.$$

Penalization and Support Vector Machines

Define support vectors as observations such that

$$|\boldsymbol{\omega}^{\mathsf{T}}\boldsymbol{x}_i + b| = 1$$



The margin is the distance between hyperplanes defined by support vectors. The distance from support vectors to $H_{\boldsymbol{\omega},b}$ is $\|\boldsymbol{\omega}\|^{-1}$

Now, what about the non-separable case?

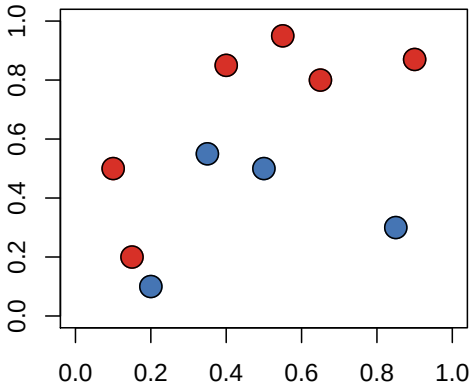
Here, we cannot have $y_i(\boldsymbol{\omega}^{\mathsf{T}}\boldsymbol{x}_i + b) \geq 1 \; \forall i$.

Penalization and Support Vector Machines

Thus, introduce slack variables,

$$\begin{cases} \boldsymbol{\omega}^\top \boldsymbol{x}_i + b \geq +1 - \xi_i \text{ when } y_i = +1 \\ \boldsymbol{\omega}^\top \boldsymbol{x}_i + b \leq -1 + \xi_i \text{ when } y_i = -1 \end{cases}$$

where $\xi_i \geq 0 \ \forall i$. There is a classification error when $\xi_i > 1$.



The idea is then to solve

$$\min\left\{\frac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\omega} + C\mathbf{1}^\top\mathbf{1}_{\boldsymbol{\xi}>1}\right\}, \text{ instead of} \min\left\{\frac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\omega}\right\}$$

Support Vector Machines, with a Linear Kernel

So far,

$$d(\boldsymbol{x}_0, H_{\boldsymbol{\omega},b}) = \min_{\boldsymbol{x} \in H_{\boldsymbol{\omega},b}} \{\|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\ell_2}\}$$

where $\|\cdot\|_{\ell_2}$ is the Euclidean (ℓ_2) norm,

$$\|\boldsymbol{x}_0 - \boldsymbol{x}\|_{\ell_2} = \sqrt{(\boldsymbol{x}_0 - \boldsymbol{x}) \cdot (\boldsymbol{x}_0 - \boldsymbol{x})} = \sqrt{\boldsymbol{x}_0 \cdot \boldsymbol{x}_0 - 2\boldsymbol{x}_0 \cdot \boldsymbol{x} + \boldsymbol{x} \cdot \boldsymbol{x}}$$

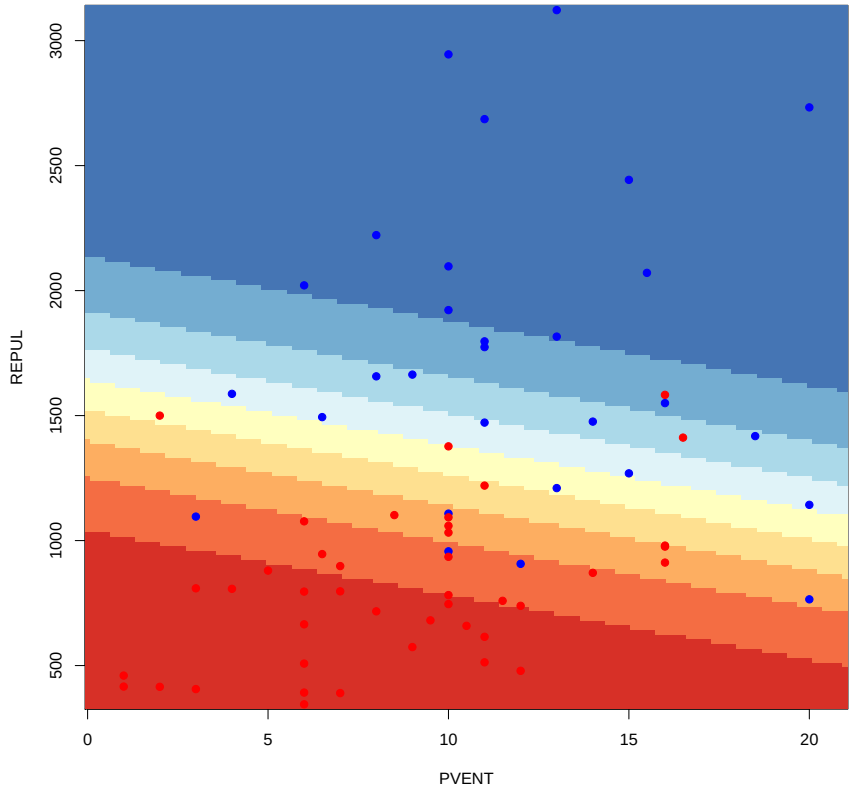
More generally,

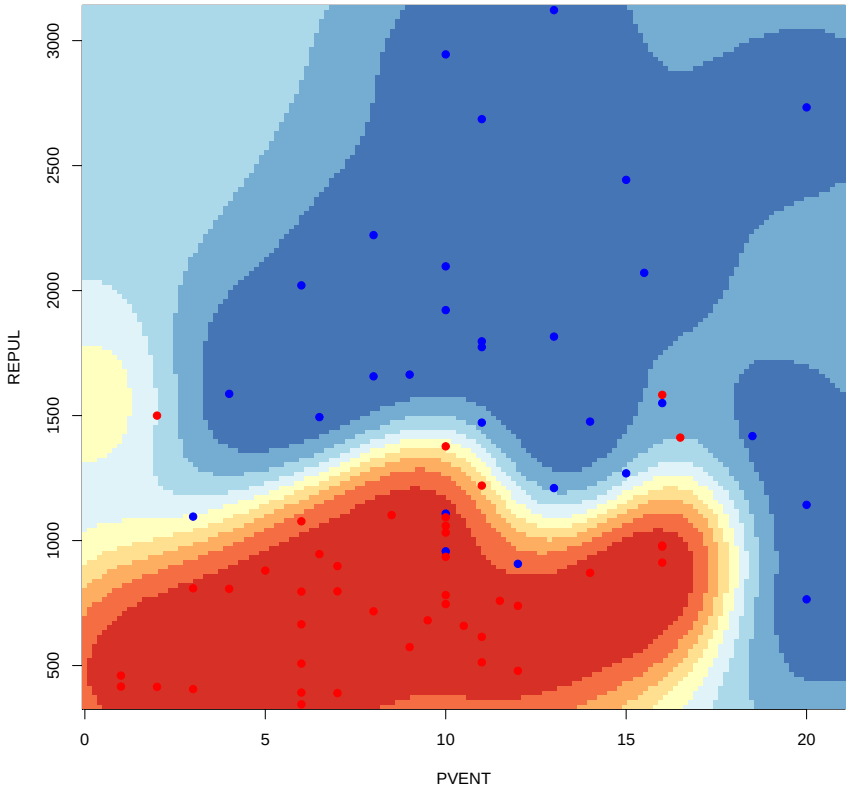
$$d(\boldsymbol{x}_0, H_{\boldsymbol{\omega},b}) = \min_{\boldsymbol{x} \in H_{\boldsymbol{\omega},b}} \{\|\boldsymbol{x}_0 - \boldsymbol{x}\|_{k}\}$$

where $\|\cdot\|_{k}$ is some kernel-based norm,

$$\|\boldsymbol{x}_0 - \boldsymbol{x}\|_{k} = \sqrt{k(\boldsymbol{x}_0, \boldsymbol{x}_0) - 2k(\boldsymbol{x}_0, \boldsymbol{x}) + k(\boldsymbol{x} \cdot \boldsymbol{x})}$$

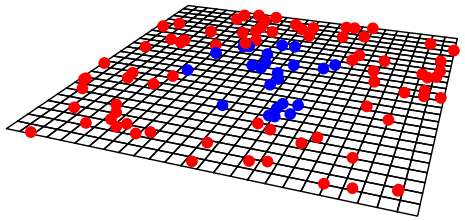
Support Vector Machines, with a Non Linear Kernel

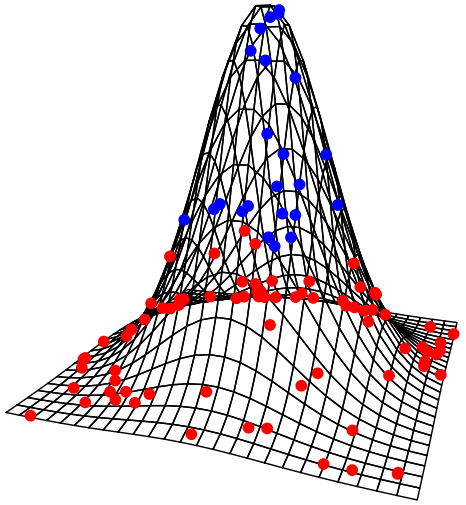




Heuristics on SVMs

An interpretation is that data aren't linearly seperable in the original space, but might be separare by some kernel transformation,





Penalization and Mean Square Error

Consider the quadratic loss function, $\ell(\theta,\hat{\theta})=(\theta-\hat{\theta})^2$, the risk function becomes the mean squared error of the estimate,

$$R(\theta,\hat{\theta})=\mathbb{E}(\theta-\hat{\theta})^2=\underbrace{[\theta-\mathbb{E}(\hat{\theta})]^2}_{\text{bias}^2}+\underbrace{\mathbb{E}(\mathbb{E}[\hat{\theta}]-\hat{\theta})^2}_{\text{variance}}$$

Get back to the intial example, $y_i\in\{0,1\}$, with $p=\mathbb{P}(Y=1)$.

Consider the estimate that minimizes the mse, that can be writen $\widehat{p}=(1-\alpha)\overline{y}$, then

$$\text{mse}(\widehat{p})=\alpha^2p^2+(1-\alpha)^2\frac{p(1-p)}{n}$$

then $\alpha^\star=\dfrac{1-p}{1+(n-1)p}$.

i.e.unbiased estimators have nice mathematical properties, but can be improved.

Linear Model

Consider some linear model $y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i$ for all $i = 1, \cdots, n$.

Assume that ε_i are i.i.d. with $\mathbb{E}(\varepsilon) = 0$ (and finite variance). Write

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\boldsymbol{y}, n\times 1} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix}}_{\boldsymbol{X}, n\times(k+1)} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}, (k+1)\times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}, n\times 1}.$$

Assuming $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathbb{I})$, the maximum likelihood estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\{\|\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|_{\ell_2}\} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$$

... under the assumtption that $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ is a full-rank matrix.

What if $\boldsymbol{X}_i^\mathsf{T}\boldsymbol{X}$ cannot be inverted? Then $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^\mathsf{T}\boldsymbol{X}]^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$ does not exist, but $\widehat{\boldsymbol{\beta}}_\lambda = [\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\mathbb{I}]^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$ always exist if $\lambda > 0$.

Ridge Regression

The estimator $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda\mathbb{I}]^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$ is the Ridge estimate obtained as solution of

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} [y_i - \beta_0 - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}]^2 + \lambda \underbrace{\|\boldsymbol{\beta}\|_{\ell_2}}_{\mathbf{1}^\mathsf{T}\boldsymbol{\beta}^2} \right\}$$

for some tuning parameter λ. One can also write

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_2}\leq s}{\text{argmin}} \{\|\boldsymbol{Y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|_{\ell_2}\}$$

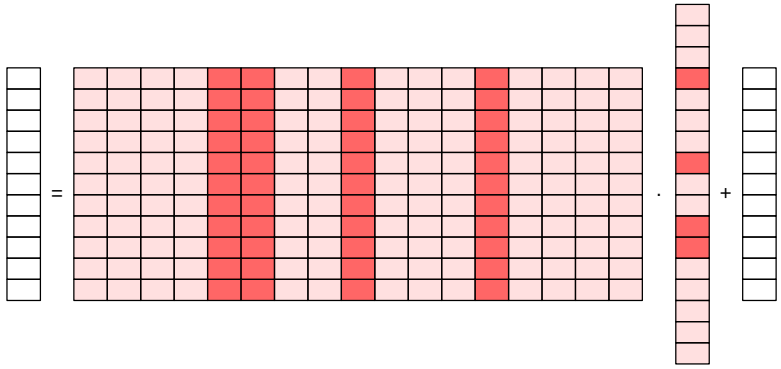
Remark Note that we solve $\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\{\text{objective}(\boldsymbol{\beta})\}$ where

$$\text{objective}(\boldsymbol{\beta}) = \underbrace{\mathcal{L}(\boldsymbol{\beta})}_{\text{training loss}} + \underbrace{\mathcal{R}(\boldsymbol{\beta})}_{\text{regularization}}$$

Going further on sparcity issues

In severall applications, k can be (very) large, but a lot of features are just noise: $\beta_j = 0$ for many j's. Let s denote the number of relevent features, with $s << k$, cf Hastie, Tibshirani & Wainwright (2015),

$$s = \text{card}\{\mathcal{S}\} \text{ where } \mathcal{S} = \{j; \beta_j \neq 0\}$$



The model is now $y = \boldsymbol{X}_{\mathcal{S}}^{\mathsf{T}}\boldsymbol{\beta}_{\mathcal{S}} + \varepsilon$, where $\boldsymbol{X}_{\mathcal{S}}^{\mathsf{T}}\boldsymbol{X}_{\mathcal{S}}$ is a full rank matrix.

Going further on sparcity issues

Define $\|\boldsymbol{a}\|_{\ell_0} = \sum \mathbf{1}(|a_i| > 0)$. Ici $\dim(\boldsymbol{\beta}) = s$.

We wish we could solve

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_0} \leq s}{\operatorname{argmin}} \{\|\boldsymbol{Y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|_{\ell_2}\}$$

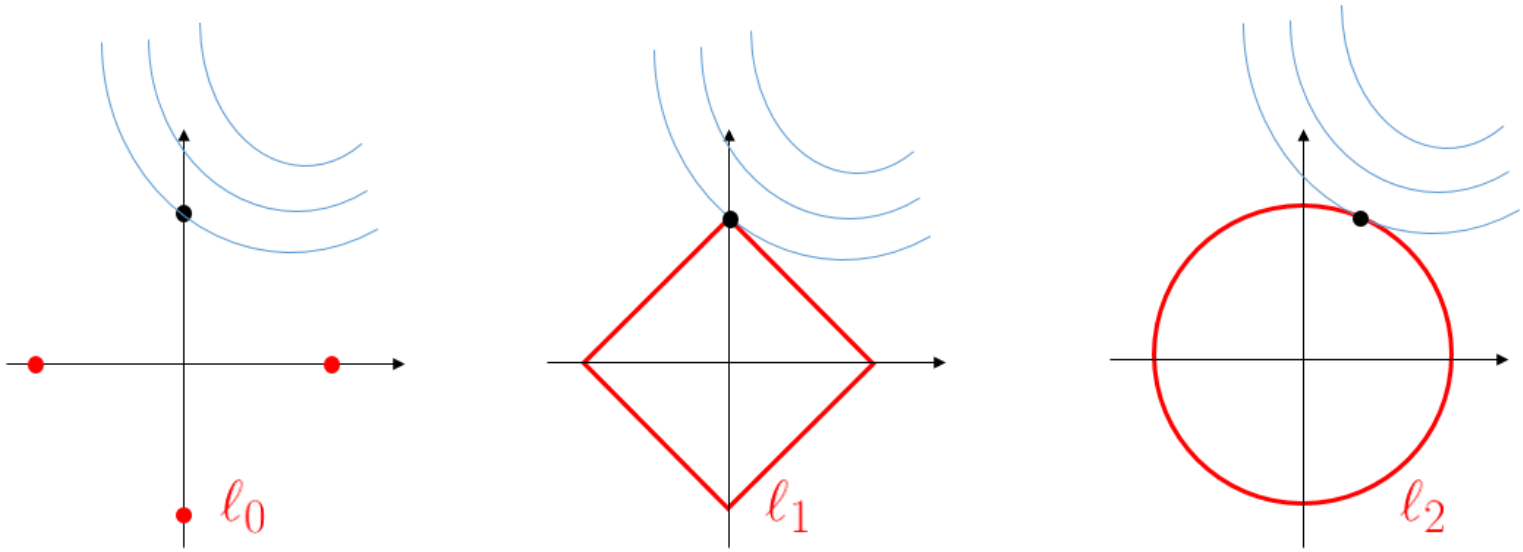
Problem: it is usually not possible to describe all possible constraints, since $\binom{s}{k}$ coefficients should be chosen here (with k (very) large).

Idea: solve the dual problem

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta};\|\boldsymbol{Y}-\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|_{\ell_2} \leq h}{\operatorname{argmin}} \{\|\boldsymbol{\beta}\|_{\ell_0}\}$$

where we might convexify the ℓ_0 norm, $\|\cdot\|_{\ell_0}$.

Regularization ℓ_0, ℓ_1 and ℓ_2



$$\min\{\|\boldsymbol{\beta}\|_{\ell_\star}\} \text{ subject to } \|\boldsymbol{Y} - \boldsymbol{X}^\top\boldsymbol{\beta}\|_{\ell_2} \leq h$$

Going further on sparcity issues

On $[-1,+1]^k$, the convex hull of $\|\boldsymbol{\beta}\|_{\ell_0}$ is $\|\boldsymbol{\beta}\|_{\ell_1}$

On $[-a,+a]^k$, the convex hull of $\|\boldsymbol{\beta}\|_{\ell_0}$ is $a^{-1}\|\boldsymbol{\beta}\|_{\ell_1}$

Hence,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_1}\leq\tilde{s}}{\operatorname{argmin}} \{\|\boldsymbol{Y}-\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|_{\ell_2}\}$$

is equivalent (Kuhn-Tucker theorem) to the Lagragian optimization problem

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}\{\|\boldsymbol{Y}-\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|_{\ell_2}+\lambda\|\boldsymbol{\beta}\|_{\ell_1}\}$$

LASSO *Least Absolute Shrinkage and Selection Operator*

$$\widehat{\boldsymbol{\beta}} \in \text{argmin}\{\|\boldsymbol{Y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|_{\ell_2} + \lambda\|\boldsymbol{\beta}\|_{\ell_1}\}$$

is a convex problem (several algorithms*), but not strictly convex (no unicity of the minimum). Nevertheless, predictions $\widehat{\boldsymbol{y}} = \boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\beta}}$ are unique

* MM, minimize majorization, coordinate descent Hunter (2003).

Optimal LASSO Penalty

Use cross validation, e.g. K-fold,

$$\widehat{\boldsymbol{\beta}}_{(-k)}(\lambda) = \text{argmin}\left\{\sum_{i\notin\mathcal{I}_k}[y_i - \boldsymbol{x}_i^\intercal\boldsymbol{\beta}]^2 + \lambda\|\boldsymbol{\beta}\|\right\}$$

then compute the sum of the squared errors,

$$Q_k(\lambda) = \sum_{i\in\mathcal{I}_k}[y_i - \boldsymbol{x}_i^\intercal\widehat{\boldsymbol{\beta}}_{(-k)}(\lambda)]^2$$

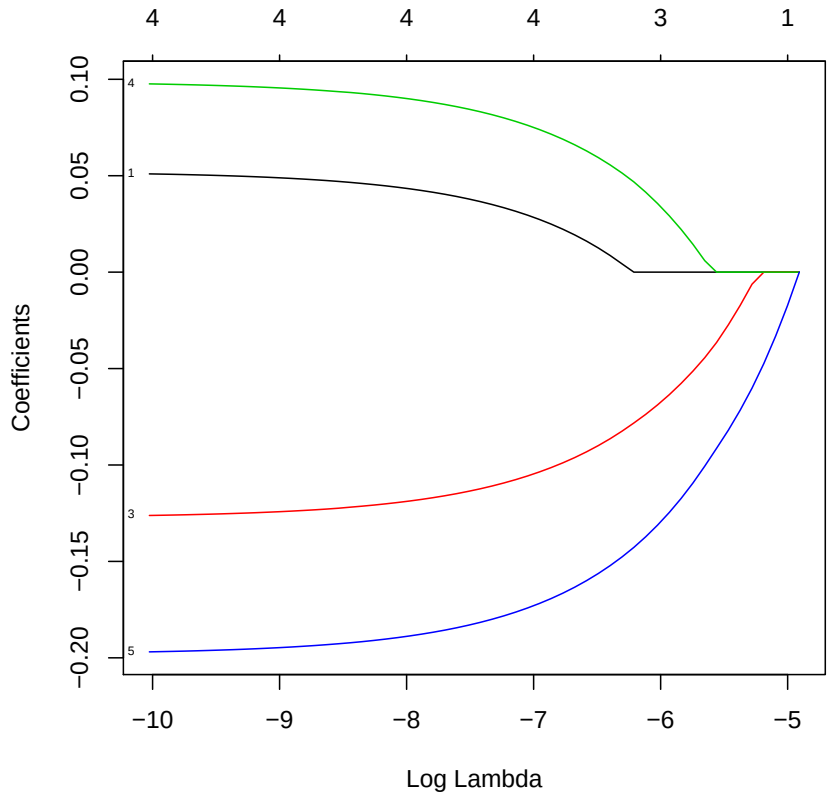
and finally solve

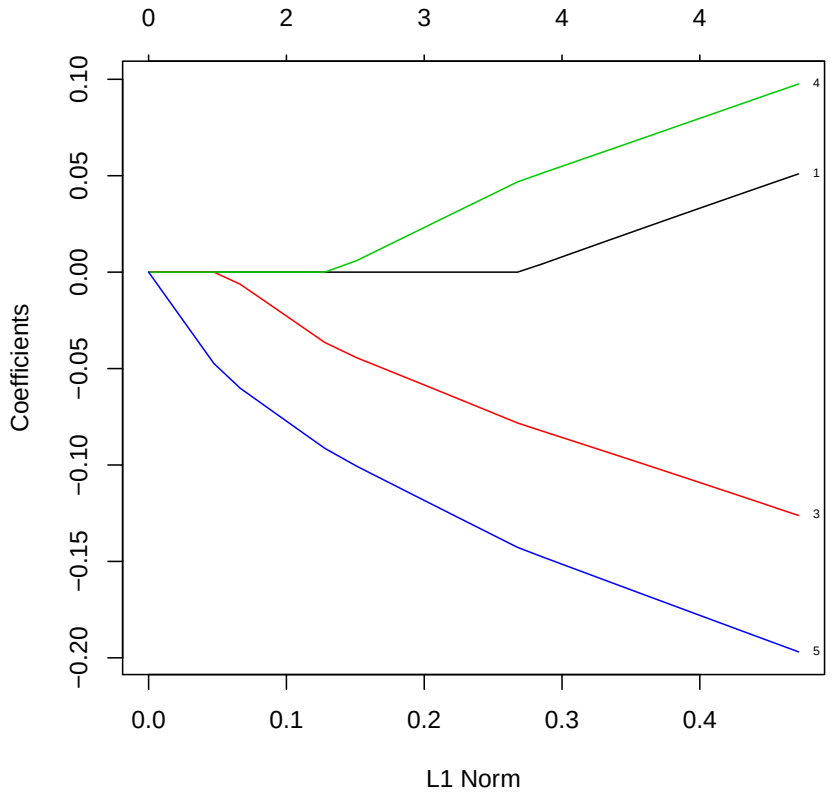
$$\lambda^\star = \text{argmin}\left\{\overline{Q}(\lambda) = \frac{1}{K}\sum_k Q_k(\lambda)\right\}$$

Note that this might overfit, so Hastie, Tibshiriani & Friedman (2009) suggest the largest λ such that

$$\overline{Q}(\lambda) \leq \overline{Q}(\lambda^\star) + \text{se}[\lambda^\star] \text{ with } \text{se}[\lambda]^2 = \frac{1}{K^2}\sum_{k=1}^K[Q_k(\lambda) - \overline{Q}(\lambda)]^2$$

LASSO


4 4 4 4 3 1
Coefficients
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
4
1
3
5
−10 −9 −8 −7 −6 −5
Log Lambda


0 2 3 4 4
Coefficients
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
4
1
3
5
0.0 0.1 0.2 0.3 0.4
L1 Norm

Penalization and GLM's

The logistic regression is based on empirical risk, when $y \in \{0, 1\}$

$$-\frac{1}{n}\sum_{i=1}^{n}\left(y_i \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} - \log[1+\exp(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})]\right)$$

or, if $y \in \{-1, +1\}$,

$$\frac{1}{n}\sum_{i=1}^{n}\log\left[1+\exp(y_i \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})\right].$$

A regularized version with the ℓ_1 norm is the LASSO logistic regression

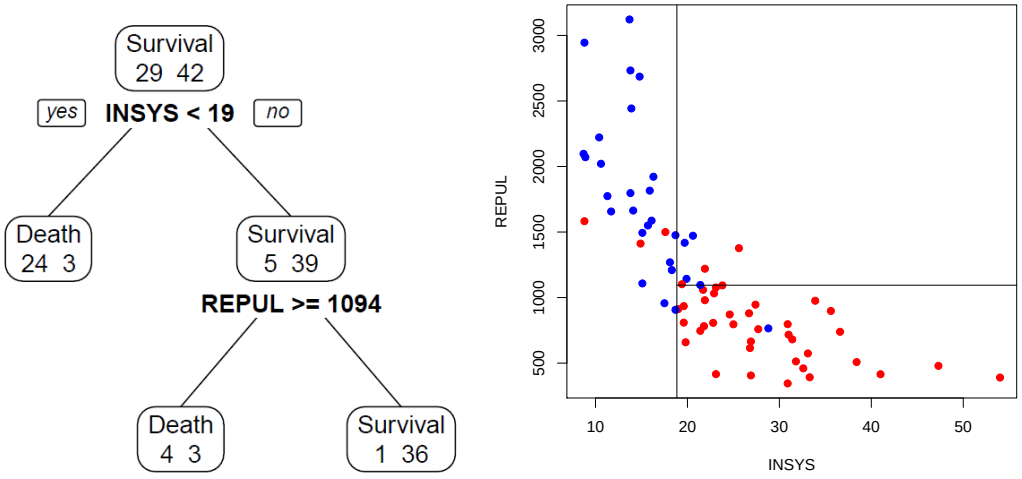
$$\frac{1}{n}\sum_{i=1}^{n}\log\left[1+\exp(y_i \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta})\right] + \lambda\|\boldsymbol{\beta}\|_1$$

or more generaly, with smoothing functions

$$\frac{1}{n}\sum_{i=1}^{n}\log\left[1+\exp(y_i g(\boldsymbol{x}_i))\right] + \lambda\|g\|$$

Classification (and Regression) Trees, CART

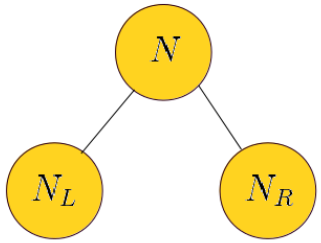
one of the predictive modelling approaches used in statistics, data mining and machine learning [...] In tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. (Source: wikipedia).



Classification (and Regression) Trees, CART

To split N into two $\{N_L, N_R\}$, consider

$$\mathcal{I}(N_L, N_R) = \sum_{x \in \{L,R\}} \frac{n_x}{n} \mathcal{I}(N_x)$$



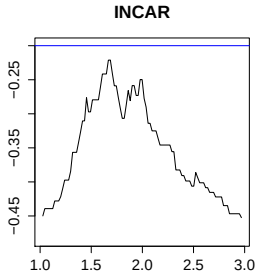
e.g. Gini index (used originally in CART, see Breiman *et al.* (1984))

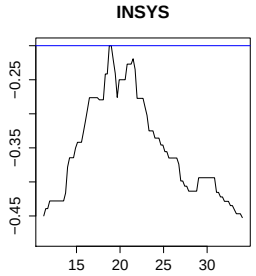
$$\text{gini}(N_L, N_R) = - \sum_{x \in \{L,R\}} \frac{n_x}{n} \sum_{y \in \{0,1\}} \frac{n_{x,y}}{n_x} \left(1 - \frac{n_{x,y}}{n_x}\right)$$

and the cross-entropy (used in C4.5 and C5.0)

$$\text{entropy}(N_L, N_R) = - \sum_{x \in \{L,R\}} \frac{n_x}{n} \sum_{y \in \{0,1\}} \frac{n_{x,y}}{n_x} \log \left(\frac{n_{x,y}}{n_x}\right)$$

Classification (and Regression) Trees, CART

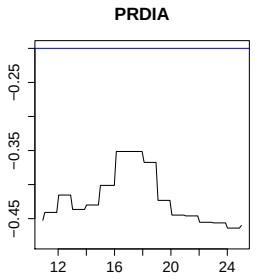


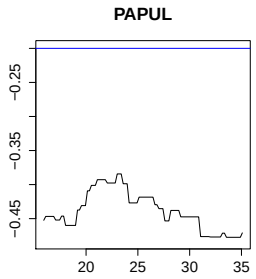


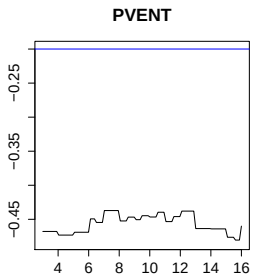
$$N_L\colon \{x_{i,j} \leq s\} \quad N_R\colon \{x_{i,j} > s\}$$

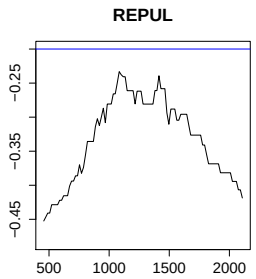
$$\text{solve} \max_{j\in\{1,\cdots,k\},s} \{\mathcal{I}(N_L, N_R)\}$$

$\longleftarrow$ first split

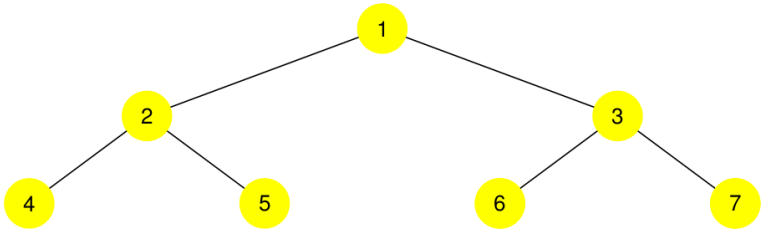


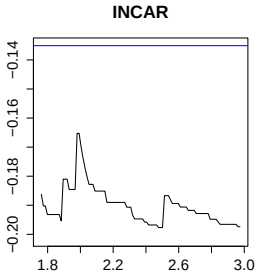


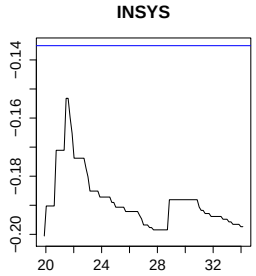


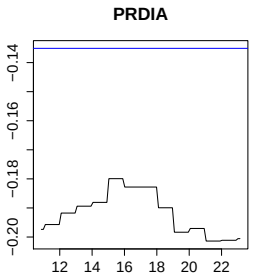


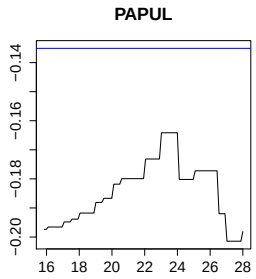
second split $\longrightarrow$

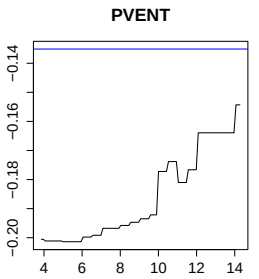


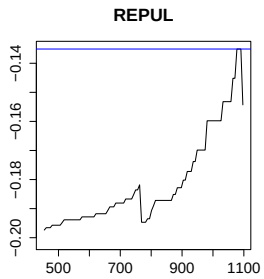












Pruning Trees

One can grow a big tree, until leaves have a (preset) small number of observations, and then possibly go back and prune branches (or leaves) that do not improve gains on good classification sufficiently.

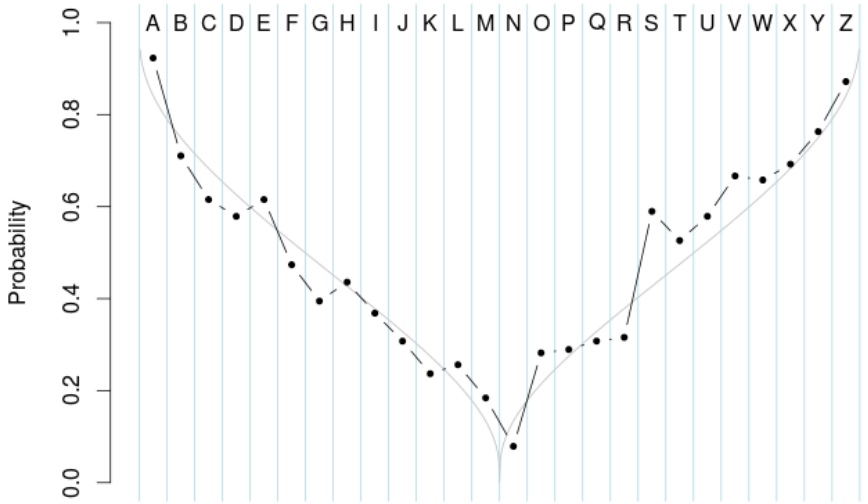
Or we can decide, at each node, whether we split, or not.

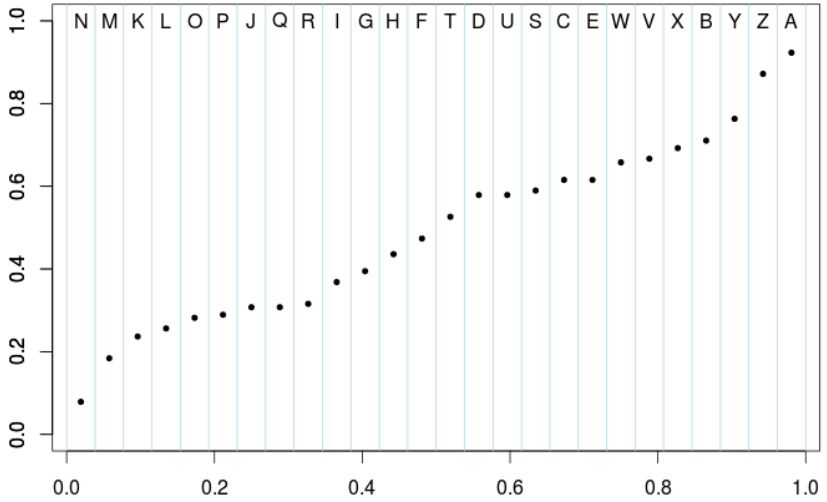
In trees, overfitting increases with the number of steps, and leaves. Drop in impurity at node N is defined as

$$\Delta\mathcal{I}(N_L,N_R)=\mathcal{I}(N)-\mathcal{I}(N_L,N_R)=\mathcal{I}(N)-\left(\frac{n_L}{n}\mathcal{I}(N_L)-\frac{n_R}{n}\mathcal{I}(N_R)\right)$$

(Fast) Trees with Categorical Features

Consider some simple categorical covariate, $x \in \{A, B, C, \cdots, Y, Z\}$, defined from a continuous latent variable $\widetilde{x} \sim \mathcal{U}([0,1])$.





Compute $\overline{y}(x) = \dfrac{1}{n_x} \displaystyle\sum_{i:x_i=x} y_i \approx \mathbb{E}[Y|X=x]$ and sort them

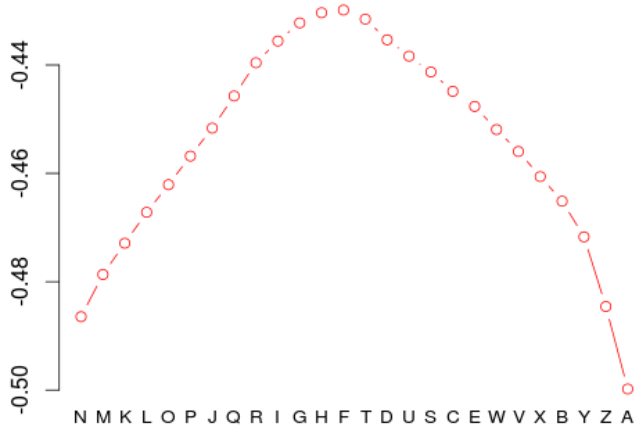
$$\overline{y}(x_{1:26}) \leq \overline{y}(x_{2:26}) \leq \cdots \leq \overline{y}(x_{25:26}) \leq \overline{y}(x_{26:26}).$$

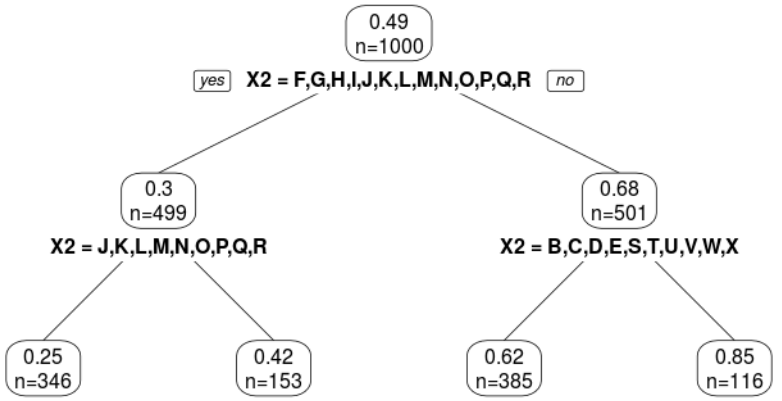
(Fast) Trees with Categorical Features

Then the split is done base on sample

$$x \in \{x_{1:26}, \cdots, x_{j:26}\}$$

$$\text{vs. } x \in \{x_{j+1:26}, \cdots, x_{26:26}\}$$





Bagging

Bootstrapped Aggregation (Bagging) , is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification (Source: wikipedia).

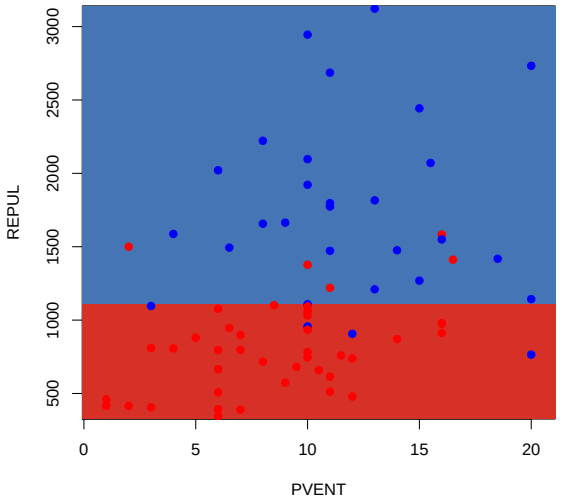
It is an ensemble method that creates multiple models of the same type from different sub-samples of the same dataset [boostrap]. The predictions from each separate model are combined together to provide a superior result [aggregation].

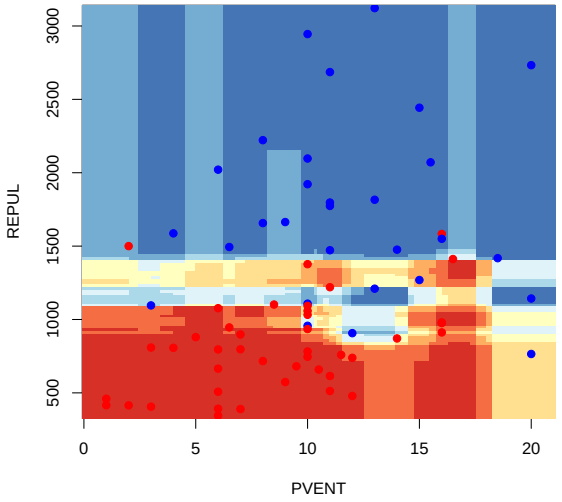
$\rightarrow$ can be used on any kind of model, but interesting for trees, see Breiman (1996)

Boostrap can be used to define the concept of margin,

$$\text{margin}_i = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(\widehat{y}_i = y_i) - \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(\widehat{y}_i \neq y_i)$$

Remark Probability that ith raw is not selection $(1-n^{-1})^n \rightarrow e^{-1} \sim 36.8\%$, cf training / validation samples (2/3-1/3)





Bagging : Bootstrap Aggregation

For classes, $\tilde{m}(\boldsymbol{x}) = \underset{y}{\text{argmax}} \sum_{b=1}^{B} \mathbf{1}(y = \widehat{m}^{(b)})$.

For probabilities,

$$\tilde{m}(\boldsymbol{x}) = \frac{1}{n} \sum_{b=1}^{B} \widehat{m}^{(b)}(\boldsymbol{x}) = \frac{1}{n} \sum_{b=1}^{B} \sum_{j=1}^{k_b} y_i \mathbf{1}(\boldsymbol{x}_i \in C_j).$$

Model Selection and Gini/Lorentz (on incomes)

Consider an ordered sample $\{y_1, \cdots, y_n\}$, then Lorenz curve is

$$\{F_i, L_i\} \text{ with } F_i = \frac{i}{n} \text{ and } L_i = \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{n} y_j}$$

The theoretical curve, given a distribution F, is

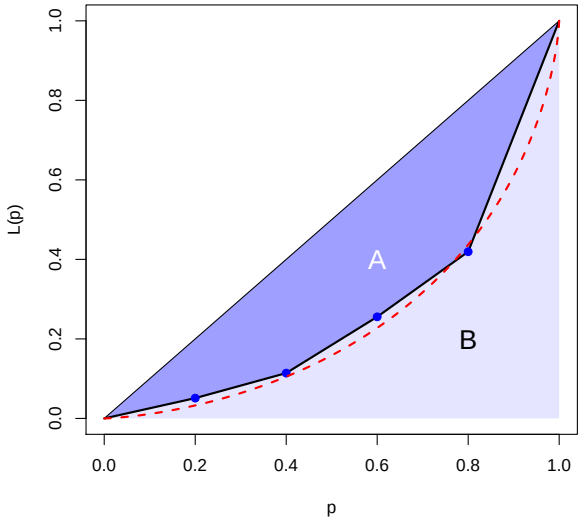
$$u \mapsto L(u) = \frac{\int_{-\infty}^{F^{-1}(u)} t dF(t)}{\int_{-\infty}^{+\infty} t dF(t)}$$

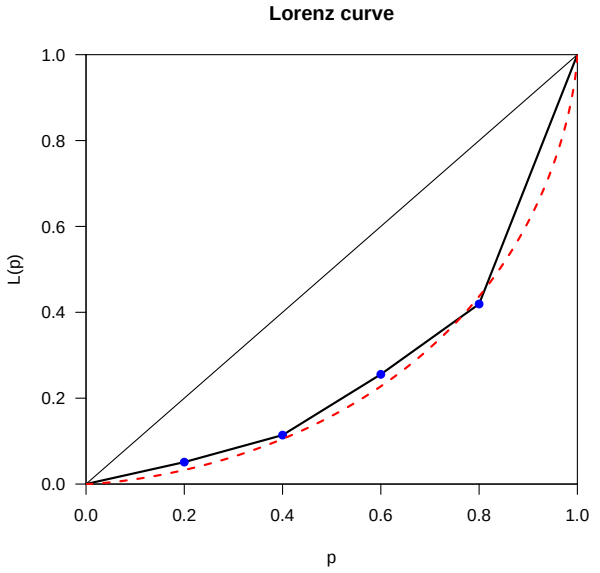
see Gastwirth (1972)

Model Selection and Gini/Lorentz

Gini index is the ratio of the areas $\frac{\mathsf{A}}{\mathsf{A}+\mathsf{B}}$. Thus,

$$\begin{aligned} G &= \frac{2}{n(n-1)\overline{x}} \sum_{i=1}^{n} i \cdot x_{i:n} - \frac{n+1}{n-1} \\ &= \frac{1}{\mathbb{E}(Y)} \int_0^{\infty} F(y)(1-F(y))dy \end{aligned}$$

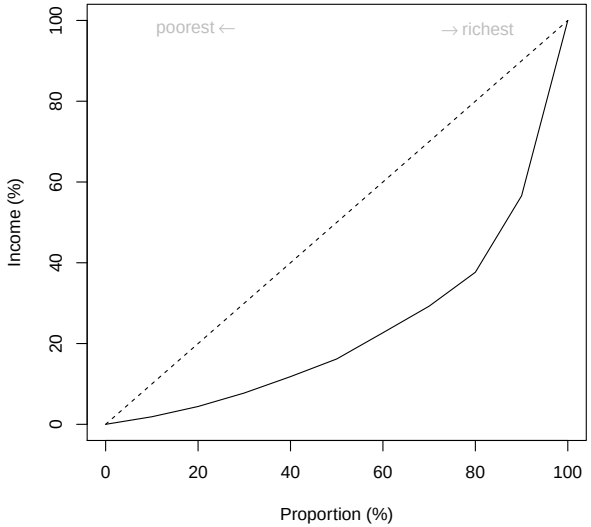




Model Selection

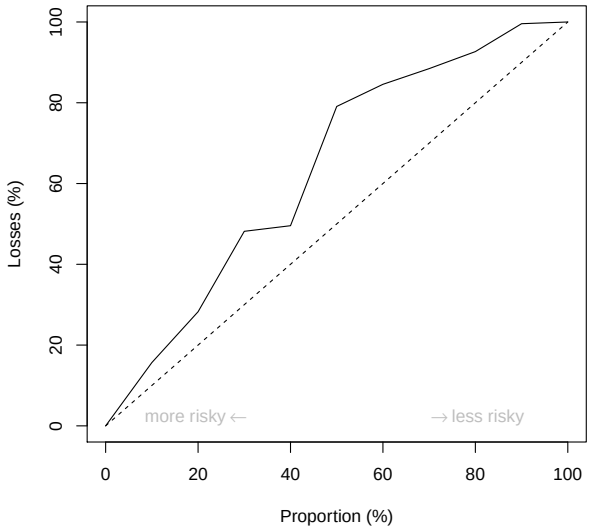
Consider an ordered sample $\{y_1, \cdots, y_n\}$ of incomes, with $y_1 \leq y_2 \leq \cdots \leq y_n$, then Lorenz curve is

$$\{F_i, L_i\} \text{ with } F_i = \frac{i}{n} \text{ and } L_i = \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{n} y_j}$$

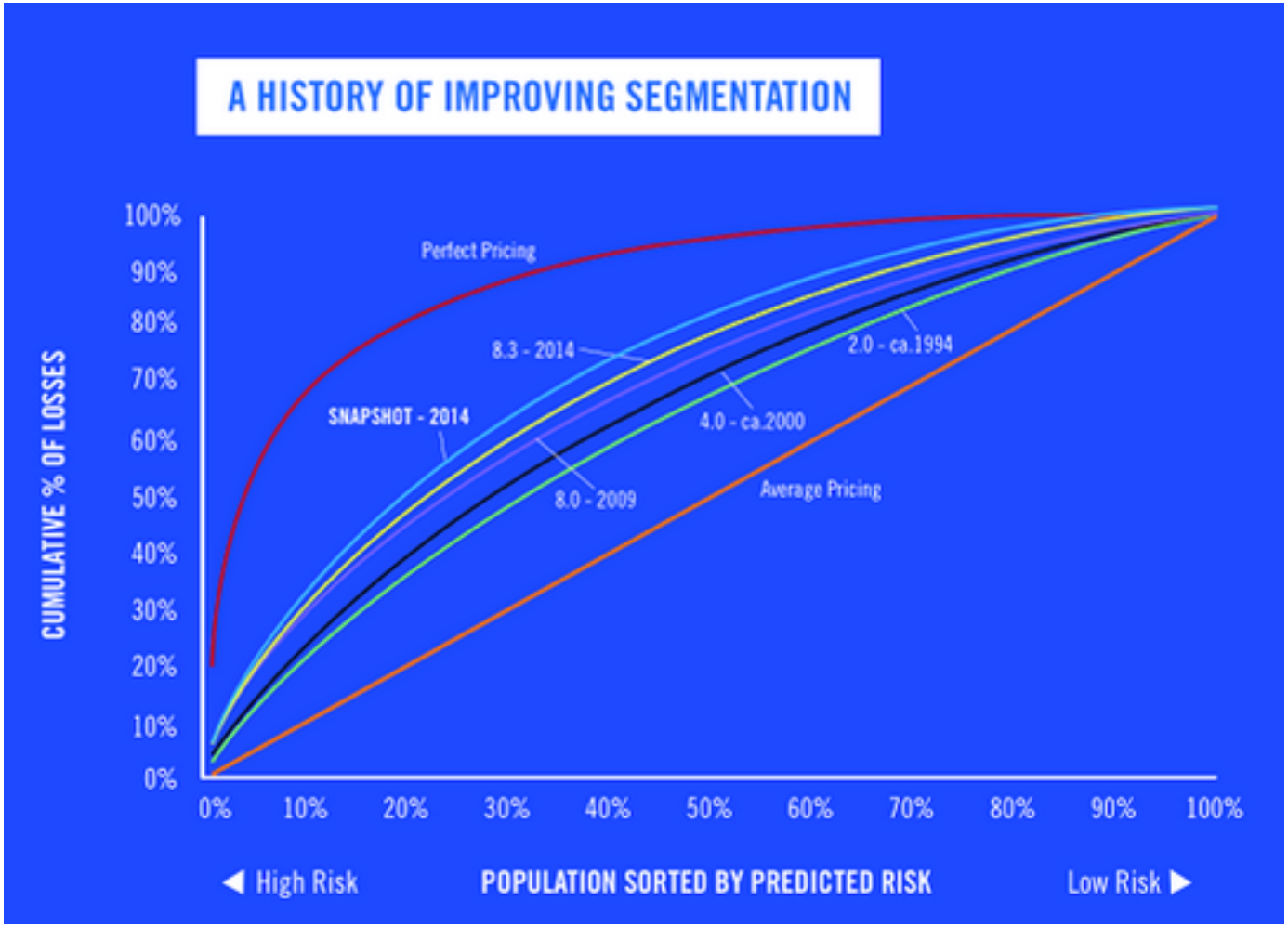


We have observed losses y_i and premiums $\widehat{\pi}(\boldsymbol{x}_i)$. Consider an ordered sample by the model, see Frees, Meyers & Cummins (2014), $\widehat{\pi}(\boldsymbol{x}_1) \geq \widehat{\pi}(\boldsymbol{x}_2) \geq \cdots \geq \widehat{\pi}(\boldsymbol{x}_n)$, then plot

$$\{F_i, L_i\} \text{ with } F_i = \frac{i}{n} \text{ and } L_i = \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{n} y_j}$$



Model Selection



See Frees *et al.* (2010) or Tevet (2013).

Part 4.
Small Data and Bayesian Philosophy

"it's time to adopt modern Bayesian data analysis as standard procedure in our scientific practice and in our educational curriculum. Three reasons:

1. Scientific disciplines from astronomy to zoology are moving to Bayesian analysis.

   We should be leaders of the move, not followers.

2. Modern Bayesian methods provide richer information, with greater flexibility and broader applicability than 20th century methods. Bayesian methods are intellectually coherent and intuitive.

   Bayesian analyses are readily computed with modern software and hardware.

3. Null-hypothesis significance testing (NHST), with its reliance on p values, has many problems.

   There is little reason to persist with NHST now that Bayesian methods are accessible to everyone.

My conclusion from those points is that we should do whatever we can to encourage the move to Bayesian data analysis." John Kruschke,

(quoted in Meyers & Guszcza (2013))

Bayes vs. Frequentist, inference on heads/tails

Consider some Bernoulli sample $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$, where $x_i \in \{0, 1\}$.

X_i's are i.i.d. $\mathcal{B}(p)$ variables, $f_X(x) = p^x[1-p]^{1-x}$, $x \in \{0, 1\}$.

Standard frequentist approach

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \underset{p\in(0,1)}{\operatorname{argmax}}\Big\{ \underbrace{\prod_{i=1}^{n} f_X(x_i)}_{\mathcal{L}(p;\boldsymbol{x})} \Big\}$$

From the central limit theorem

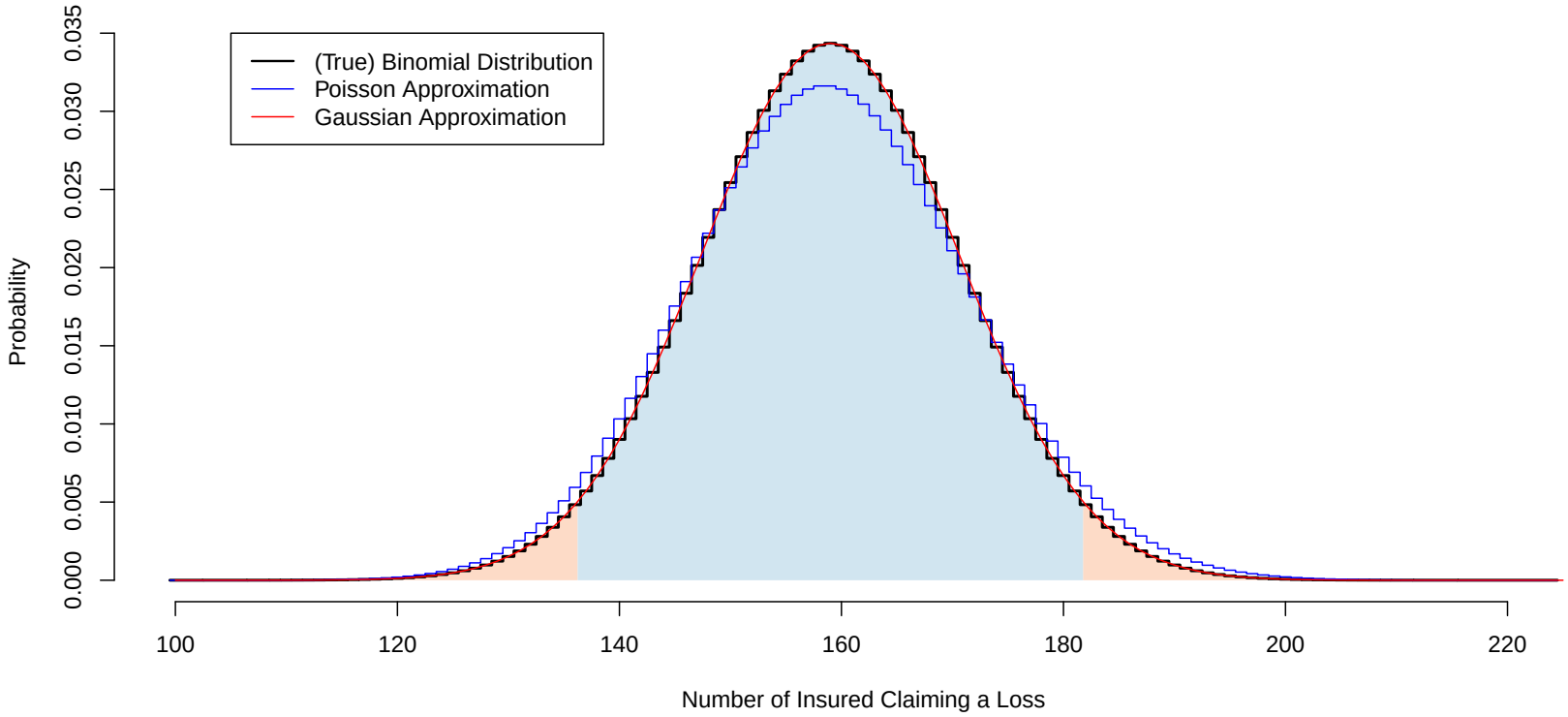
$$\sqrt{n}\frac{\widehat{p}-p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ as } n \to \infty$$

we can derive an approximated 95% confidence interval

$$\left[\widehat{p} \pm \frac{1.96}{\sqrt{n}}\sqrt{\widehat{p}(1-\widehat{p})}\right]$$

Bayes vs. Frequentist, inference on heads/tails

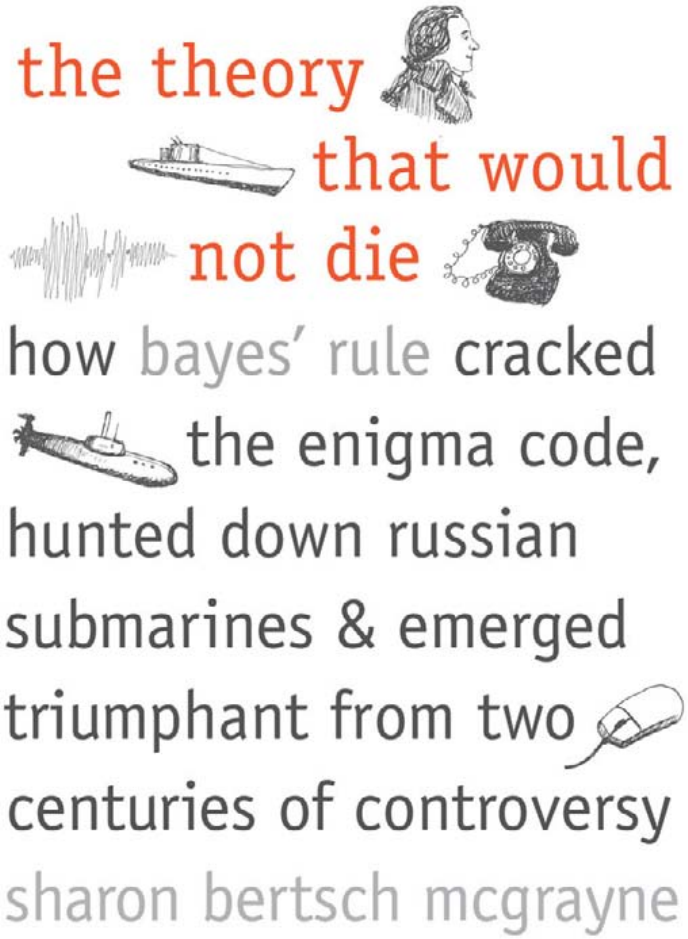
Example out of 1,047 contracts, 159 claimed a loss



Small Data and Black Swans

Example [Operational risk] What if our sample is $\boldsymbol{x} = \{0, 0, 0, 0, 0\}$?
How would we derive a confidence interval for p ?

"INA's chief executive officer, dressed as Santa Claus, asked an unthinkable question: Could anyone predict the probability of two planes colliding in midair? Santa was asking his chief actuary, L. H. Longley-Cook, to make a prediction based on no experience at all. There had never been a serious midair collision of commercial planes. Without any past experience or repetitive experimentation, any orthodox statistician had to answer Santa's question with a resounding no."

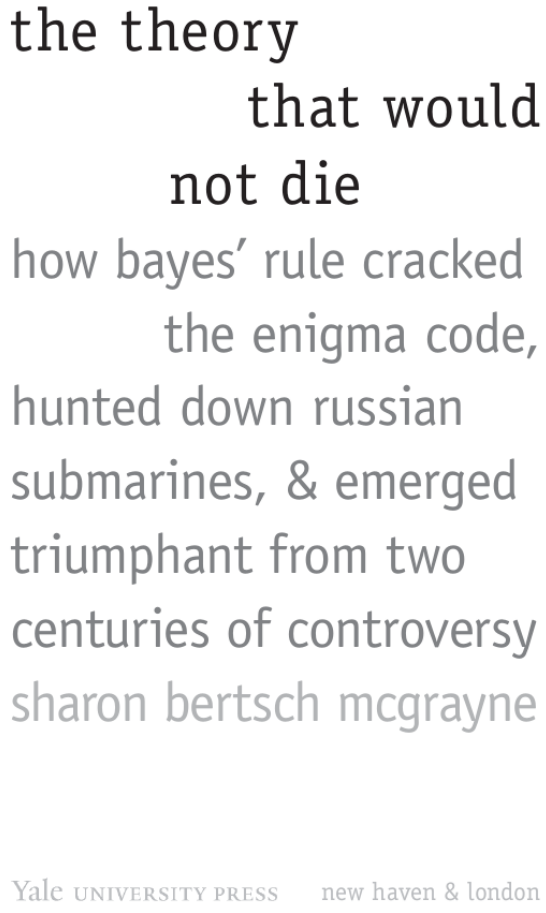


Bayes, *the theory that would not die*

Liu et al. (1996) claim that " Statistical methods with a Bayesian flavor [...] have long been used in the insurance industry".

History of Bayesian statistics, the theory that would not die by Sharon Bertsch McGrayne

"[Arthur] Bailey spent his first year in New York [in 1918] trying to prove to himself that 'all of the fancy actuarial [Bayesian] procedures of the casualty business were mathematically unsound.' After a year of intense mental struggle, however, realized to his consternation that actuarial sledgehammering worked" [...]



Bayes, *the theory that would not die*

[...] " He even preferred it to the elegance of frequentism. He positively liked formulae that described 'actual data . . . I realized that the hard-shelled underwriters were recognizing certain facts of life neglected by the statistical theorists.' He wanted to give more weight to a large volume of data than to the frequentists small sample; doing so felt surprisingly 'logical and reasonable'. He concluded that only a 'suicidal' actuary would use Fishers method of maximum likelihood, which assigned a zero probability to nonevents. Since many businesses file no insurance claims at all, Fishers method would produce premiums too low to cover future losses."

6.

arthur bailey

After the Second World War the first public challenge to the anti-Bayesian status quo came not from the military or university mathematicians and statisticians but from a Bible-quoting business executive named Arthur L. Bailey.

Bailey was an insurance actuary whose father had been fired and blackballed by every bank in Boston for telling his employers they should not be lending large sums of money to local politicians. So ostracized was the family that even Arthur's schoolmates stopped inviting him and his sister to parties. Turning his back on the New England establishment, Bailey enrolled at the University of Michigan in Ann Arbor. There he studied statistics in the mathematics department's actuarial program, earned a bachelor of science degree in 1928, and met his wife, Helen, who became an actuary for John Hancock Mutual Life before their children were born.[1]

Bailey's first job was, he liked to say, "in bananas," that is, in the statistics department of the United Fruit Company headquarters in Boston. When the department was eliminated during the Depression, Bailey wound up driving a fruit truck and chasing escaped tarantulas down Boston streets. He was lucky to have the job, and his family never lacked for bananas and oranges.

In 1937, after nine years in bananas, Bailey got a job in an unrelated field in New York City. There he was in charge of setting premium rates to cover risks involving automobiles, aircraft, manufacturing, burglary, and theft for the American Mutual Alliance, a consortium of mutual insurance companies.

Preferring church and community connections to the fair-weather friends of his youth, Bailey hid his growing professional success by living quietly in unpretentious New York suburbs. He relaxed by gardening, hiking

91

Bayes's theorem

Consider some hypothesis H and some evidence E, then

$$\mathbb{P}_E(H) = \mathbb{P}(H|E) = \frac{\mathbb{P}(H \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(H) \cdot \mathbb{P}(E|H)}{\mathbb{P}(E)}$$

Bayes rule,

$$\begin{cases} \text{prior probability } \mathbb{P}(H) \\ \text{versus posterior probability after receiving evidence } E,\ \mathbb{P}_E(H) = \mathbb{P}(H|E). \end{cases}$$

In Bayesian (parametric) statistics, $H = \{\theta \in \Theta\}$ and $E = \{\boldsymbol{X} = \boldsymbol{x}\}$.

Bayes' Theorem,

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta) \cdot f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x})} = \frac{\pi(\theta) \cdot f(\boldsymbol{x}|\theta)}{\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta} \propto \pi(\theta) \cdot f(\boldsymbol{x}|\theta)$$

Small Data and Black Swans

Consider sample $\boldsymbol{x} = \{0, 0, 0, 0, 0\}$.

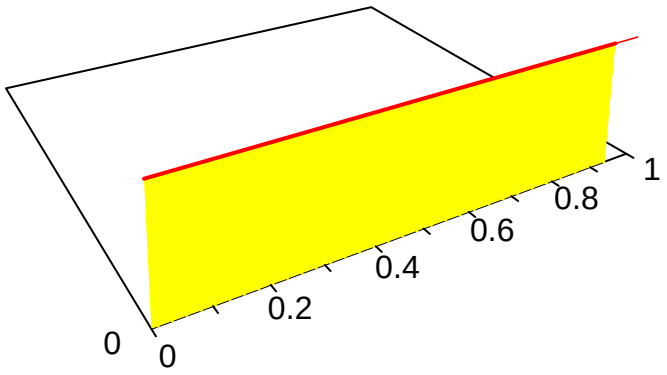
Here the likelihood is

$$\begin{cases} f(x_i|\theta) = \theta^{x_i}[1-\theta]^{1-x_i} \\ f(\boldsymbol{x}|\theta) = \theta^{\boldsymbol{x}^\intercal \mathbf{1}}[1-\theta]^{n-\boldsymbol{x}^\intercal \mathbf{1}} \end{cases}$$

and we need a priori distribution $\pi(\cdot)$ e.g. a beta distribution

$$\pi(\theta) = \frac{\theta^\alpha[1-\theta]^\beta}{B(\alpha,\beta)}$$

$$\pi(\theta|\boldsymbol{x}) = \frac{\theta^{\alpha+\boldsymbol{x}^\intercal \mathbf{1}}[1-\theta]^{\beta+n-\boldsymbol{x}^\intercal \mathbf{1}}}{B(\alpha+\boldsymbol{x}^\intercal \mathbf{1}, \beta+n-\boldsymbol{x}^\intercal \mathbf{1})}$$



On Bayesian Philosophy, Confidence vs. Credibility

for frequentists, a probability is a measure of the the frequency of repeated events

$\rightarrow$ parameters are fixed (but unknown), and data are random

for Bayesians, a probability is a measure of the degree of certainty about values

$\rightarrow$ parameters are random and data are fixed

"Bayesians : Given our observed data, there is a 95% probability that the true value of θ falls within the credible region

vs. Frequentists : There is a 95% probability that when I compute a confidence interval from data of this sort, the true value of θ will fall within it." in Vanderplas (2014)

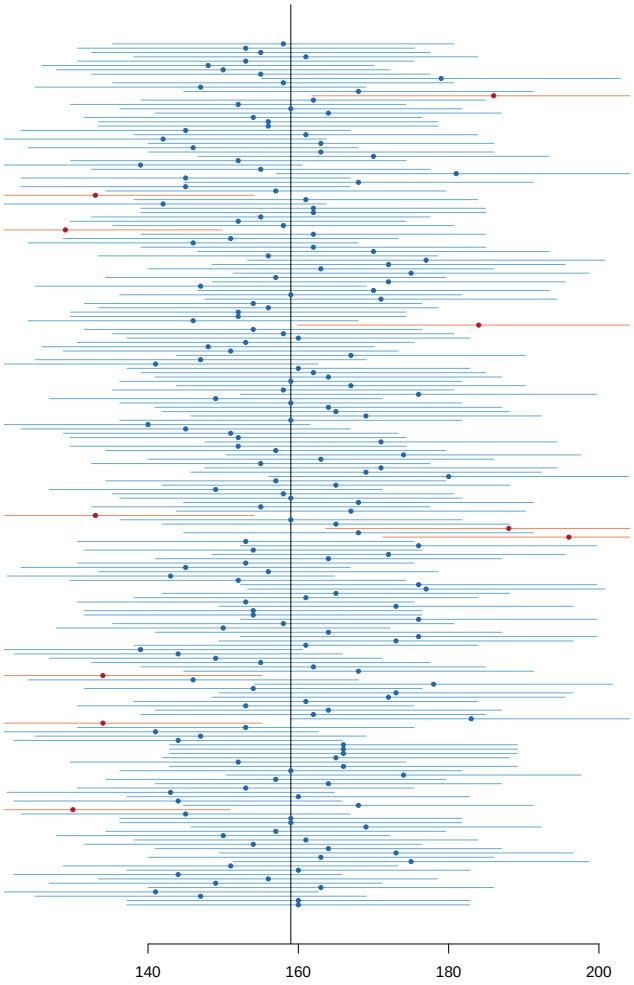
Example see Jaynes (1976), e.g. the truncated exponential

On Bayesian Philosophy, Confidence vs. Credibility

Example What is a 95% confidence interval of a proportion ? Here $\overline{x} = 159$ and $n = 1047$.

1. draw sets $(\tilde{x}_1, \cdots, \tilde{x}_n)_k$ with $X_i \sim \mathcal{B}(\overline{x}/n)$
2. compute for each set of values confidence intervals
3. determine the fraction of these confidence interval that contain $\overline{x}$

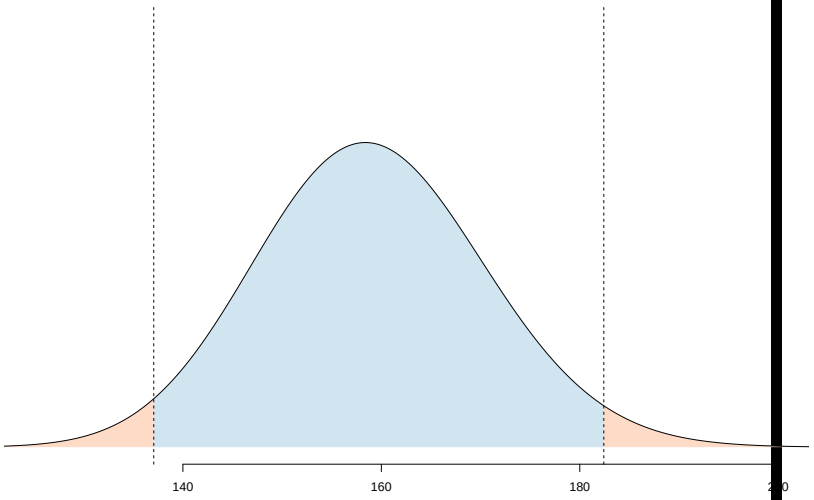
$\rightarrow$ the parameter is fixed, and we guarantee that 95% of the confidence intervals will contain it.



On Bayesian Philosophy, Confidence vs. Credibility

Example What is 95% credible region of a proportion ? Here $\overline{x} = 159$ and $n = 1047$.

1. draw random parameters p_k with from the posterior distribution, $\pi(\cdot|\boldsymbol{x})$
2. sample sets $(\tilde{x}_1, \cdots, \tilde{x}_n)_k$ with $X_{i,k} \sim \mathcal{B}(p_k)$
3. compute for each set of values means $\overline{x}_k$
4. look at the proportion of those $\overline{x}_k$ that are within this credible region $[\Pi^{-1}(.025|\boldsymbol{x}); \Pi^{-1}(.975|\boldsymbol{x})]$



$\rightarrow$ the credible region is fixed, and we guarantee that 95% of possible values of $\overline{x}$ will fall within it it.

Difficult concepts ? Difficult computations ?

We have a sample $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ i.i.d. from distribution $f_\theta(\cdot)$.

In predictive modeling, we need $\mathbb{E}(g(X)|\boldsymbol{x}) = \int g(x) f_{\theta|\boldsymbol{x}}(x) dx$ where

$$f_{\theta|\boldsymbol{x}}(x) = \int f_\theta(x) \cdot \pi(\theta|\boldsymbol{x}) d\theta$$

while prior density (without information $\boldsymbol{x}$) was

$$f_\theta(x) = \int f_\theta(x) \cdot \pi(\theta) d\theta$$

How can we derive $\pi(\theta|\boldsymbol{x})$?

Can we sample from $\pi(\theta|\boldsymbol{x})$ (use monte carlo technique to approximate the integral) ?

Computations not that simple... until the 90's : MCMC

Markov Chain

Stochastic process, $(X_t)_{t\in\mathbb{N}_\star}$, on some discrete space Ω

$$\mathbb{P}(X_{t+1}=y|X_t=x,\underline{\boldsymbol{X}}_{t-1}=\underline{\boldsymbol{x}}_{t-1})=\mathbb{P}(X_{t+1}=y|X_t=x)=P(x,y)$$

where P is a transition probability, that can be stored in a transition matrix, $\boldsymbol{P}=[P_{x,y}]=[P(x,y)]$.

Observe that $\mathbb{P}(X_{t+k}=y|X_t=x)=P_k(x,y)$ where $\boldsymbol{P}^k=[P_k(x,y)]$.

Under some condition, $\lim_{n\to\infty}\boldsymbol{P}^n=\boldsymbol{\Lambda}=[\boldsymbol{\lambda}^\intercal]$,

Problem given a distribution $\boldsymbol{\lambda}$, is it possible to generate a Markov Chain that converges to this distribution ?

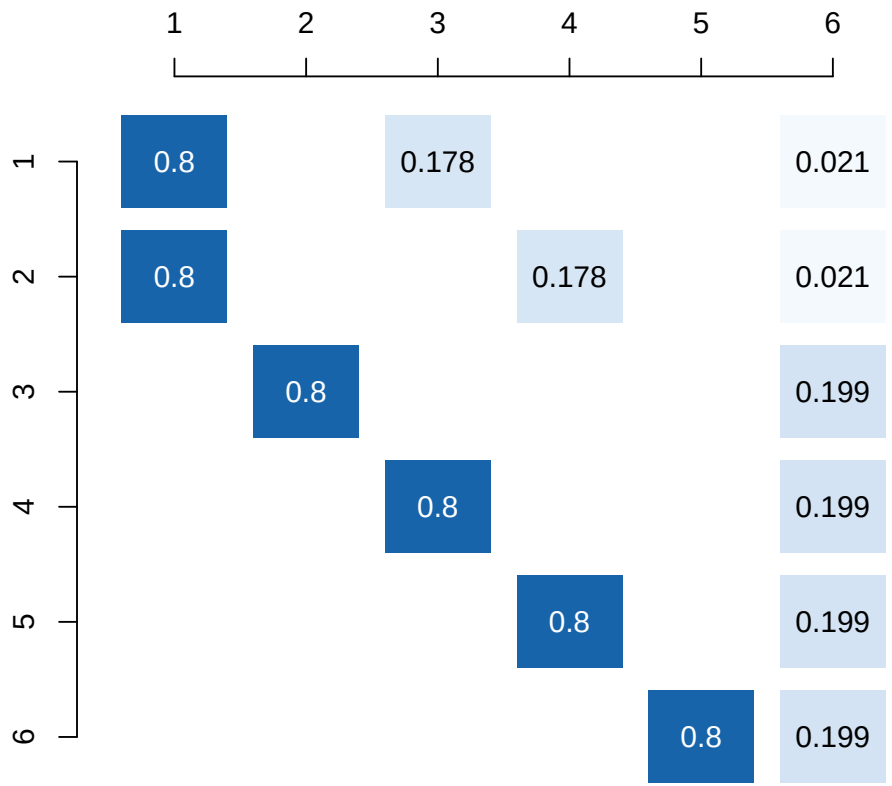
Bonus Malus and Markov Chains

Ex no-claim bonus, see Lemaire (1995).

HONG KONG
Table B-9. Hong Kong System

Class	Premium	Class After 0 Claims	Class After 1 Claim	Class After ≥2 Claims
6	100	5	6	6
5	80	4	6	6
4	70	3	6	6
3	60	2	6	6
2	50	1	4	6
1	40	1	3	6

Starting class: 6.



Assume that the number of claims is $N \sim \mathcal{P}(21.7\%)$, so that $\mathbb{P}(N = 0) = 80\%$.

Hastings-Metropolis

Back to our problem, we want to sample from $\pi(\theta|\boldsymbol{x})$

i.e. generate $\theta_1, \cdots, \theta_n, \cdots$ from $\pi(\theta|\boldsymbol{x})$.

Hastings-Metropolis sampler will generate a Markov Chain (θ_t) as follows,

- generate θ_1
- generate $\theta^\star$ and $U \sim \mathcal{U}([0,1])$,

  compute $R = \dfrac{\pi(\theta^\star|\boldsymbol{x})}{\pi(\theta_t|\boldsymbol{x})} \dfrac{P(\theta_t|\theta^\star)}{P(\theta^\star|\theta_{t-1})}$

  if $U < R$ set $\theta_{t+1} = \theta^\star$

  if $U \geq R$ set $\theta_{t+1} = \theta_t$

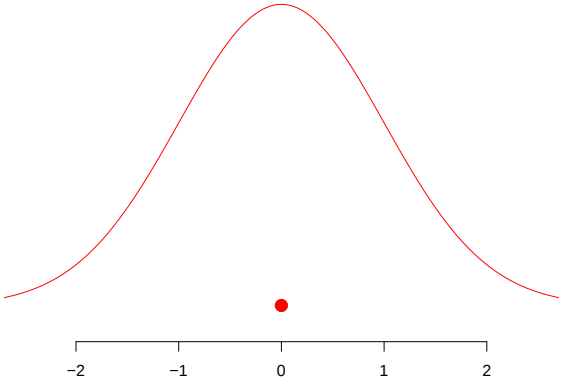
R is the acceptance ratio, we accept the new state $\theta^\star$ with probability $\min\{1, R\}$.

Hastings-Metropolis

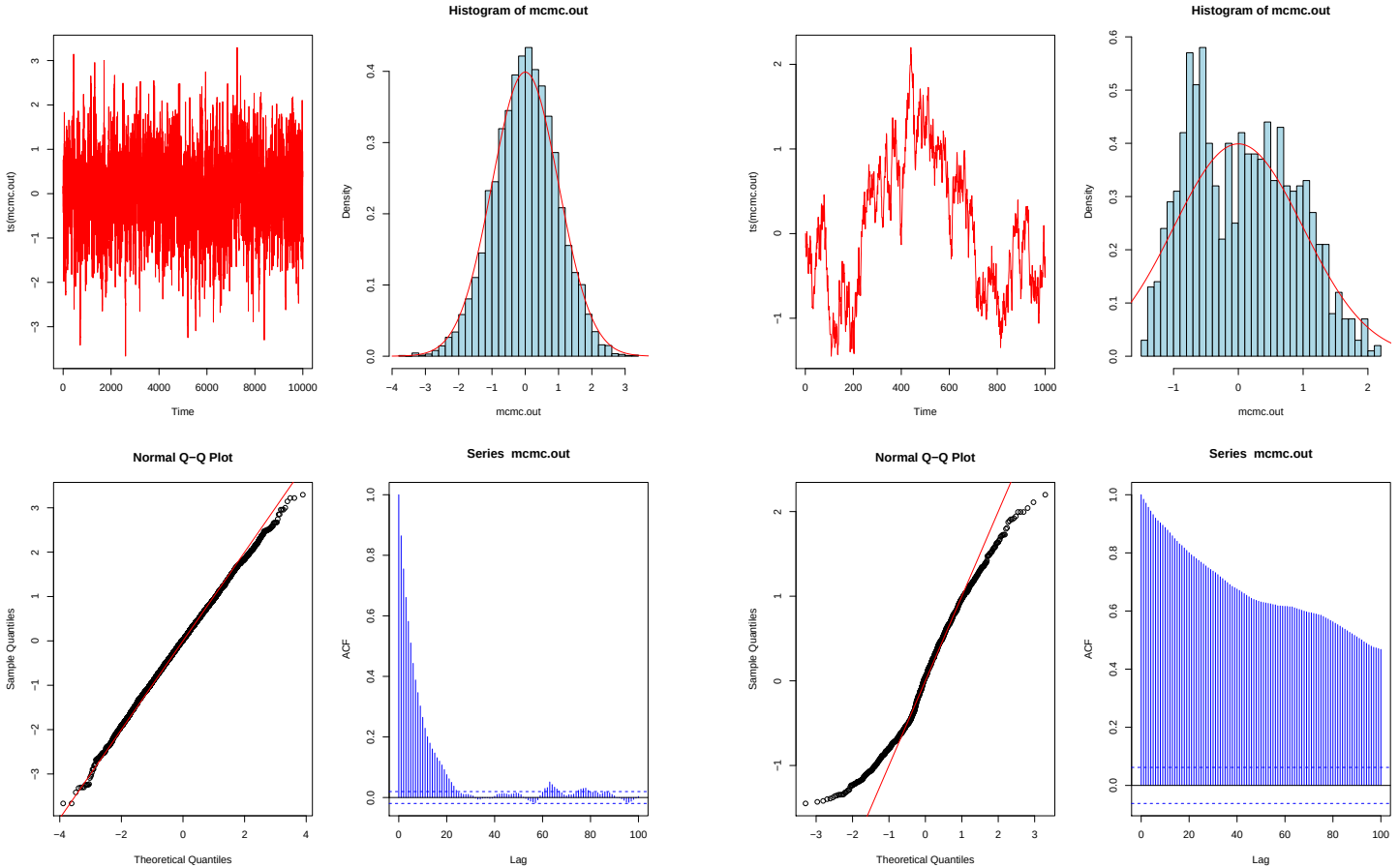
Observe that

$$R = \frac{\pi(\theta^\star) \cdot f(\boldsymbol{x}|\theta^\star)}{\pi(\theta_t) \cdot f(\boldsymbol{x}|\theta_t)} \frac{P(\theta_t|\theta^\star)}{P(\theta^\star|\theta_{t-1})}$$

In a more general case, we can have a Markov process, not a Markov chain.
E.g. $P(\theta^\star|\theta_t) \sim \mathcal{N}(\theta_t, 1)$



Using MCMC to generate Gaussian values



Heuristics on Hastings-Metropolis

In standard Monte Carlo, generate θ_i's i.i.d., then

$$\frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \to \mathbb{E}[g(\theta)] = \int g(\theta)\pi(\theta)d\theta$$

(strong law of large numbers).

Well-behaved Markov Chains ($\boldsymbol{P}$ aperiodic, irreducible, positive recurrent) can satisfy some ergodic property, similar to that LLN. More precisely,

- $\boldsymbol{P}$ has a unique stationary distribution λ, i.e. $\lambda = \lambda \times \boldsymbol{P}$
- ergodic theorem

$$\frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \to \int g(\theta)\lambda(\theta)d\theta$$

even if θ_i's are not independent.

Heuristics on Hastings-Metropolis

Remark The conditions mentioned above are

- aperiodic, the chain does not regularly return to any state in multiples of some k.
- irreducible, the state can go from any state to any other state in some finite number of steps
- positively recurrent, the chain will return to any particular state with probability 1, and finite expected return time

Gibbs Sampler

For a multivariate problem, it is possible to use Gibbs sampler.

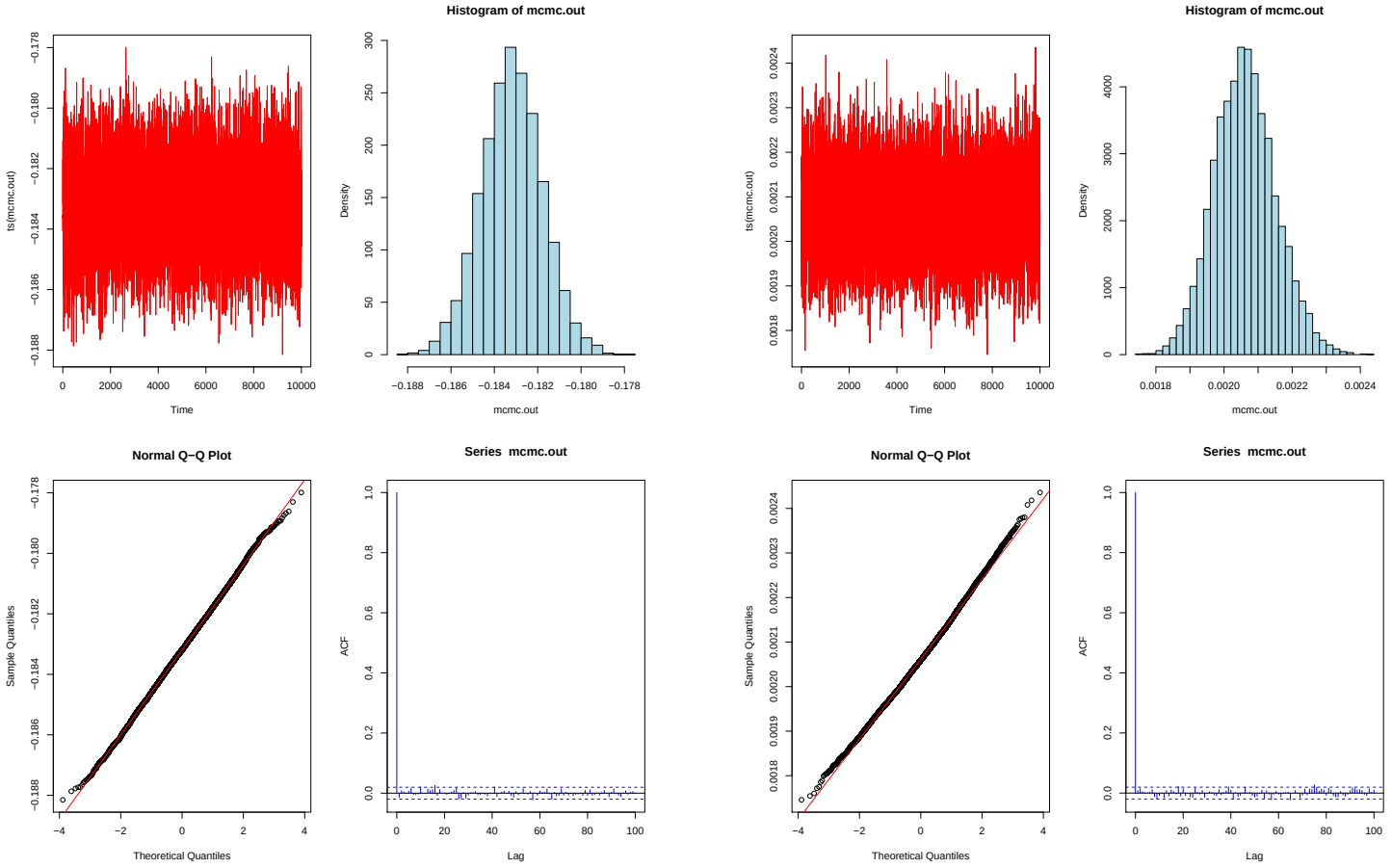
Example Assume that the loss ratio of a company has a lognormal distribution, $LN(\mu, \sigma^2)$, .e.g

Example Assume that we have a sample $\boldsymbol{x}$ from a $\mathcal{N}(\mu, \sigma^2)$. We want the posterior distribution of $\boldsymbol{\theta} = (\mu, \sigma^2)$ given $\boldsymbol{x}$. Observe here that if priors are Gaussian $\mathcal{N}\left(\mu_0, \tau^2\right)$ and the inverse Gamma distribution $IG(a, b)$, them

$$\begin{cases} \mu|\sigma^2, \boldsymbol{x} \sim \mathcal{N}\left(\dfrac{\sigma^2}{\sigma^2 + n\tau^2}\mu_0 + \dfrac{n\tau^2}{\sigma^2 + n\tau^2}\overline{x}, \dfrac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right) \\ \sigma^2|\mu, \boldsymbol{x} \sim IG\left(\dfrac{n}{2} + a, \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}[x_i - \mu]^2 + b\right) \end{cases}$$

More generally, we need the conditional distribution of $\theta_k|\boldsymbol{\theta}_{-k}, \boldsymbol{x}$, for all k.

Gibbs Sampler



Gibbs Sampler

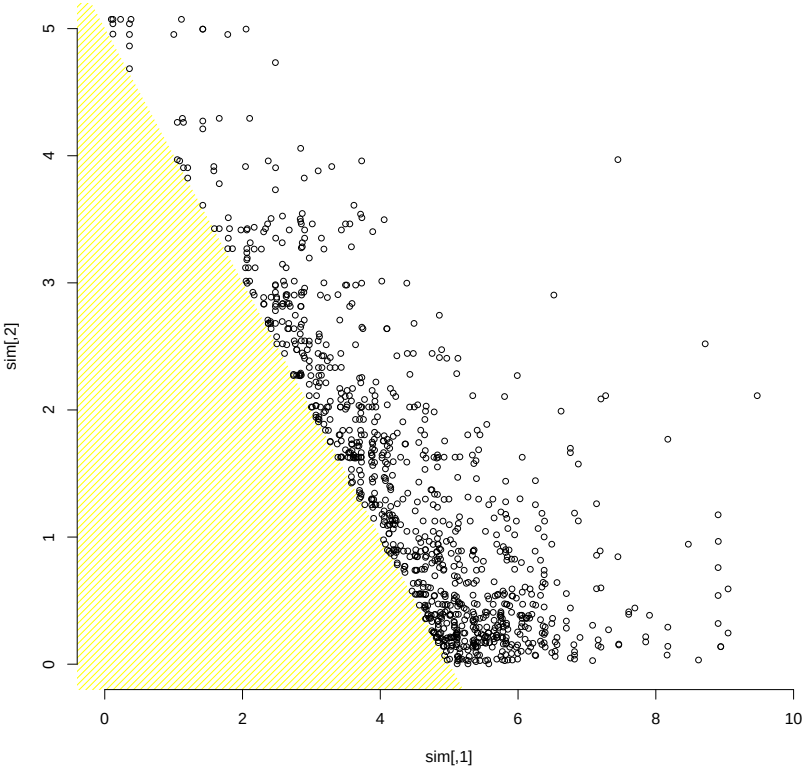
Example Consider some vector $\boldsymbol{X} = (X_1, \cdots, X_d)$ with indépendent components, $X_i \sim \mathcal{E}(\lambda_i)$. To sample from $\boldsymbol{X}$ given $\boldsymbol{X}^\mathsf{T}\mathbf{1} > s$ for some $s > 0$:

start with some starting point $\boldsymbol{x}_0$ such that $\boldsymbol{x}_0^\mathsf{T}\mathbf{1} > s$

pick up (randomly) $i \in \{1, \cdots, d\}$

X_i given $X_i > s - \boldsymbol{x}_{(-i)}^\mathsf{T}\mathbf{1}$ has an Exponential distribution $\mathcal{E}(\lambda_i)$

draw $Y \sim \mathcal{E}(\lambda_i)$ and set $x_i = y + (s - \boldsymbol{x}_{(-i)}^\mathsf{T}\mathbf{1})_+$ until $\boldsymbol{x}_{(-i)}^\mathsf{T}\mathbf{1} + x_i > s$



JAGS and STAN

Martyn Plummer developed JAGS *Just another Gibbs sampler* in 2007 (stable since 2013). It is an open-source, enhanced, cross-platform version of an earlier engine BUGS (Bayesian inference Using Gibbs Sampling).

STAN is a newer tool that uses the Hamiltonian Monte Carlo (HMC) sampler.

HMC uses information about the derivative of the posterior probability density to improve the algorithm. These derivatives are supplied by algorithm differentiation in C/C++ codes.

MCMC and Claims Reserving

Consider the following (cumulated) triangle, $\{C_{i,j}\}$,

	0	1	2	3	4	5
0	3209	4372	4411	4428	4435	4456
1	3367	4659	4696	4720	4730	4752.4
2	3871	5345	5398	5420	5430.1	5455.8
3	4239	5917	6020	6046.1	6057.4	6086.1
4	4929	6794	6871.7	6901.5	6914.3	6947.1
5	5217	7204.3	7286.7	7318.3	7331.9	7366.7
λ_j		1.3809	1.0114	1.0043	1.0018	1.0047
σ_j		0.7248	0.3203	0.04587	0.02570	0.02570

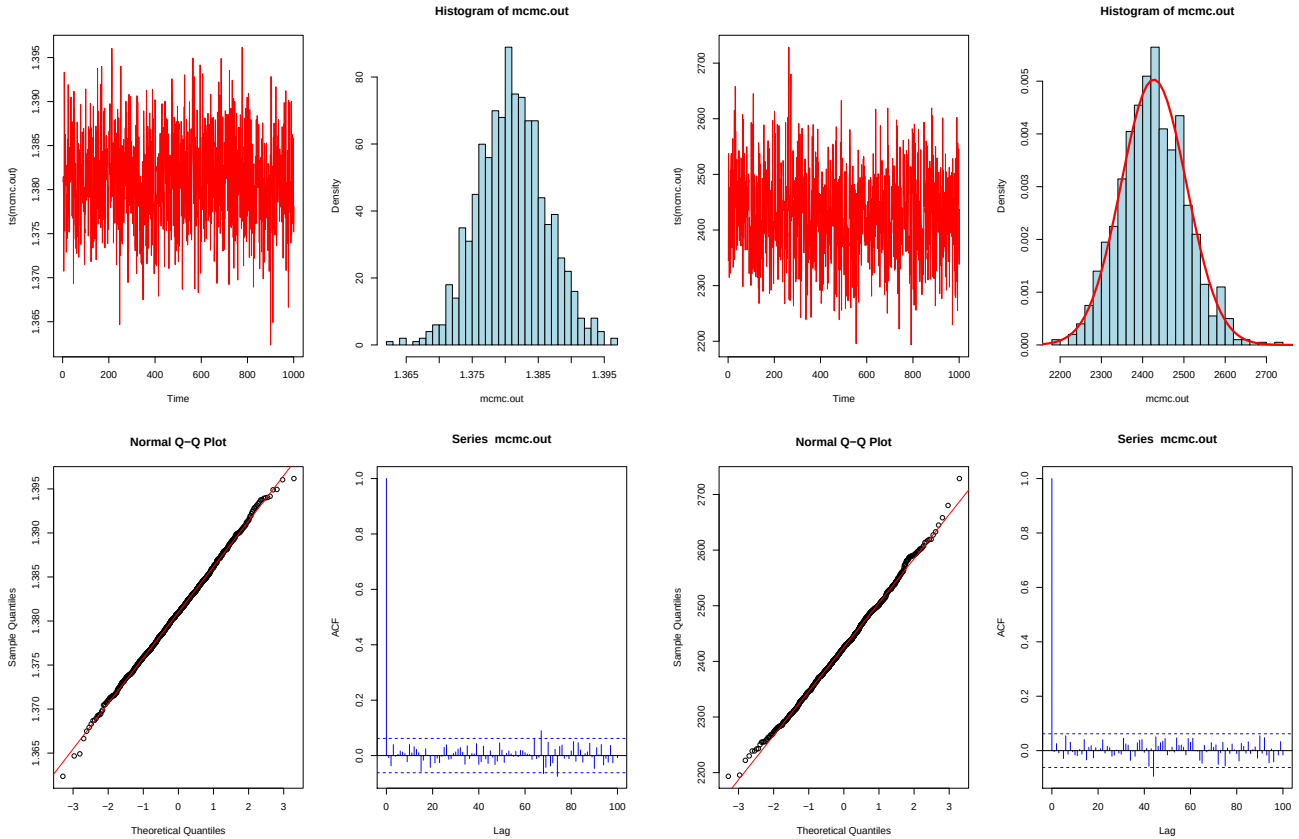
A Bayesian version of Chain Ladder

	0	1	2	3	4	5
0		1.362418	1.008920	1.003854	1.001581	1.004735
1		1.383724	1.007942	1.005111	1.002119	
2		1.380780	1.009916	1.004076		
3		1.395848	1.017407			
4		1.378373				
λ_j		1.3809	1.0114	1.0043	1.0018	1.0047
σ_j		0.7248	0.3203	0.04587	0.02570	0.02570

Assume that $\lambda_{i,j} \sim \mathcal{N}\left(\mu_j, \frac{\tau_j}{C_{i,j}}\right)$.

We can use Gibbs sampler to get the distribution of the transition factors, as well as a distribution for the reserves,

A Bayesian version of Chain Ladder

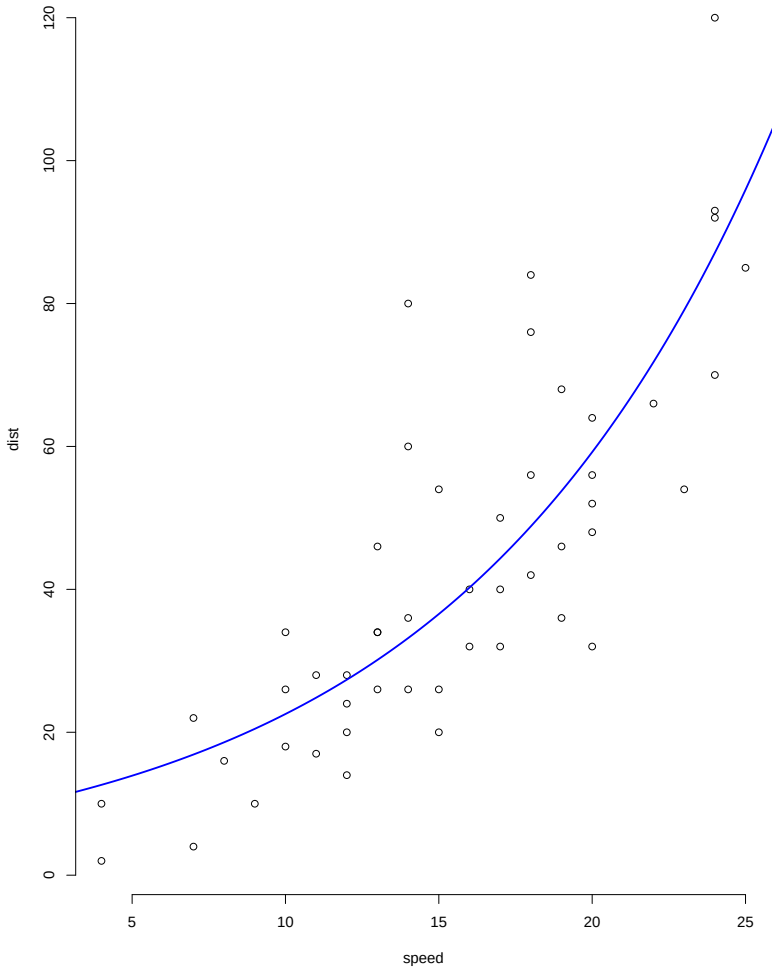


A Bayesian analysis of the Poisson Regression Model

In a Poisson regression model, we have a sample $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i)\}$,

$$y_i \sim \mathcal{P}(\mu_i) \text{ with } \log \mu_i = \beta_0 + \beta_1 x_i.$$

In the Bayesian framework, β_0 and β_1 are random variables.



Other alternatives to classical statistics

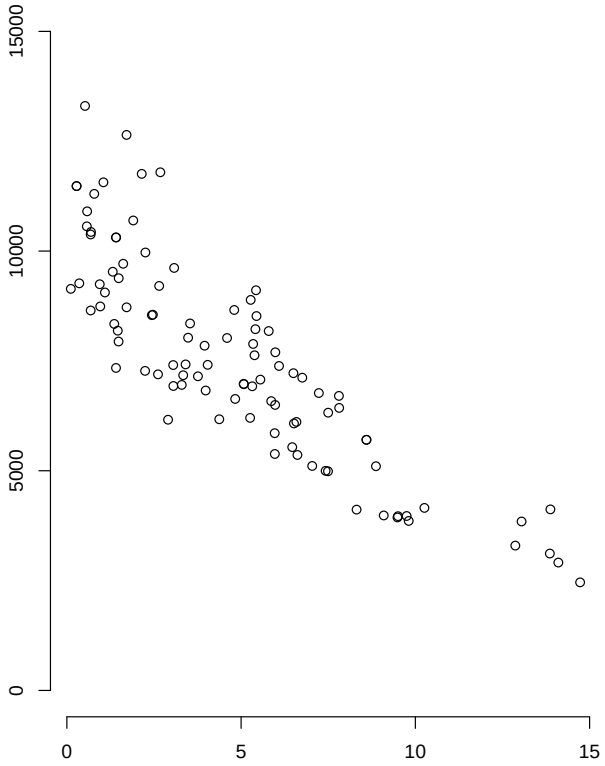
Consider a regression problem, $\mu(x) = \mathbb{E}(Y|X = x)$, and assume that smoothed splines are used,

$$\mu(x) = \sum_{i=1}^{k} \beta_j h_j(x)$$

Let $\boldsymbol{H}$ be the $n \times k$ matrix, $\boldsymbol{H} = [h_j(x_i)] = [\boldsymbol{h}(x_i)]$, then $\widehat{\boldsymbol{\beta}} = (\boldsymbol{H}^\mathsf{T}\boldsymbol{H})^{-1}\boldsymbol{H}^\mathsf{T}\boldsymbol{y}$, and

$$\widehat{\text{se}}(\widehat{\mu}(x)) = [\boldsymbol{h}(x)^\mathsf{T}(\boldsymbol{H}^\mathsf{T}\boldsymbol{H})^{-1}\boldsymbol{h}(x)]^{\frac{1}{2}}\widehat{\sigma}$$

With a Gaussian assumption on the residuals, we can derive (approximated) confidence bands for predictions $\widehat{\mu}(x)$.



Bayesian interpretation of the regression problem

Assume here that $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau\boldsymbol{\Sigma})$ as the priori distribution for $\boldsymbol{\beta}$.

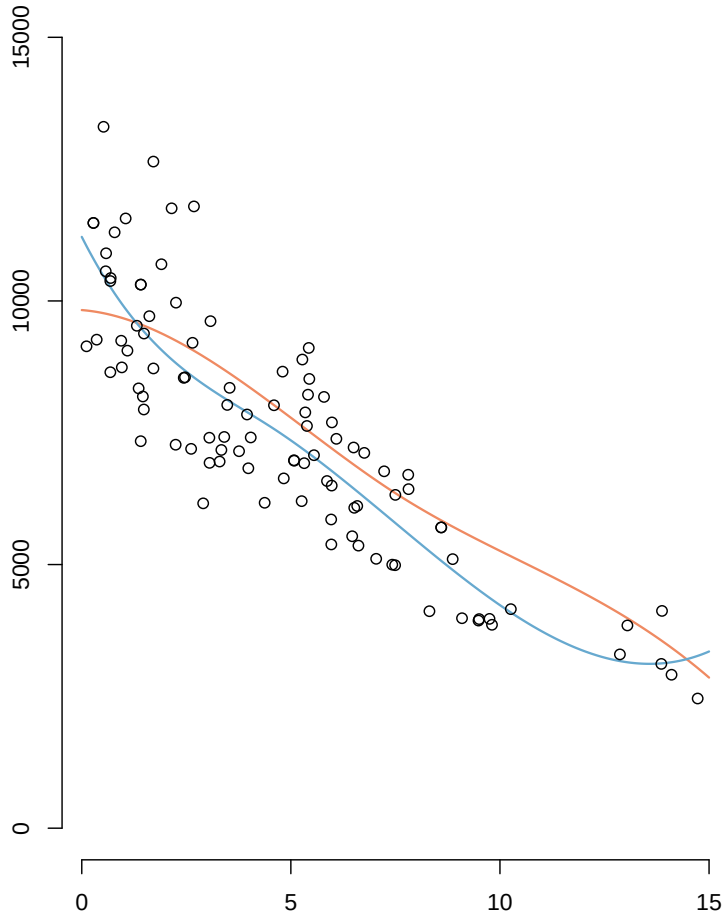
Then, if $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i), i = 1, \cdots, n\}$, the posterior distribution of $\mu(x)$ will be Gaussian, with

$$\mathbb{E}(\mu(x)|\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{h}(x)^\mathsf{T} \left(\boldsymbol{H}^\mathsf{T}\boldsymbol{H} + \frac{\sigma^2}{\tau}\boldsymbol{\Sigma}^{-1} \right)^{-1} \boldsymbol{H}^\mathsf{T}\boldsymbol{y}$$

$$\text{cov}(\mu(x), \mu(x')|\boldsymbol{x}, \boldsymbol{y})$$

$$= \boldsymbol{h}(x)^\mathsf{T} \left(\boldsymbol{H}^\mathsf{T}\boldsymbol{H} + \frac{\sigma^2}{\tau}\boldsymbol{\Sigma}^{-1} \right)^{-1} \boldsymbol{h}(x')\sigma^2$$

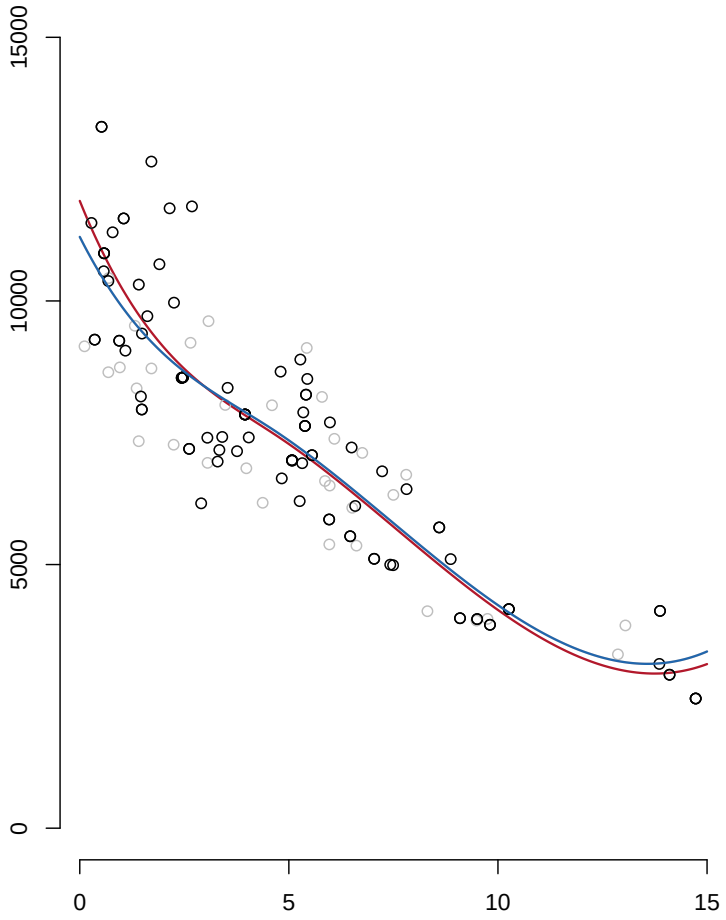
Example $\boldsymbol{\Sigma} = \mathbb{I}$



Bootstrap strategy

Assume that $Y = \mu(x) + \varepsilon$, and based on the estimated model, generate pseudo observations, $y_i^\star = \widehat{\mu}(x_i) + \widehat{\varepsilon}_i^\star$.
Based on $(\boldsymbol{x}, \boldsymbol{y}^\star) = \{(x_i, y_i^\star), i = 1, \cdots, n\}$, derive the estimator $\widehat{\mu}^\star(\cdot)$
(and repeat)

Observe that the bootstrap is the Bayesian case, when $\tau \to \infty$.



Part 5.
Data, Models & Actuarial Science (some sort of conclusion)

The Privacy-Utility Trade-Off

In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees

GIC has to publish the data: GIC(zip, date of birth, sex, diagnosis, procedure, ...)

Sweeney paid \$20 and bought the voter registration list for Cambridge Massachusetts, VOTER(name, party, ..., zip, date of birth, sex)

William Weld (former governor) lives in Cambridge, hence is in VOTER

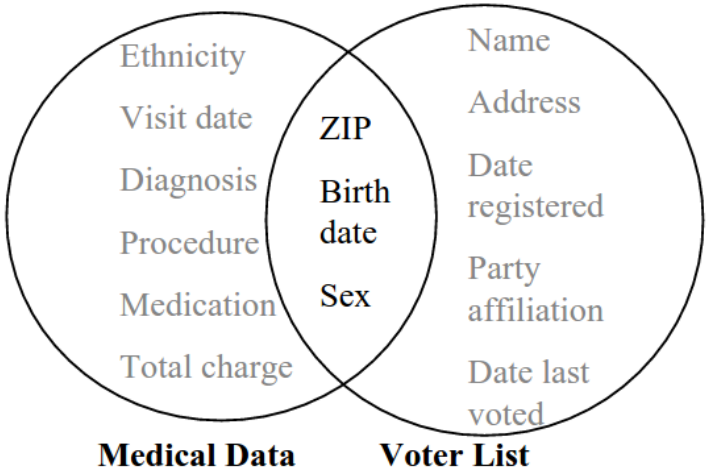


Figure 1 Linking to re-identify data

The Privacy-Utility Trade-Off

- 6 people in VOTER share his date of birth
- only 3 of them were man (same sex)
- Weld was the only one in that zip
- Sweeney learned Weld's medical records

All systems worked as specified, yet an important data was leaked.

"87% of Americans are uniquely identified by their zip code, gender and birth date", see Sweeney (2000).

A dataset is considered k-anonymous if the information for each person contained in the release cannot be distinguished from at least $k-1$ individuals whose information also appear in the release

manière des caractéristiques $\mathbf{\Omega}$ de l'assuré, et lui réclame donc une prime pure de montant $\mathbb{E}[S]$, la même que celle qu'il réclame à tous les assurés du portefeuille. Dans ce cas, la situation est telle que présentée au Tableau 3.7.

	Assurés	Assureur
Dépense	$\mathbb{E}[S]$	$S - \mathbb{E}[S]$
Dépense moyenne	$\mathbb{E}[S]$	0
Variance	0	$\mathbb{V}[S]$

TAB. 3.7 – *Situation des assurés et de l'assureur en l'absence de segmentation.*

L'assureur prend donc l'entièreté de la variance des sinistres $\mathbb{V}[S]$ à sa charge, que celle-ci soit due à l'hétérogénéité du portefeuille, ou à la variabilité intrinsèque des montants des sinistres.

Transfert de risque en information complète

A l'autre extrême, supposons que l'assureur incorpore toute l'information $\mathbf{\Omega}$ dans la tarification. On serait alors dans la situation décrite au Tableau 3.8.

	Assurés	Assureur
Dépense	$\mathbb{E}[S\|\mathbf{\Omega}]$	$S - \mathbb{E}[S\|\mathbf{\Omega}]$
Dépense moyenne	$\mathbb{E}[S]$	0
Variance	$\mathbb{V}\Big[\mathbb{E}[S\|\mathbf{\Omega}]\Big]$	$\mathbb{V}\Big[S - \mathbb{E}[S\|\mathbf{\Omega}]\Big]$

TAB. 3.8 – *Situation des assurés et de l'assureur dans le cas où la segmentation est opérée sur base de* $\mathbf{\Omega}$.

Contrairement au cas précédent, la prime payée par un assuré prélevé au hasard dans le portefeuille est à présent une variable aléatoire: $\mathbb{E}[S|\mathbf{\Omega}]$ dépend des caractéristiques $\mathbf{\Omega}$ de cet assuré. Comme la variable aléatoire $S - \mathbb{E}[S|\mathbf{\Omega}]$ est centrée, le risque assumé par l'assureur la variance du résultat financier de l'opération d'assurance, i.e.

$$\mathbb{V}\Big[S - \mathbb{E}[S|\mathbf{\Omega}]\Big] = \mathbb{E}\Big[\Big(S - \mathbb{E}[S|\mathbf{\Omega}]\Big)^2\Big]$$

No segmentation

	Insured	Insurer
Loss	$\mathbb{E}[S]$	$S - \mathbb{E}[S]$
Average Loss	$\mathbb{E}[S]$	0
Variance	0	$\text{Var}[S]$

Perfect Information: $\mathbf{\Omega}$ observable

	Insured	Insurer
Loss	$\mathbb{E}[S\|\mathbf{\Omega}]$	$S - \mathbb{E}[S\|\mathbf{\Omega}]$
Average Loss	$\mathbb{E}[S]$	0
Variance	$\text{Var}\Big[\mathbb{E}[S\|\mathbf{\Omega}]\Big]$	$\text{Var}\Big[S - \mathbb{E}[S\|\mathbf{\Omega}]\Big]$

$$\text{Var}[S] = \underbrace{\mathbb{E}\Big[\text{Var}[S|\mathbf{\Omega}]\Big]}_{\to \text{ insurer}} + \underbrace{\text{Var}\Big[\mathbb{E}[S|\mathbf{\Omega}]\Big]}_{\to \text{ insured}}.$$

On assiste dans ce cas à un partage de la variance totale de S (c'est-à-dire du risque) entre les assurés et l'assureur, matérialisé par la formule

$$\mathbb{V}[S] = \underbrace{\mathbb{E}\Big[\mathbb{V}[S|\boldsymbol{\Omega}]\Big]}_{\rightarrow\text{assureur}} + \underbrace{\mathbb{V}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big]}_{\rightarrow\text{assurés}}.$$

Ainsi, lorsque toutes les variables pertinentes $\boldsymbol{\Omega}$ ont été prises en compte, l'intervention de l'assureur se limite à la part des sinistres due exclusivement au hasard; en effet, $\mathbb{V}[S|\boldsymbol{\Omega}]$ représente les fluctuations de S dues au seul hasard. Dans cette situation idéale, l'assureur mutualise le risque et il n'y a donc aucune solidarité induite entre les assurés du portefeuille: chacun paie en fonction de son propre risque.

Transfert des risques en information partielle

Bien entendu, la situation décrite au paragraphe précédent est purement théorique puisque parmi les variables explicatives $\boldsymbol{\Omega}$ nombreuses sont celles qui ne peuvent pas être observées par l'assureur. En assurance automobile par exemple, l'assureur ne peut pas observer la vitesse à laquelle roule l'assuré, son agressivité au volant, ni le nombre de kilomètres qu'il parcourt chaque année[2]. Dès lors, l'assureur ne peut utiliser qu'un sous-ensemble $\boldsymbol{X}$ des variables explicatives contenues dans $\boldsymbol{\Omega}$, i.e. $\boldsymbol{X} \subset \boldsymbol{\Omega}$. La situation est alors semblable à celle décrite au Tableau 3.9.

	Assuré	Assureur
Dépense	$\mathbb{E}[S\|\boldsymbol{X}]$	$S - \mathbb{E}[S\|\boldsymbol{X}]$
Dépense moyenne	$\mathbb{E}[S]$	0
Variance	$\mathbb{V}\Big[\mathbb{E}[S\|\boldsymbol{X}]\Big]$	$\mathbb{E}\Big[\mathbb{V}[S\|\boldsymbol{X}]\Big]$

TAB. 3.9 – *Situation de l'assuré et de l'assureur dans le cas où la segmentation est opérée sur base de* $\boldsymbol{X} \subset \boldsymbol{\Omega}$.

Il est intéressant de constater que

$$\begin{aligned}\mathbb{E}\Big[\mathbb{V}[S|\boldsymbol{X}]\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{V}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big] + \mathbb{E}\Big[\mathbb{V}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big] \\ &= \underbrace{\mathbb{E}\Big[\mathbb{V}[S|\boldsymbol{\Omega}]\Big]}_{\text{mutualisation}} + \underbrace{\mathbb{E}\Big\{\mathbb{V}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big\}}_{\text{solidarité}}. \end{aligned} \quad (3.22)$$

Non-Perfect Information: $\boldsymbol{X} \subset \boldsymbol{\Omega}$ is observable

	Insured	Insurer
Loss	$\mathbb{E}[S\|\boldsymbol{X}]$	$S - \mathbb{E}[S\|\boldsymbol{X}]$
Average Loss	$\mathbb{E}[S]$	0
Variance	$\text{Var}\Big[\mathbb{E}[S\|\boldsymbol{X}]\Big]$	$\mathbb{E}\Big[\text{Var}[S\|\boldsymbol{X}]\Big]$

$$\begin{aligned}\mathbb{E}\Big[\text{Var}[S|\boldsymbol{X}]\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\text{Var}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big] \\ &+ \mathbb{E}\Big[\text{Var}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big] \\ &= \underbrace{\mathbb{E}\Big[\text{Var}[S|\boldsymbol{\Omega}]\Big]}_{\text{pooling}} \\ &+ \underbrace{\mathbb{E}\Big\{\text{Var}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big\}}_{\text{solidarity}}.\end{aligned}$$

SEGMENTATION ET MUTUALISATION
LES DEUX FACES D'UNE MÊME PIÈCE ?

Arthur Charpentier
Professeur à l'Université du Québec, Montréal

Michel Denuit
Professeur à l'Université catholique de Louvain

Romuald Elie
Professeur à l'Université de Marne-la-Vallée

L'assurance repose fondamentalement sur l'idée que la mutualisation des risques entre des assurés est possible. Cette mutualisation, qui peut être vue comme une relecture actuarielle de la loi des grands nombres, n'a de sens qu'au sein d'une population de risques « homogènes » [Charpentier, 2011]. Cette condition (actuarielle) impose aux assureurs de segmenter, ce que confirment plusieurs travaux économiques (1). Avec l'explosion du nombre de données, et donc de variables tarifaires possibles, certains assureurs évoquent l'idée d'un tarif individuel, semblant remettre en cause l'idée même de mutualisation des risques. Entre cette force qui pousse à segmenter et la force de rappel qui tend (pour des raisons sociales mais aussi actuarielles, ou au moins de robustesse statistique (2)) à imposer une solidarité minimale entre les assurés, quel équilibre va en résulter dans un contexte de forte concurrence entre les sociétés d'assurance ?

Tarification sans segmentation

Sans segmentation, le « prix juste » d'un risque est l'espérance mathématique de la charge annuelle. C'est l'idée du théorème fondamental de la valorisation actuarielle : en moyenne, la somme des primes doit permettre d'indemniser l'intégralité des sinistres survenus dans l'année. Afin d'illustrer les différents aspects de la construction du tarif et ses conséquences, on va utiliser les données présentées dans le tableau 1 (voir p. xx), qui indique la fréquence annuelle de sinistres.

Les facteurs de risque sont ici le lieu d'habitation et l'âge de l'assuré, et on observe la fréquence de sinistre par classe. Le coût unitaire, supposé fixe, équivaut à 1 000 euros. La prime pure est alors E [*S*] = 1 000 x E [*N*]. Dans cet exemple, la prime pure sans segmentation sera de 82,30 euros.

Simple model $\boldsymbol{\Omega} = \{\boldsymbol{X}_1, \boldsymbol{X}_2\}$.

Four Models

$$\begin{cases} \widehat{m}_0(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}[S] \\ \widehat{m}_1(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}[S|\boldsymbol{X}_1 = \boldsymbol{x}_1] \\ \widehat{m}_2(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}[S|\boldsymbol{X}_2 = \boldsymbol{x}_2] \\ \widehat{m}_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E}[S|\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2] \end{cases}$$

PolNum	CalYear	Gender	Type	Category	Occupation	Age	Group1	Bonus	Poldur	Value	Adind	SubGroup2	Group2	Density
200285786	2010	Male	E	Large	Employed	48	14	40	0	32345	1	O31	O	35.43401501
200285787	2010	Male	B	Medium	Employed	30	8	-30	9	8995	0	Q29	Q	239.4551701
200285788	2010	Female	B	Large	Housewife	47	2	-50	2	9145	1	U21	U	88.29014956
200285789	2010	Female	D	Large	Self-employed	48	13	-30	15	22075	1	R21	R	275.2822626
200285790	2010	Male	C	Medium	Housewife	57	12	-50	1	24985	1	Q5	Q	99.6400095
200285791	2010	Male	D	Medium	Self-employed	21	15	50	1	12100	1	R11	R	259.0040603
200285792	2010	Male	B	Small	Employed	44	5	-40	15	9820	1	Q10	Q	169.7885554
200285793	2010	Male	F	Small	Self-employed	37	17	-50	5	28680	1	Q5	Q	99.6400095
200285794	2010	Female	C	Large	Retired	49	3	20	4	28470	0	L94	L	84.22903844
200285795	2010	Female	A	Medium	Unemployed	35	5	20	5	8590	0	L112	L	66.06668352
200285796	2010	Male	E	Large	Self-employed	50	10	-30	3	20490	1	Q10	Q	169.7885554
200285797	2010	Female	B	Medium	Housewife	31	8	140	1	8385	1	P28	P	41.2451199
200285798	2010	Female	E	Medium	Self-employed	41	11	90	3	6410	1	L47	L	66.76541883
200285799	2010	Female	A	Medium	Housewife	44	10	-30	8	8485	0	P29	P	20.86448407
200285800	2010	Male	B	Large	Retired	69	8	-40	11	9380	1	U14	U	123.0152076
200285801	2010	Male	F	Medium	Housewife	45	11	30	0	19700	0	L40	L	76.05272599
200285802	2010	Male	E	Large	Retired	53	8	-30	6	10980	1	U19	U	61.79475865
200285803	2010	Male	C	Small	Employed	47	10	-10	9	21980	0	L96	L	45.66982293
200285804	2010	Female	D	Large	Retired	46	7	-50	1	28925	1	U12	U	54.93181221
200285805	2010	Female	C	Large	Retired	67	17	-50	9	14525	1	L52	L	73.25249905

$X_{1,i}$ $X_{k,i}$

Numtppd	Numtpbi	Indtppd	Indtpbi
0	1	0	1056.0334927
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
3	1	5800.0189068	16.507641942
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Y_i

PolNum	CalYear	Gender	Type	Category	Occupation	Age	Group1	Bonus	Poldur	Value	Adind	SubGroup2	Group2	Density
200375666	2011	Female	A	Large	Employed	46	11	50	0	42975	0	L18	L	58.91132801
200375667	2011	Male	B	Large	Unemployed	31	8	80	11	14835	0	U8	U	125.1320458
200375668	2011	Female	D	Medium	Employed	27	7	-40	13	19000	1	R30	R	296.4319078
200375670	2011	Male	B	Small	Self-employed	22	7	-10	14	33305	0	Q33	Q	129.6690079
200375672	2011	Male	B	Small	Employed	21	17	-20	14	25995	0	T25	T	28.51184808
200375674	2011	Male	C	Medium	Employed	45	19	-50	0	8320	1	N21	N	71.18027901
200375675	2011	Male	C	Medium	Housewife	51	19	30	3	8445	0	L110	L	83.90453994
200375676	2011	Male	E	Large	Self-employed	49	16	-50	3	19545	0	L58	L	64.53563007
200375677	2011	Male	C	Small	Housewife	31	11	-20	5	5030	1	Q7	Q	83.76263662
200375678	2011	Female	A	Medium	Housewife	31	9	-50	14	15480	1	P7	P	25.62227499
200375679	2011	Male	B	Large	Housewife	69	13	-50	7	29580	0	Q23	Q	205.4307964
200375682	2011	Male	A	Medium	Self-employed	43	13	140	3	3735	0	U16	U	91.54176264
200375683	2011	Female	A	Medium	Self-employed	64	18	-20	6	13670	1	O35	O	21.45273029
200375685	2011	Male	E	Large	Employed	25	8	-10	6	17315	0	O22	O	32.18545326
200375688	2011	Male	B	Small	Retired	55	7	-40	3	19410	1	R49	R	208.8164363
200375689	2011	Female	F	Medium	Self-employed	54	9	-40	14	4165	0	U12	U	54.93181221
200375690	2011	Male	D	Large	Housewife	42	9	80	0	11970	1	L125	L	44.16537902
200375692	2011	Male	E	Large	Employed	36	12	-20	7	28415	0	L48	L	71.62174491
200375693	2011	Male	F	Medium	Self-employed	26	10	-30	6	4300	0	L97	L	63.82886936
200375694	2011	Female	B	Small	Unemployed	24	6	-40	7	24005	0	M17	M	201.6569069

Market Competition

Decision Rule: the insured selects the cheapeast premium,

	A	B	C	D	E	F
	787.93	706.97	1032.62	907.64	822.58	**603.83**
	**170.04**	197.81	285.99	212.71	177.87	265.13
	473.15	447.58	**343.64**	410.76	414.23	425.23
	337.98	**336.20**	468.45	339.33	383.55	672.91

Market Competition

Decision Rule: the insured selects randomly from the three cheapeast premium

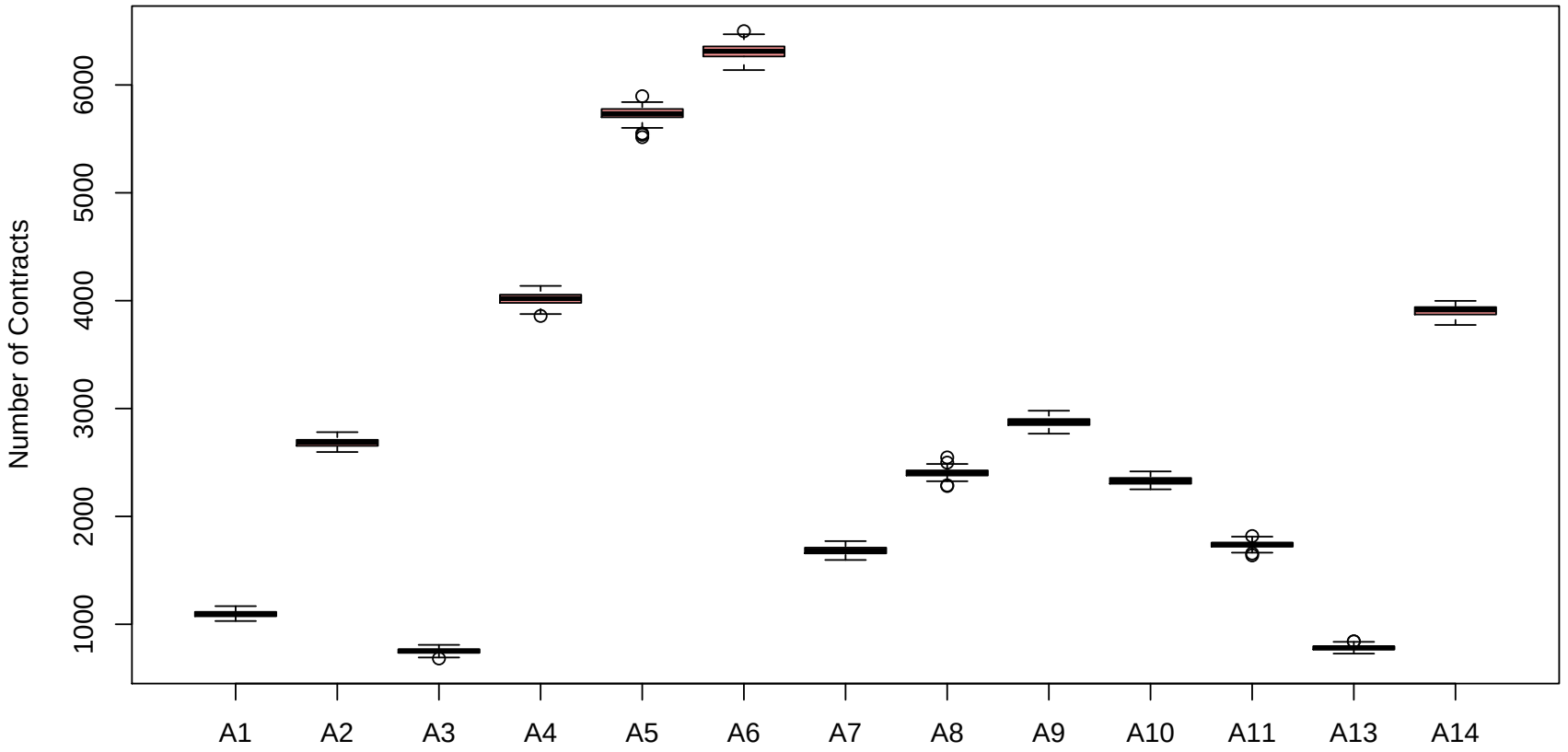
	A	B	C	D	E	F
	787.93	706.97	1032.62	907.64	822.58	603.83
	170.04	197.81	285.99	212.71	177.87	265.13
	473.15	447.58	343.64	410.76	414.23	425.23
	337.98	336.20	468.45	339.33	383.55	672.91

Market Competition

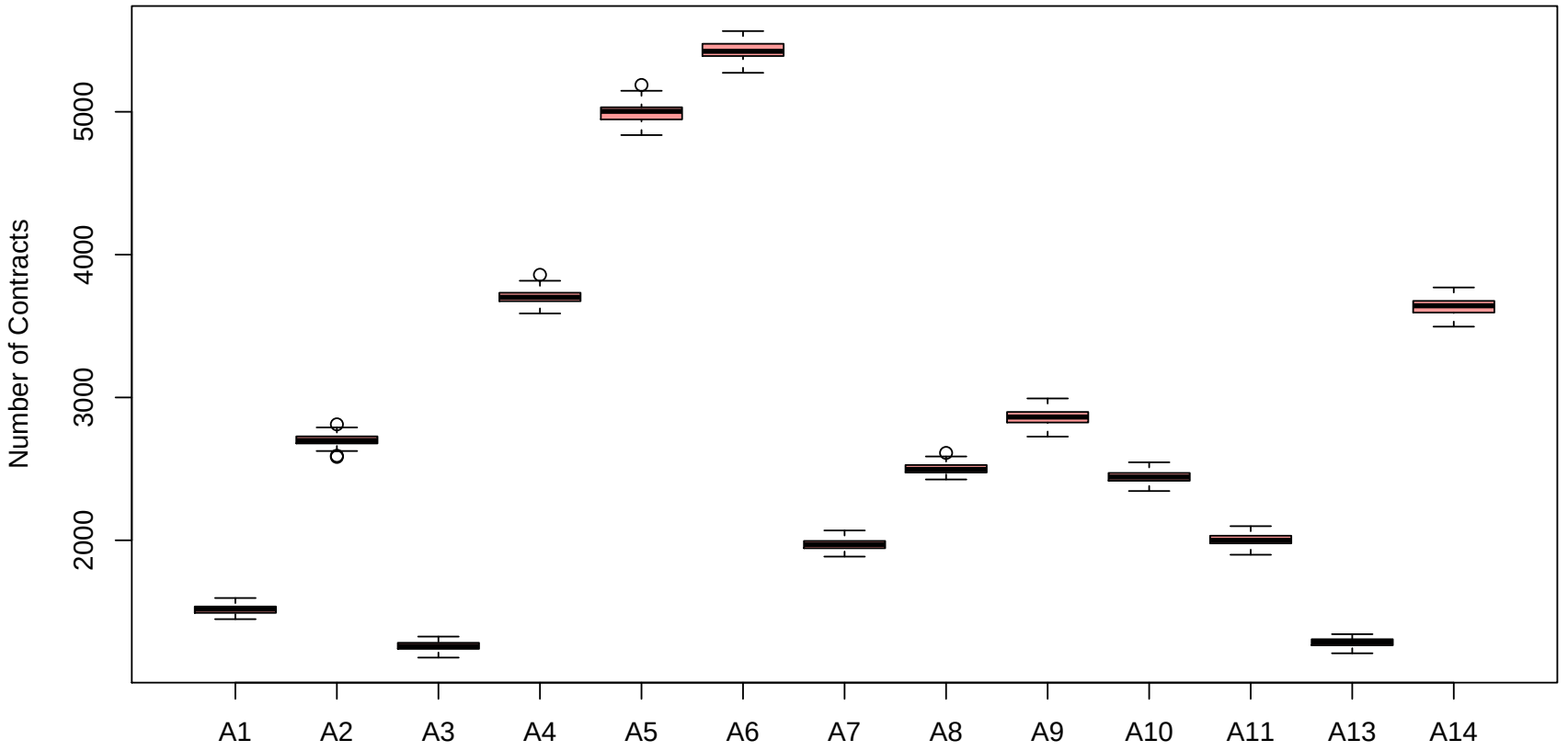
Decision Rule: the insured were assigned randomly to some insurance company for year $n-1$. For year n, they stay with their company if the premium is one of the three cheapeast premium, if not, random choice among the four

	A	B	C	D	E	F
	787.93	706.97	1032.62	907.64	822.58	603.83
	170.04	197.81	285.99	212.71	**177.87**	265.13
	473.15	447.58	343.64	410.76	414.23	425.23
	**337.98**	336.20	468.45	339.33	383.55	672.91

Market Shares (rule 2)

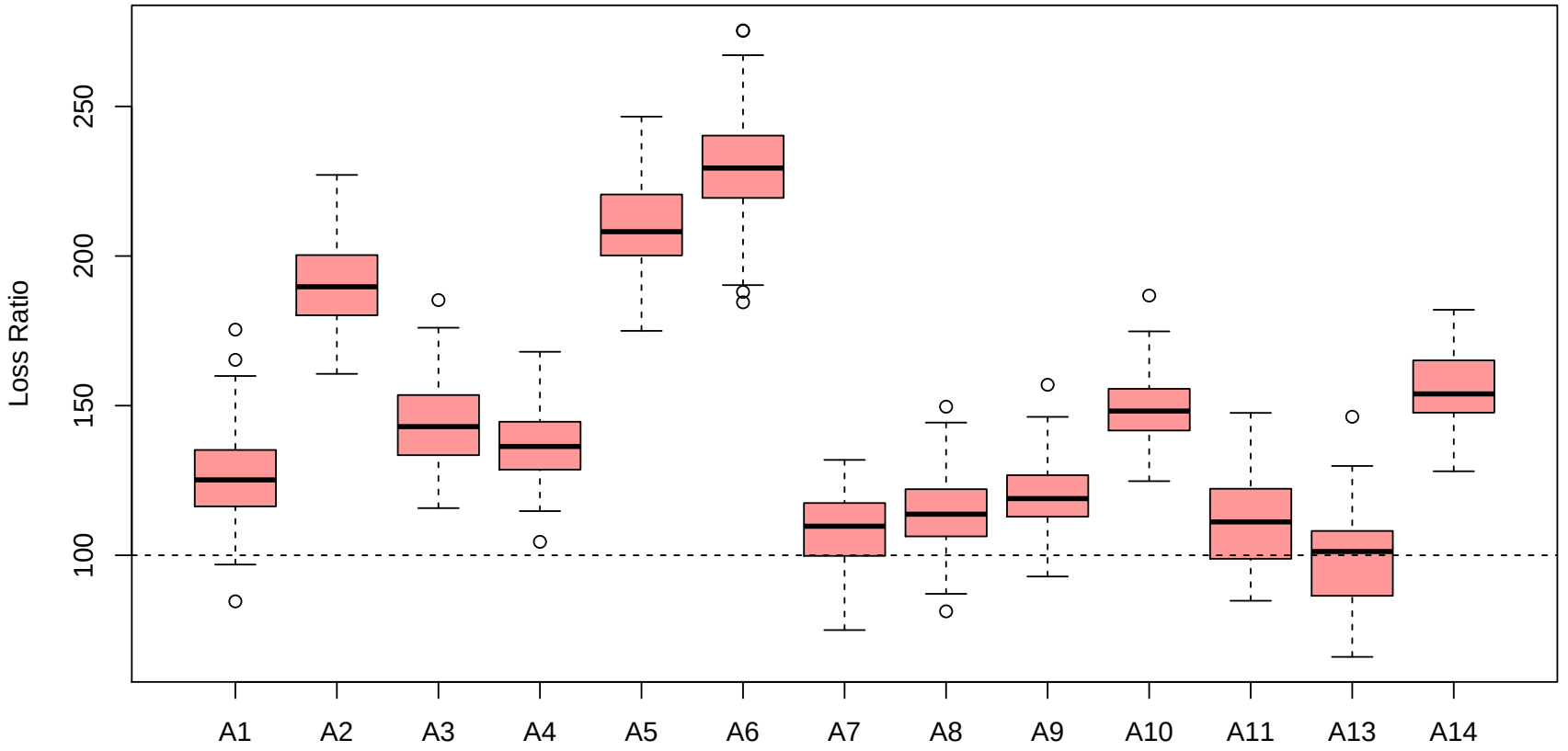


Market Shares (rule 3)



Loss Ratio, Loss / Premium (rule 2)

Market Loss Ratio $\sim 154\%$.

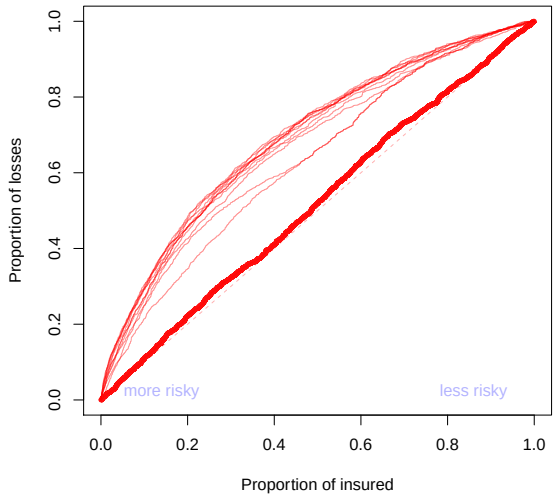


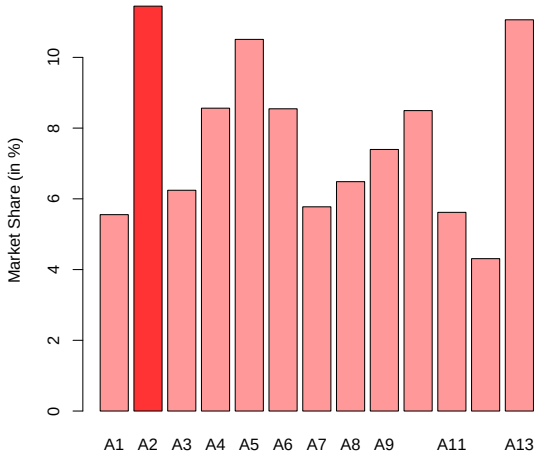
Insurer A2

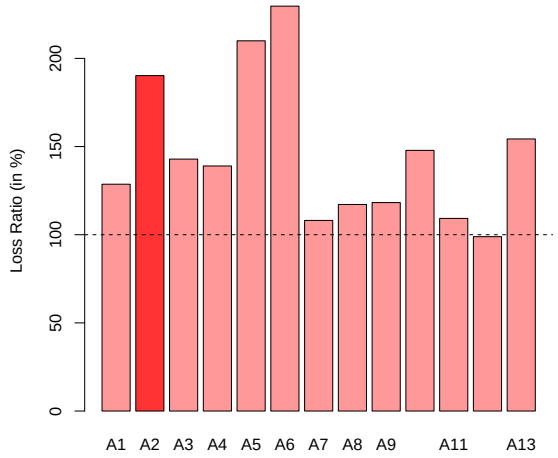
No segmentation, unique premium

Remark on normalized premiums,

$$\pi_2 = m_2(\boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^{n} m_j(\boldsymbol{x}_i) \ \forall j$$







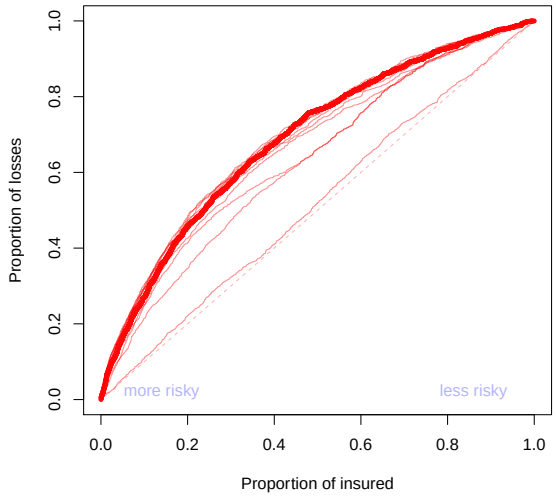
Insured A1

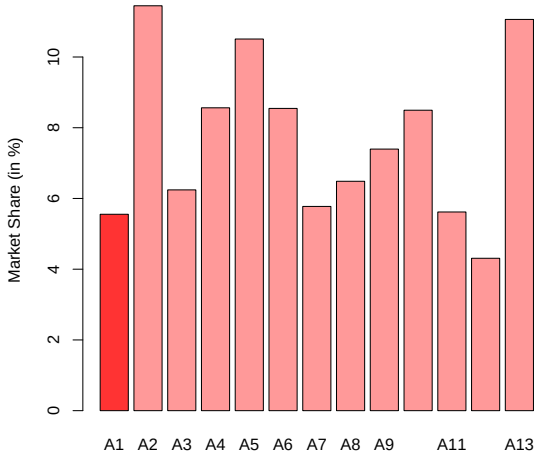
GLM, frequency material / bodily injury, individual losses material

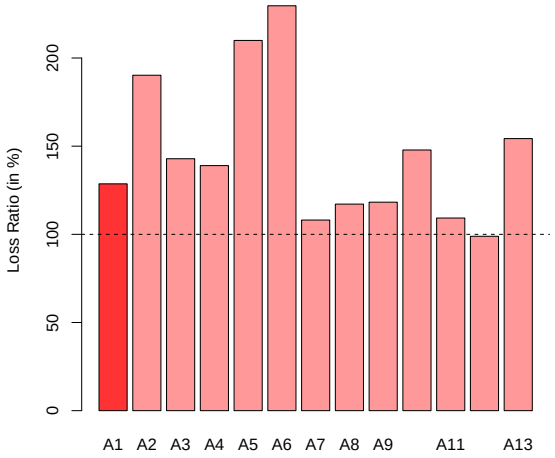
Ages in classes [18-30], [30-45], [45-60] and [60+], crossed with occupation

Manual smoothing, SAS and Excel

Actuaries in a Mutual Fund (in France)







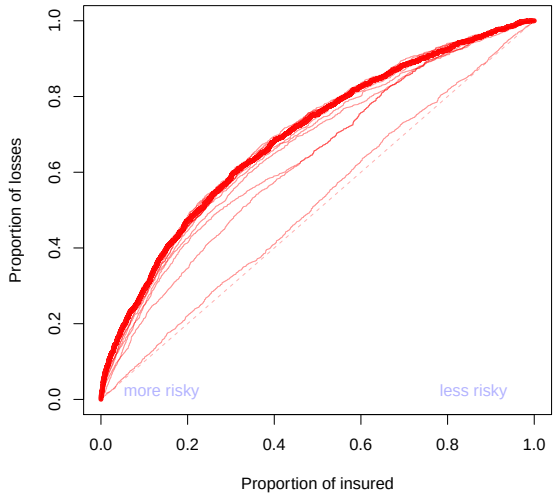
Insurer A8/A9

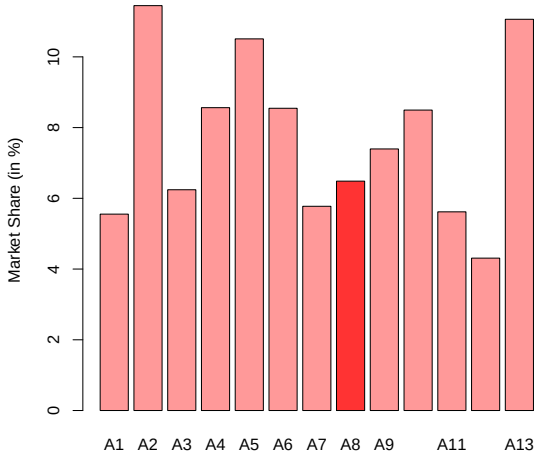
GLM, frequency and losses, without major losses (>15k)

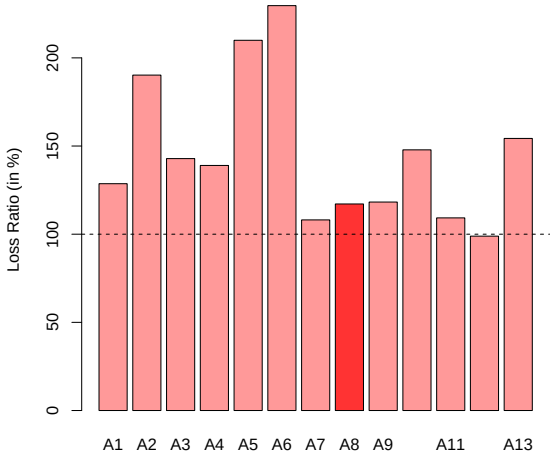
Age-gender interaction

Use of a commercial pricing software

Actuary in a French Mutual Fund





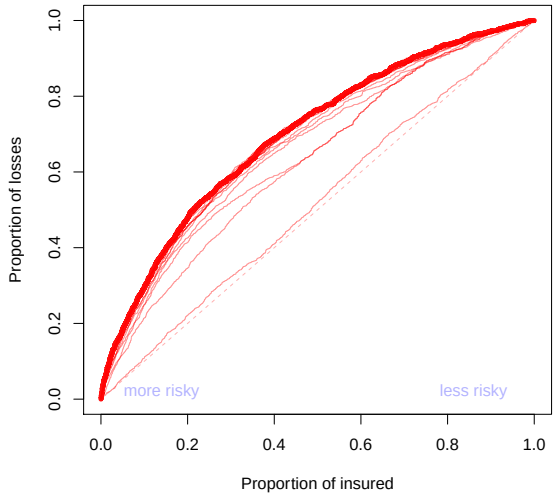


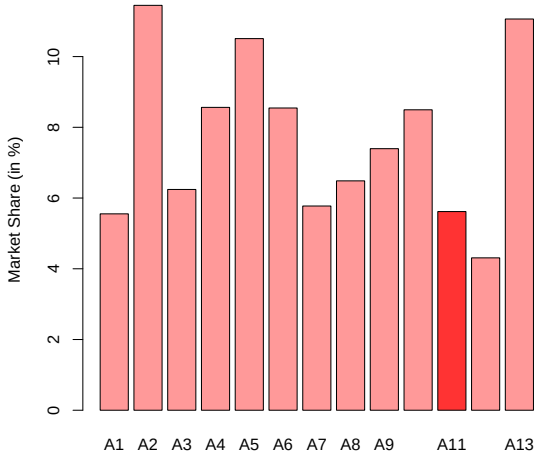
Insurer A11

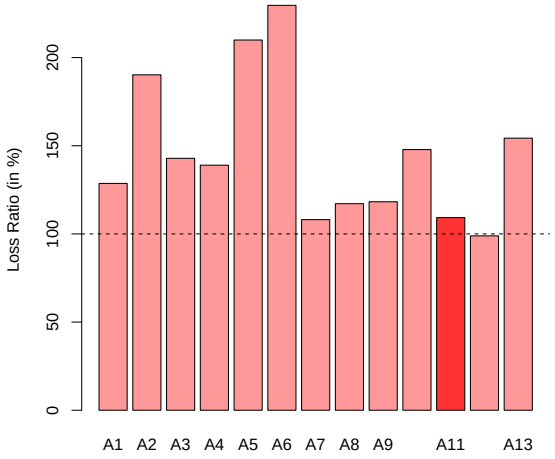
All features, but one XGBoost (gradient boosting)

Correction for negative premiums

Coded in Python actuary in an insurance company.





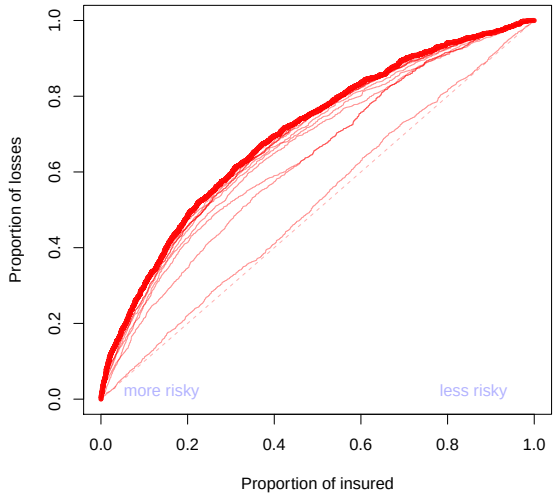


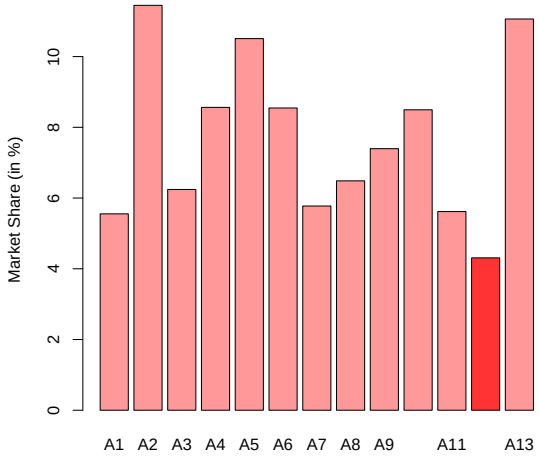
Insurer A12

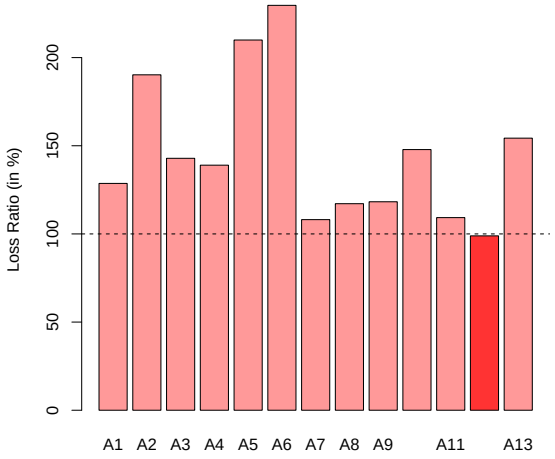
All features, use of two XGBoost (gradient boosting) models

Correction for negative premiums

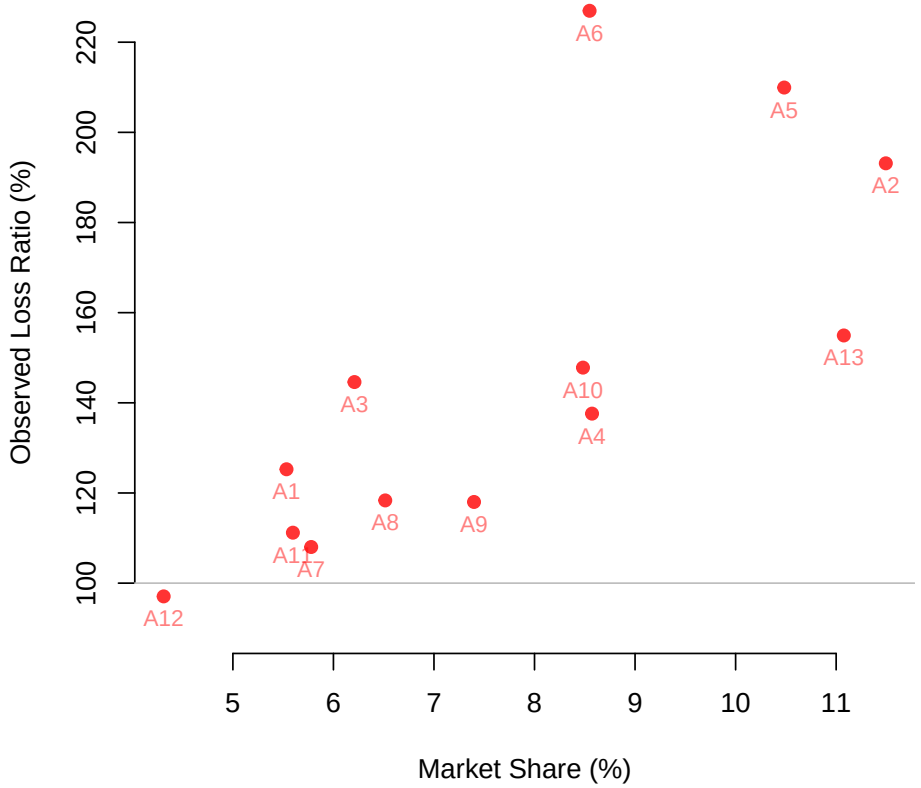
Coded in R by an actuary in an Insurance company.







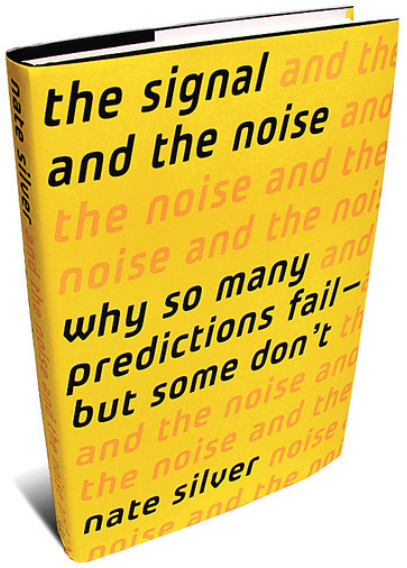
Back on the Pricing Game



Take-Away Conclusion

"People rarely succeed unless they have fun in what they are doing " D. Carnegie

- on very small datasets, it is possible to use Bayesian technique to derive robust predictions,
- on extremely large datasets, it is possible to use ideas developed in machine learning, on regression models (e.g. boostraping and aggregating)
- all those techniques require computational skills



"the numbers have no way of speaking for themselves. We speak for them. ... Before we demand more of our data, we need to demand more of ourselves " N. Silver, in Silver (2012).