Traitement statistique des données manquantes-Part I Introduction to modeling

Christophe Biernacki

Traitement statistique des données manquantes

Atelier Statistique de la SFdS – 10 et 11 mars 2021

–

Part I
Introduction to modeling

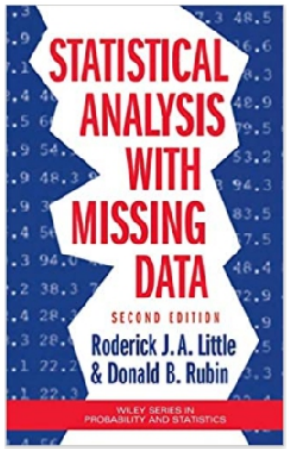
C. Biernacki

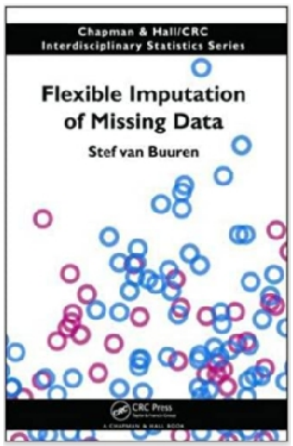
Laboratoire P. Painlevé, UMR CNRS 8524 & Université de Lille & Inria

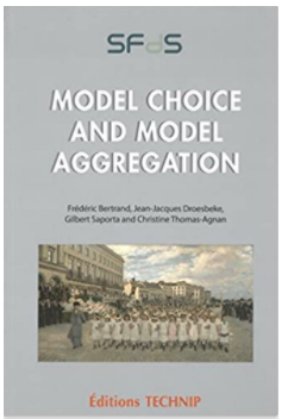
Merci à : F. Antonazzo, C. Boyer, G. Celeux, Q. Grimonprez, J. Jacques, J. Josse, C. Keribin, V. Kubicki, F. Laporte, M. Marbac, A. Sportisse, J. Vandaele, V. Vandewalle

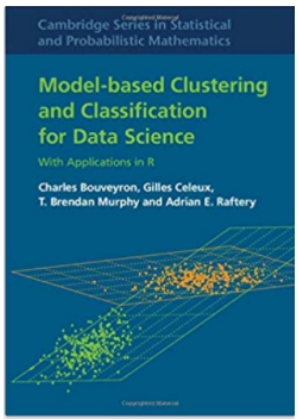
Lectures

Books about missing data. . . but also about modeling and its applications









Preamble

What is this lesson?

- Be able to perform practically some (well-established or advanced) methods on missing data
- Use them with discernment

What is not this lesson?

- Not an exhaustive list of missing data methods (and related bibliography)
- Do not make specialists of missing data methods

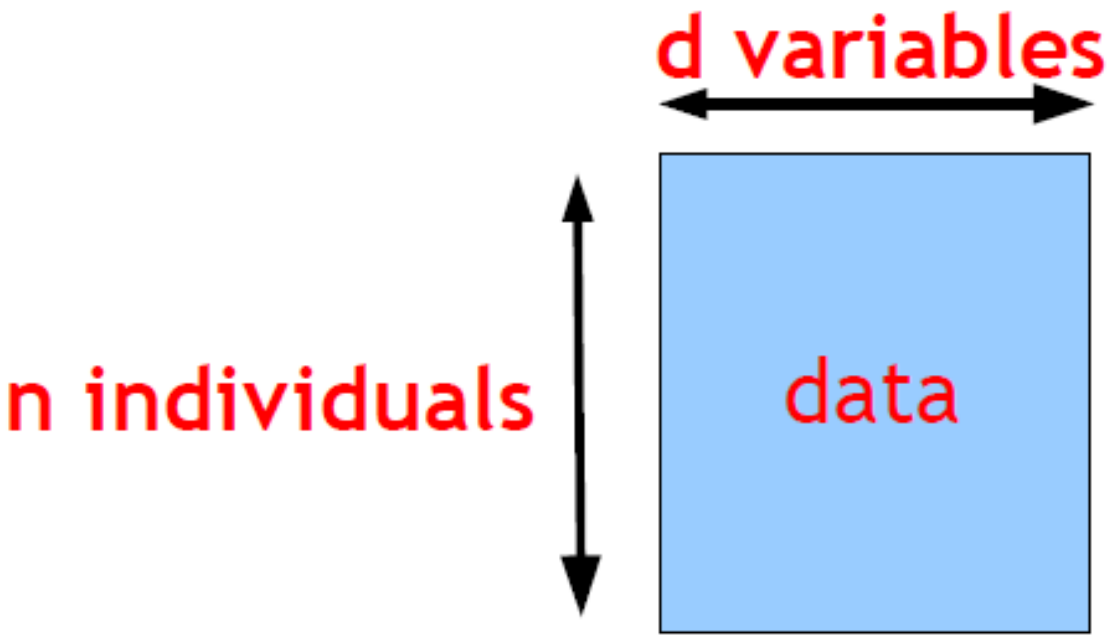
This preamble is valid for all four parts:

1. Part I: Introduction to modeling
2. Part II: Numerical and non-numerical data
3. Part III: Missing not at random data (MNAR) $\rightarrow$ ongoing research
4. Part IV: Binned data for big data analysis $\rightarrow$ ongoing research

Outline

Everything begins from data!

Data sets structure



Todays features: full mixed/missing

categorical
Marital status
maried

integer
Children
3

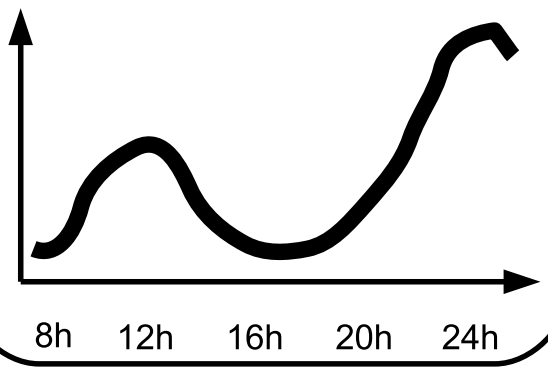
missing
Size (m)
?

rank
Drink preference
beer > soda > water

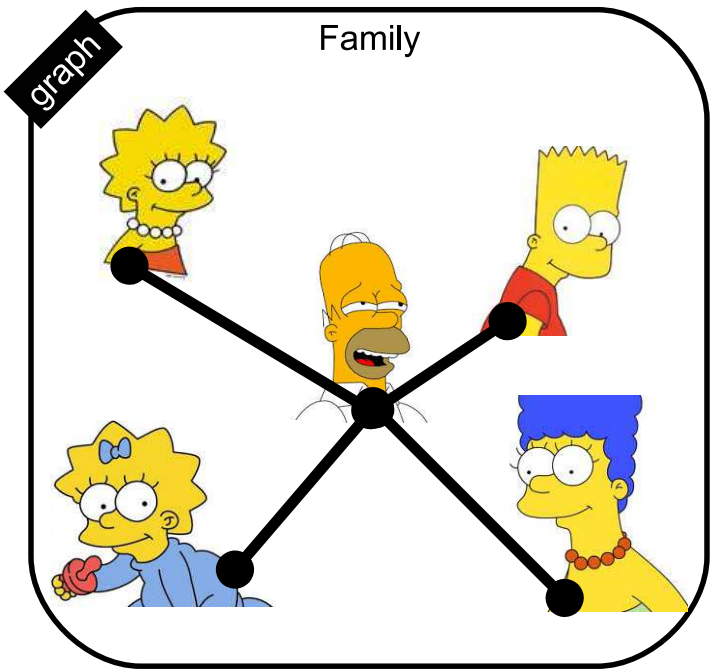
ordinal
Intelligence
low

continuous
Weight (kg)
119.5

functional

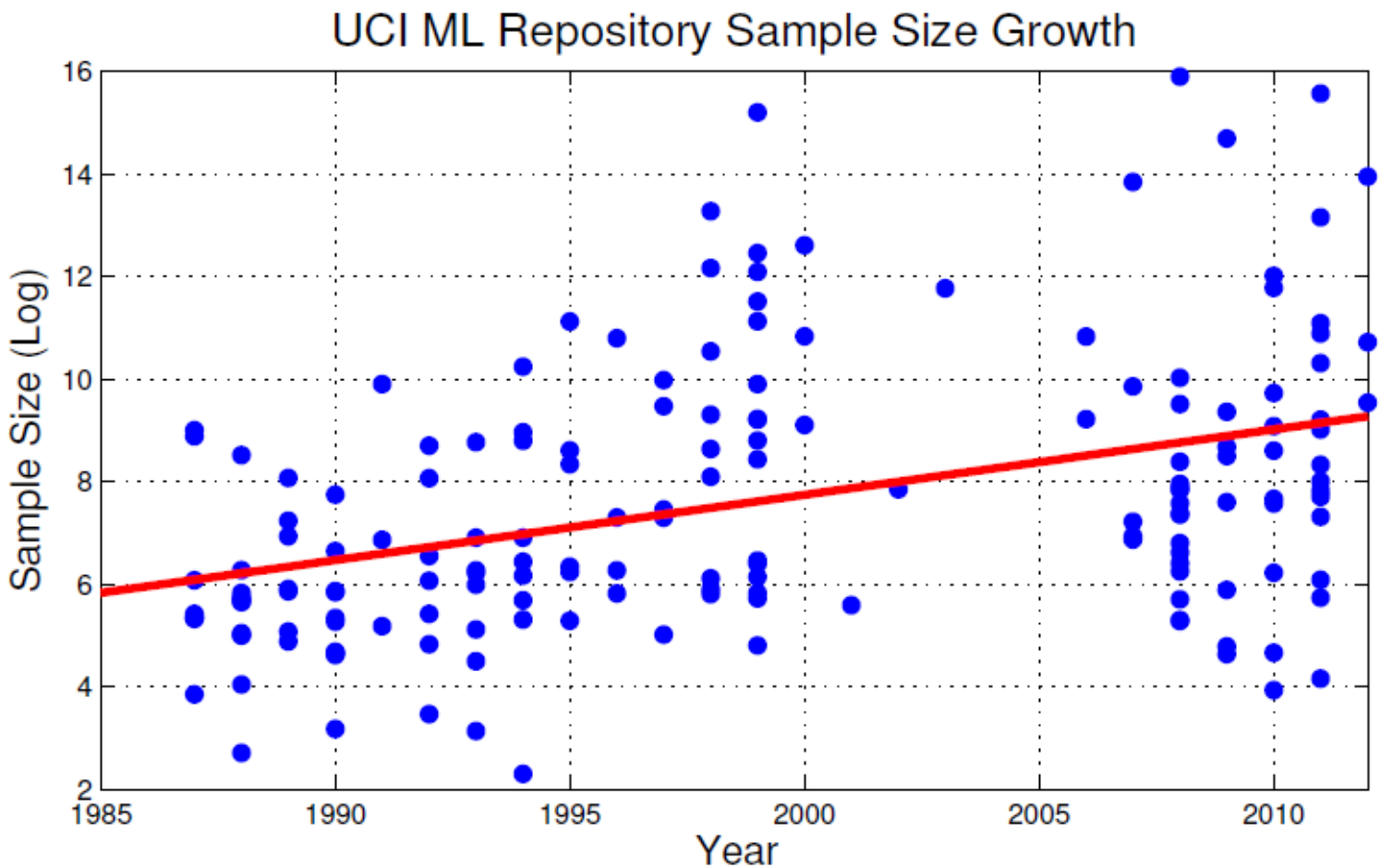


graph



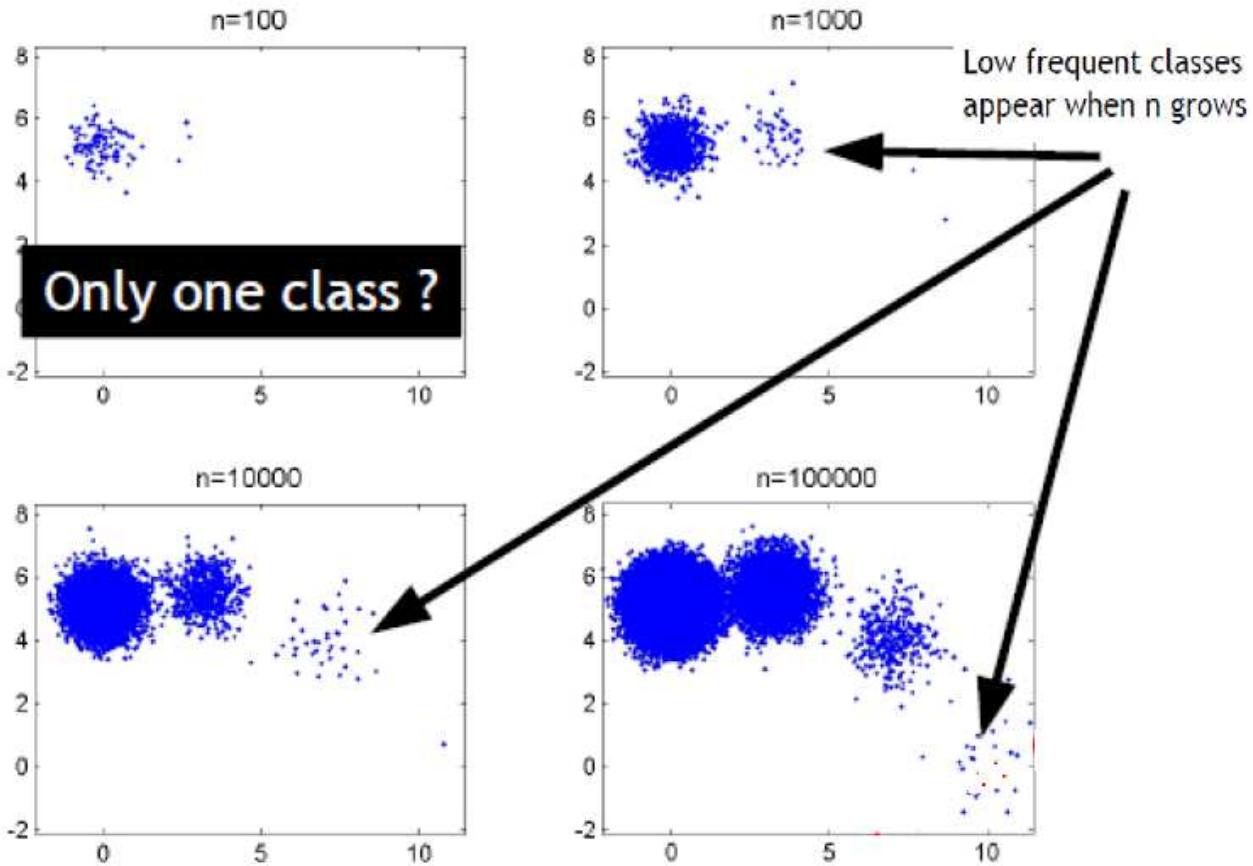
And so on...

Large data sets (n)[1]

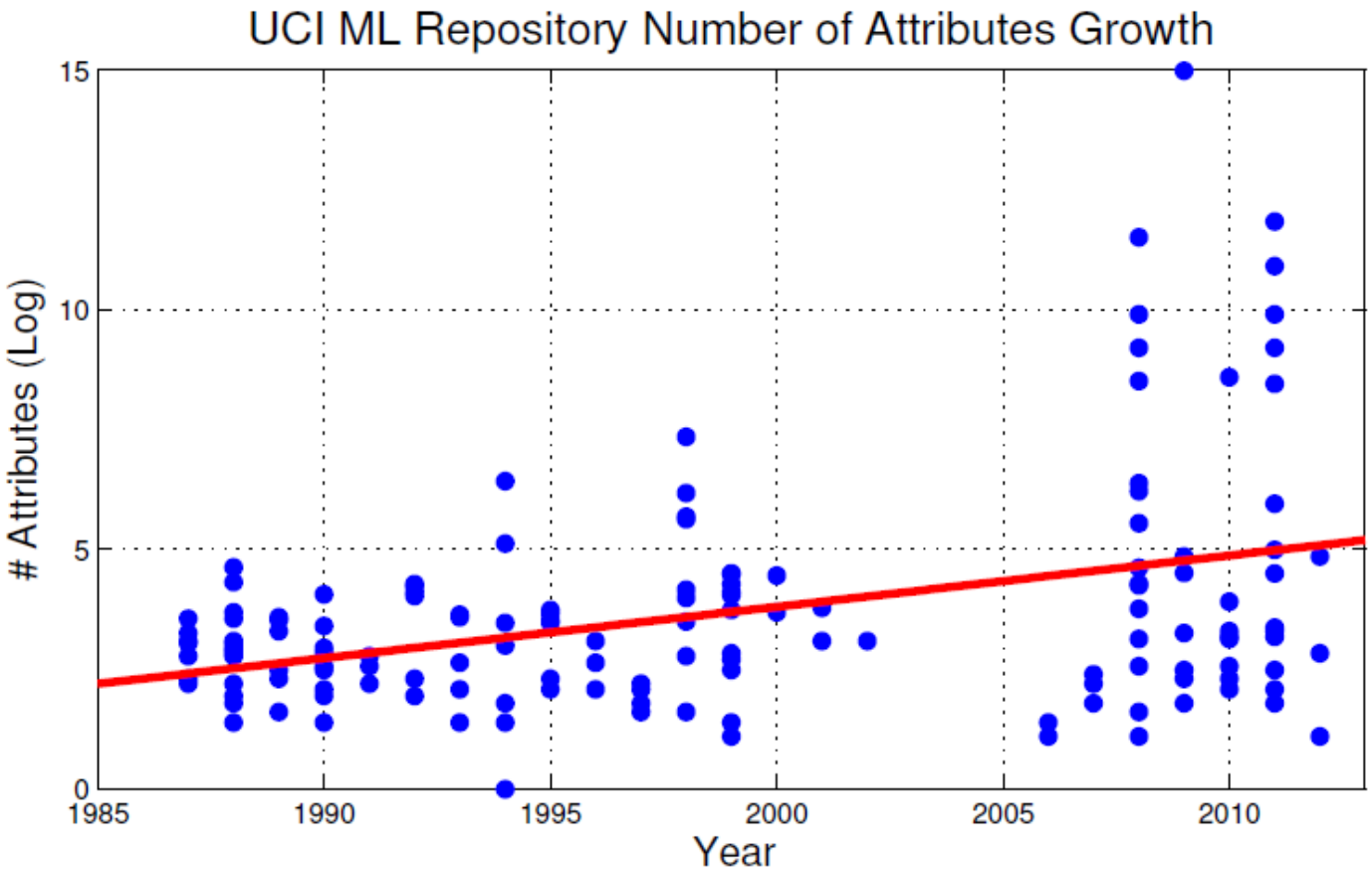


[1] S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, **29**

An opportunity for detecting weak signal



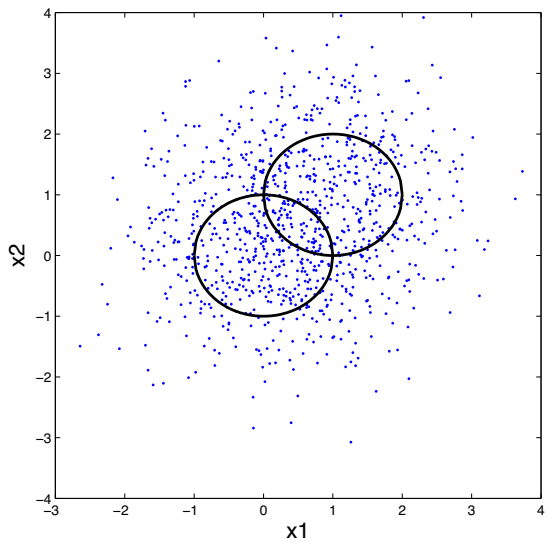
High-dimensional data $(d)^2$

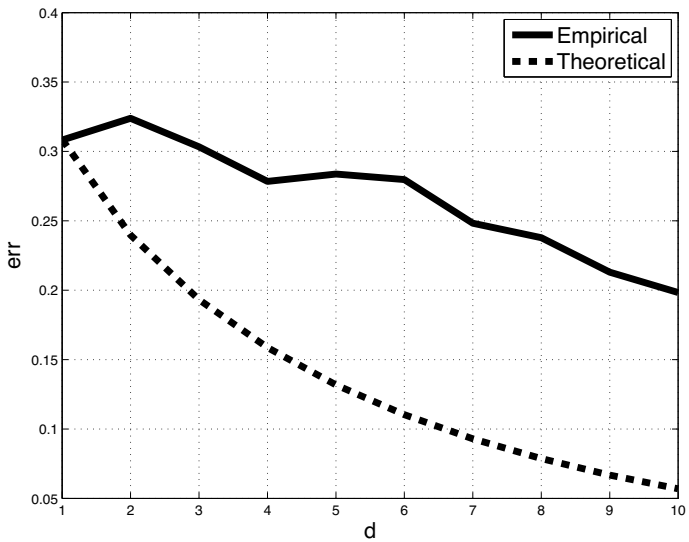


[2] S. Alelyani, J. Tang and H. Liu (2013). Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, **29**

An opportunity for amplifying the signal

Each variable provides equal and own separation information





Genesis of "Big Data"

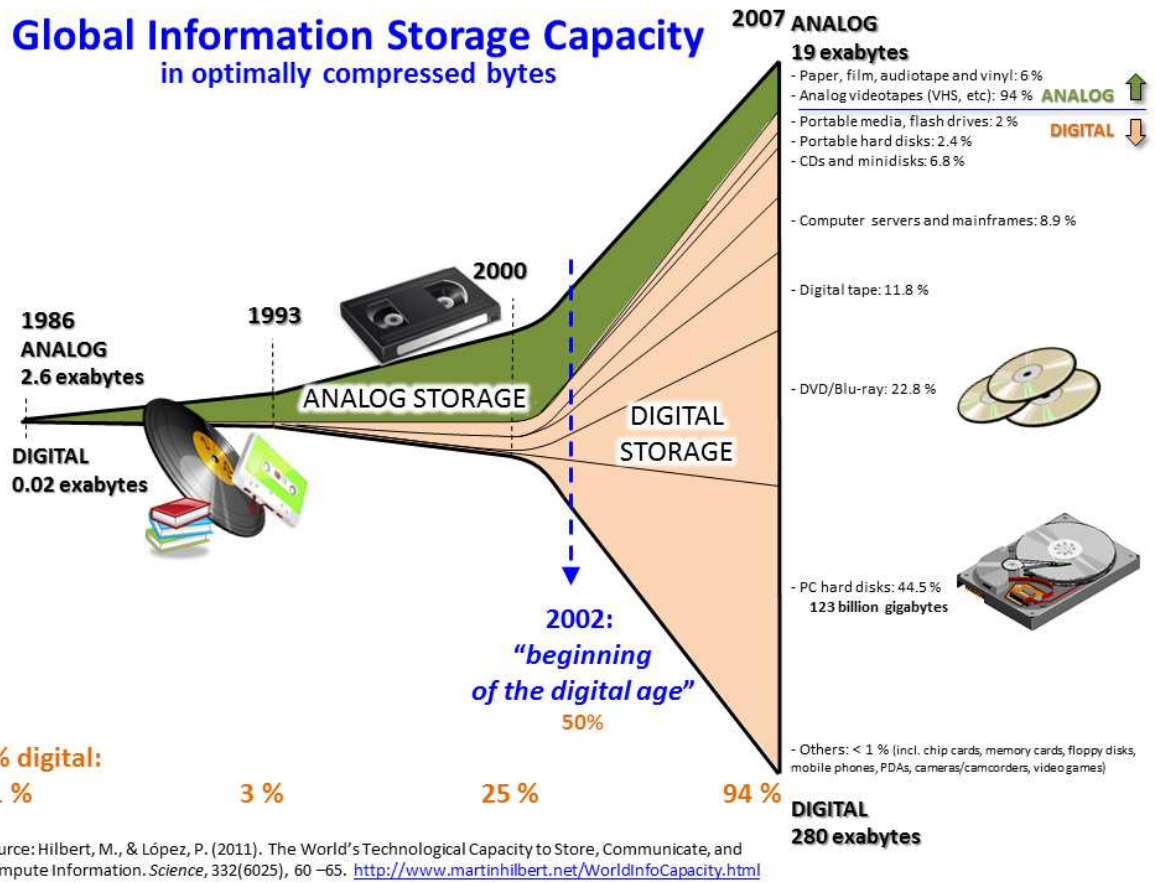
The Big Data phenomenon mainly originates in the increase of computer and digital resources at an ever lower cost

- Storage cost per MB: 700$ in 1981, 1$ in 1994, 0.01$ in 2013
  $\rightarrow$ price divided by 70,000 in thirty years
- Storage capacity of HDDs: $\approx$1.02 Go in 1982, $\approx$8 To today
  $\rightarrow$ capacity multiplied by 8,000 over the same period
- Computeur processing speed: 1 gigaFLOPS[3] in 1985, 33 petaFLOPS in 2013
  $\rightarrow$ speed multiplied by 33 million

[3] FLOP = FLoating-point Operations Per Second

Digital flow

- Digital in 1986: 1% of the stored information, 0.02 Eo[4]
- Digital in 2007: 94% of the stored information, 280 Eo (multiplied by 14,000)



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

[4] Exabyte

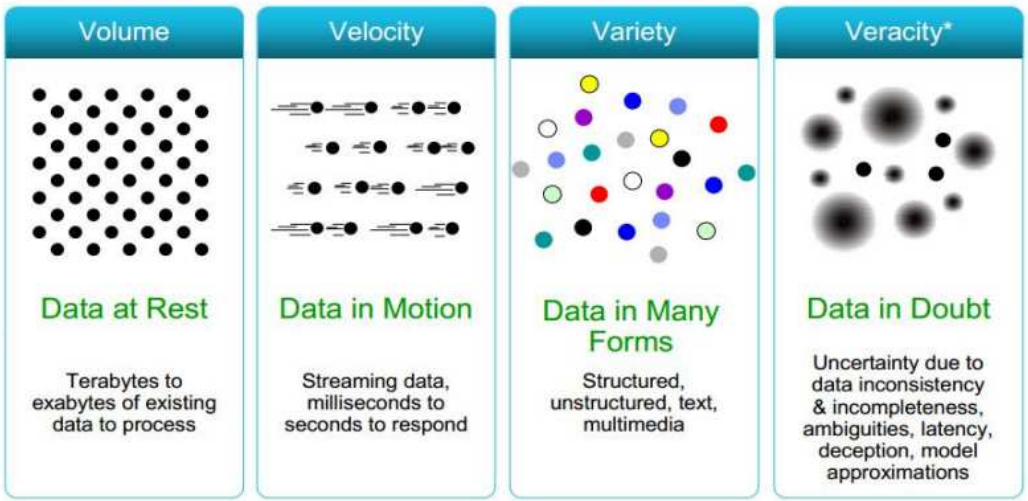
Societal phenomenon

All human activities are impacted by data accumulation

- Trade and business: corporate reporting system , banks, commercial transactions, reservation systems. . .
- Governments and organizations: laws, regulations, standardizations , infrastructure. . .
- Entertainment: music, video, games, social networks. . .
- Sciences: astronomy, physics and energy, genome,. . .
- Health: medical record databases in the social security system. . .
- Environment: climate, sustainable development , pollution, power. . .
- Humanities and Social Sciences: digitization of knowledge , literature, history , art, architecture, archaeological data. . .

Three challenges

- The storage challenge:
  - Storage, transfer, preservation, availability
  - Ex.: in Astrophysics, 50GB acquisition per day for the Euclid project (2021)
- The data analysis challenge: 3V, 4V, 5V...



- The societal and economic challenge:
  - Protection of private life, right to be forgotten, property rights, operating rights, cost of energy storage or transfer
  - Ex.: the PRISM project of the NSA

Coding for data $\boldsymbol{x}$

- A set of n individuals

$$\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$$

with $\boldsymbol{x}_i$ a set of (possibly non-scalar) d variables

$$\boldsymbol{x}_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{id}\}$$

where $\boldsymbol{x}_{ij} \in \mathcal{X}_j$

- A n-uplet of individuals

$$\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

with $\boldsymbol{x}_i$ a d-uplet of (possibly non-scalar) variables

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{id}) \in \mathcal{X}$$

where $\mathcal{X} = \mathcal{X}_1 \times \ldots\ldots \mathcal{X}_d$

We will pass from a coding to another, depending of the practical utility (useful for some calculus to have matrices or vectors for instance)

Outline

More data. . . implies more missing data!

"Classical" missing data

Today, it is easy to collect many features, so it favors

- data variety and/or mixed
- data missing
- data uncertainty (or interval data)

Mixed, missing, uncertain

Observed individuals $\boldsymbol{x}^o$			
?	0.5	?	5
0.3	0.1	green	3
0.3	0.6	{red,green}	3
0.9	[0.25 0.45]	red	?
↓	↓	↓	↓
continuous	continuous	categorical	integer

Missing data: notations

- $\boldsymbol{c} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$: pattern of missing data for the full dataset
- $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{id}) \in \{0,1\}^d$: pattern of missing data for individual $i \in \{1, \ldots, n\}$

$$c_{ij} = 1 \Leftrightarrow x_{ij} \text{ is missing}$$

- $\boldsymbol{x}^o$: the observed values in $\boldsymbol{x}$
- $\boldsymbol{x}^m$: the missing values in $\boldsymbol{x}$

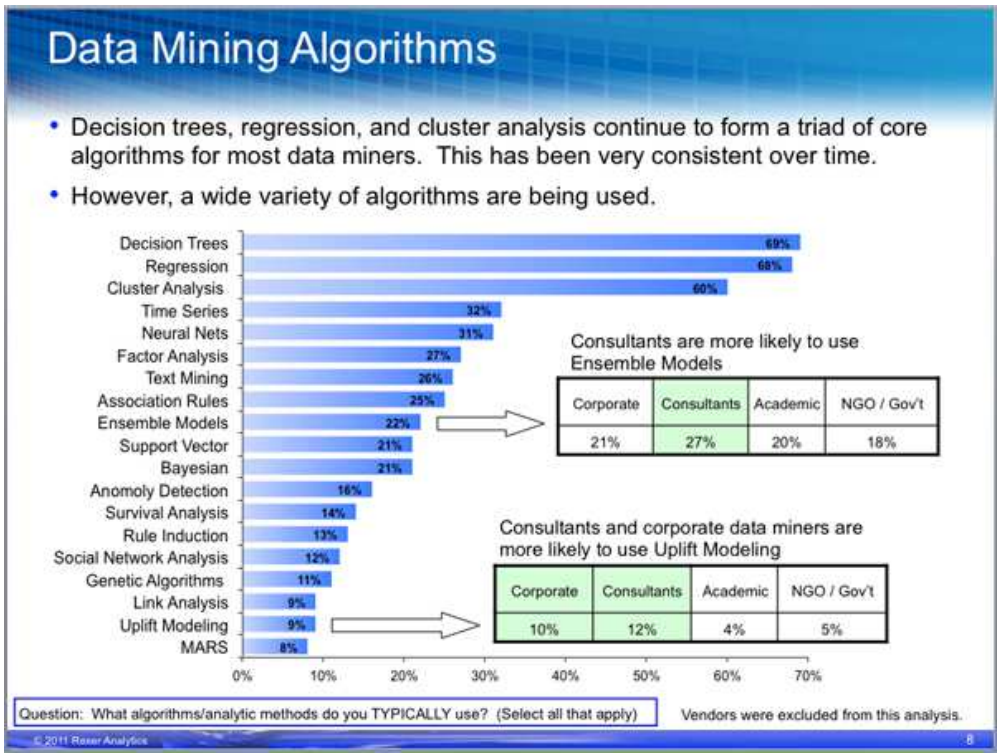
$$\boldsymbol{x} = \{\boldsymbol{x}^o, \boldsymbol{x}^m\} \text{ is the full dataset with its observed and missing parts}$$

Notation illustration

$$\boldsymbol{x}^o = \begin{pmatrix} ? & 2.6 & 5 \\ \text{blue} & 1.9 & 4 \\ \text{red} & 2.3 & ? \end{pmatrix}, \quad \boldsymbol{c} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

More advanced missing data: latent class in clustering (1/3)

Clustering everywhere[5]...



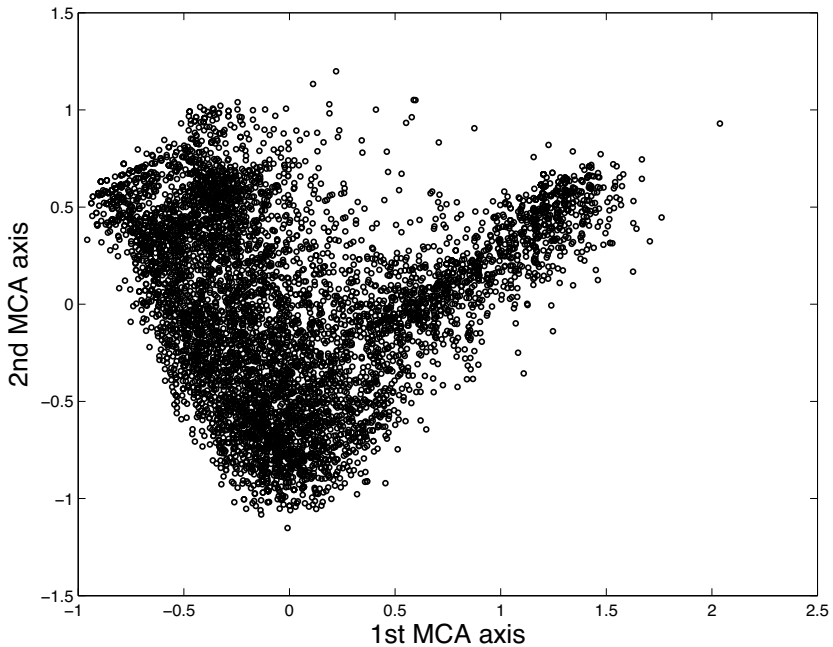
See Part II

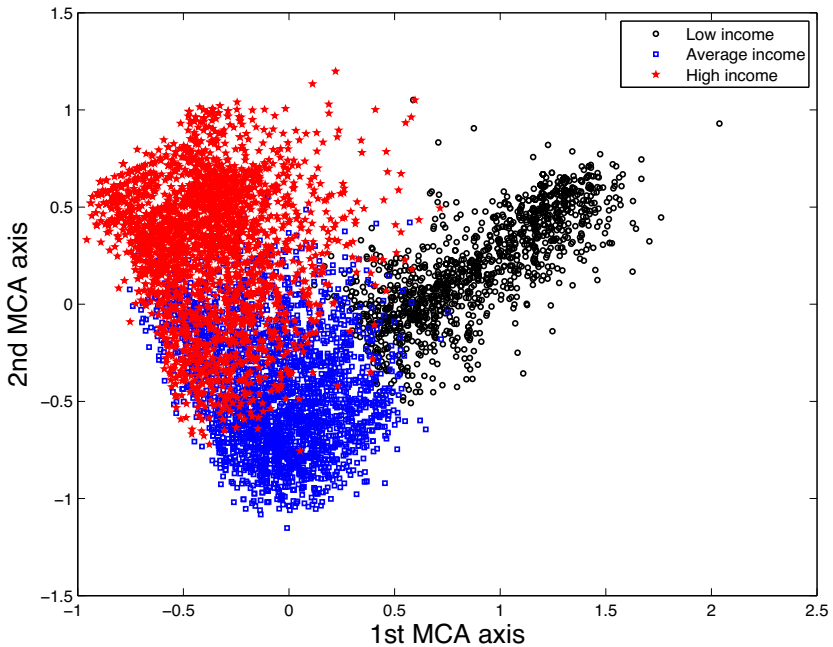
[5] Rexer Analytics's Annual Data Miner Survey is the largest survey of data mining, data science, and analytics professionals in the industry (survey of 2011)

More advanced missing data: latent class in clustering (2/3)

Clustering

Detect hidden structures in data sets: opportunity to reveal new information...





More advanced missing data: latent class in clustering (3/3)

- Aim: estimation of the partition z and the number of clusters K
- Notation : partition in K clusters $G_1, \ldots, G_K$: $z = (z_1, \ldots, z_n)$, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$

$$\boldsymbol{x}_i \in G_k \quad \Leftrightarrow \quad z_{ih} = \mathbb{I}_{\{h=k\}}$$

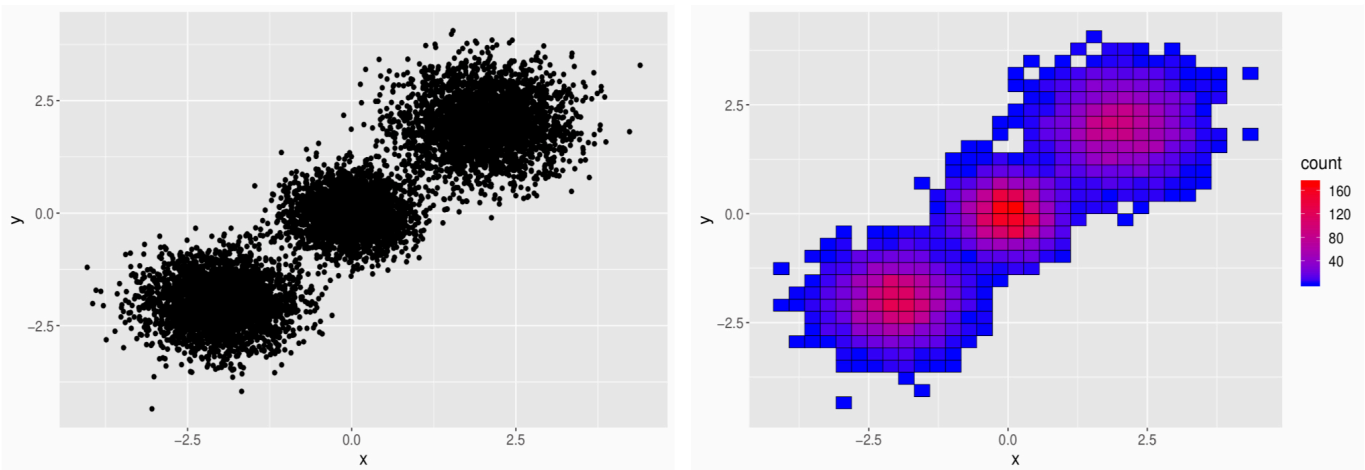
Mixed, missing, uncertain

Individuals $\boldsymbol{x}^o$				Partition z			$\Leftrightarrow$	Clusters
?	0.5	red	5	?	?	?	$\Leftrightarrow$	???
0.3	0.1	green	3	?	?	?	$\Leftrightarrow$	???
0.3	0.6	{red,green}	3	?	?	?	$\Leftrightarrow$	???
0.9	[0.25 0.45]	red	?	?	?	?	$\Leftrightarrow$	???
↓	↓	↓	↓					
continuous	continuous	categorical	integer					

All the partition z is unknown, thus it corresponds also to missing data...

Other more advanced missing data: binned data

- Binned data are equivalent to a multivariate histogram
- Data are partially missing since precise value is unknown inside each bin



(a) Raw data (b) Binned data

See Part IV

Classical strategies for dealing with missing data

Two traditional solutions (for obtaining a filled dataset)

- Discard individuals with missing data: more variance or a biased subset
- Impute missing data[6]: possible bias and underestimation of the variability

General guidelines

- Obtaining a complete dataset is not the final goal
- Missing data management should take into account the initial analysis goal

Purpose of Part I

Embed missing data management into the final goal paradigm...

[6] See others lessons for classical methods of imputation

Illustration of the risk of imputation

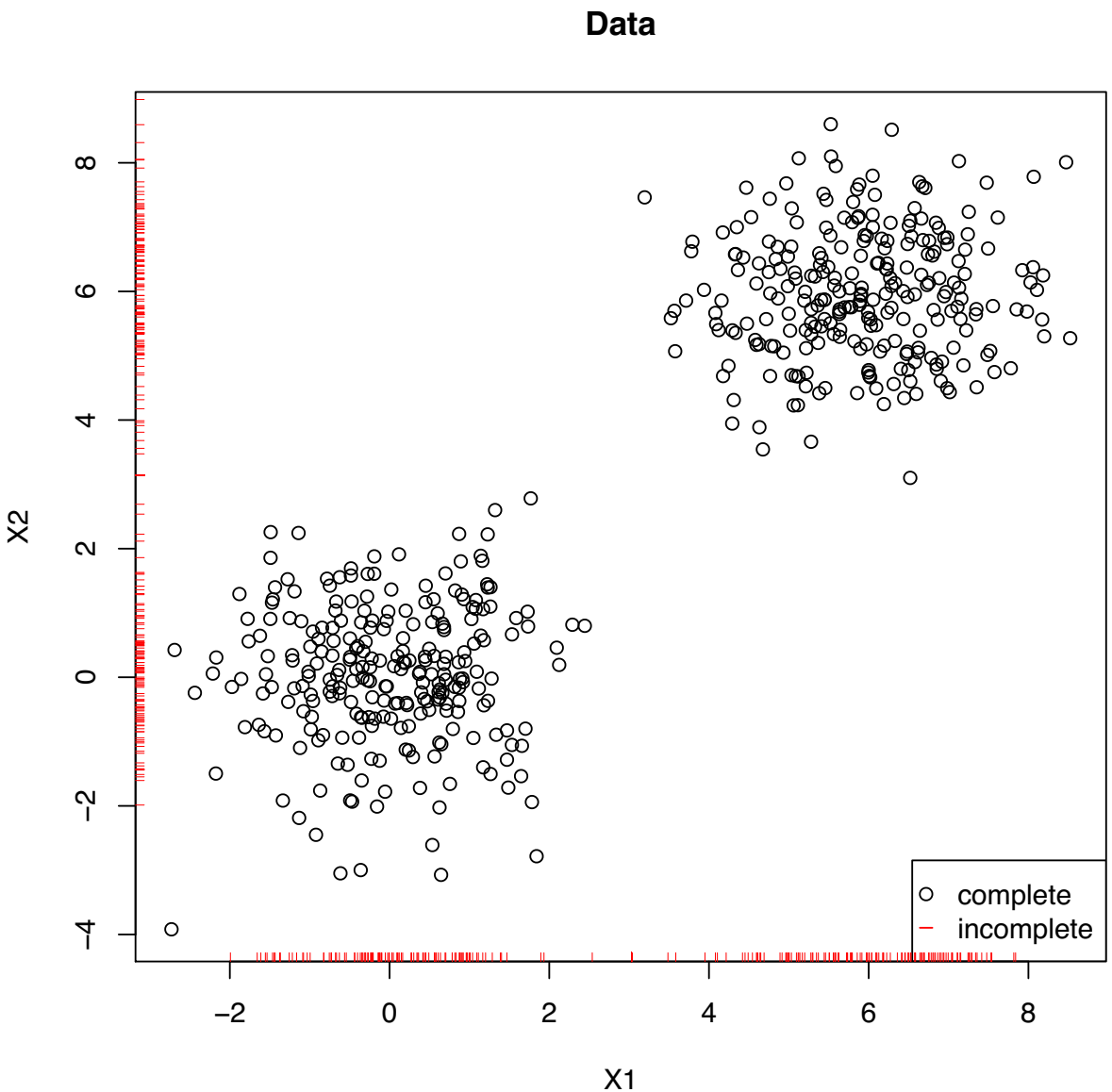


Illustration of the risk of imputation

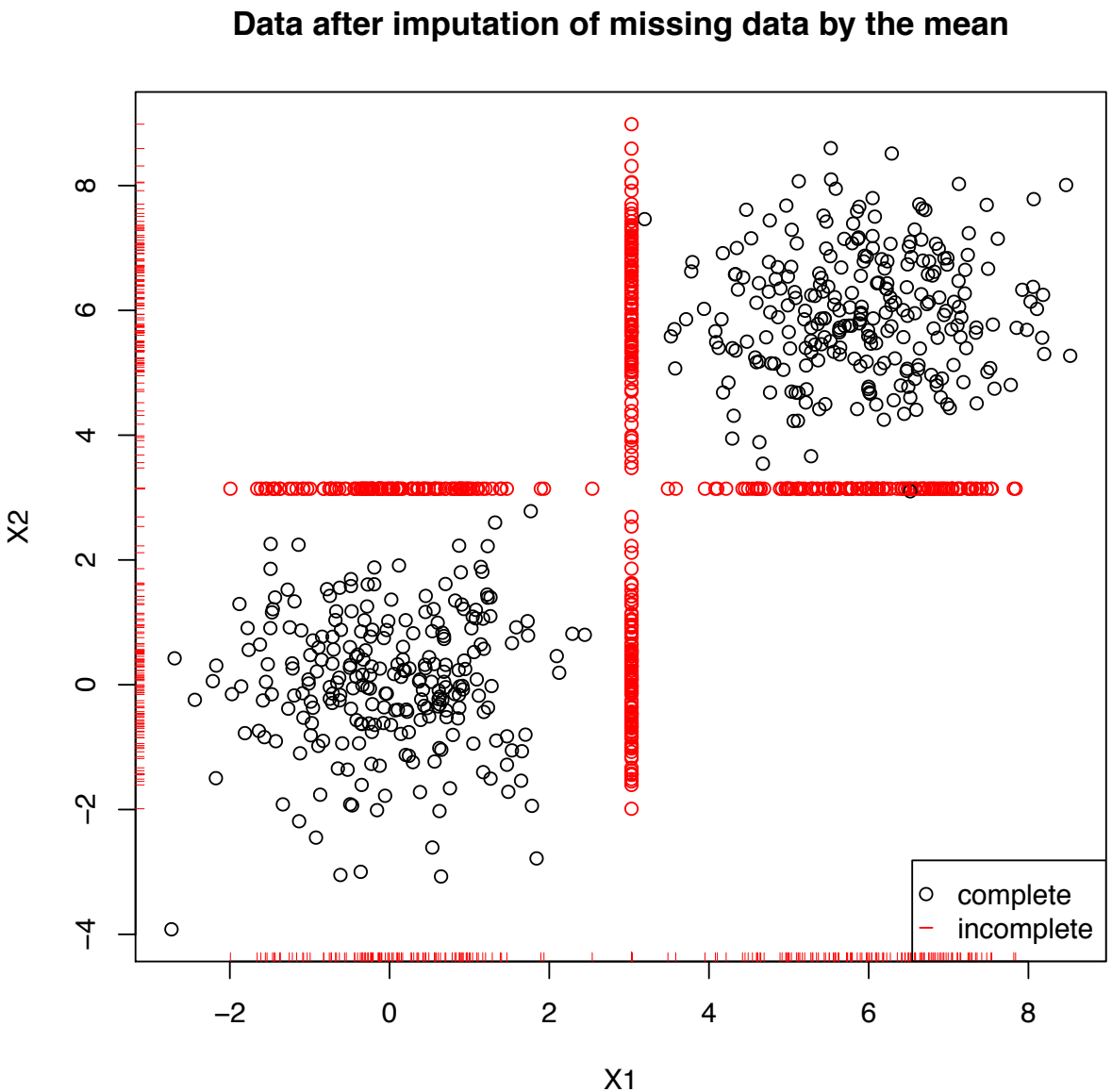
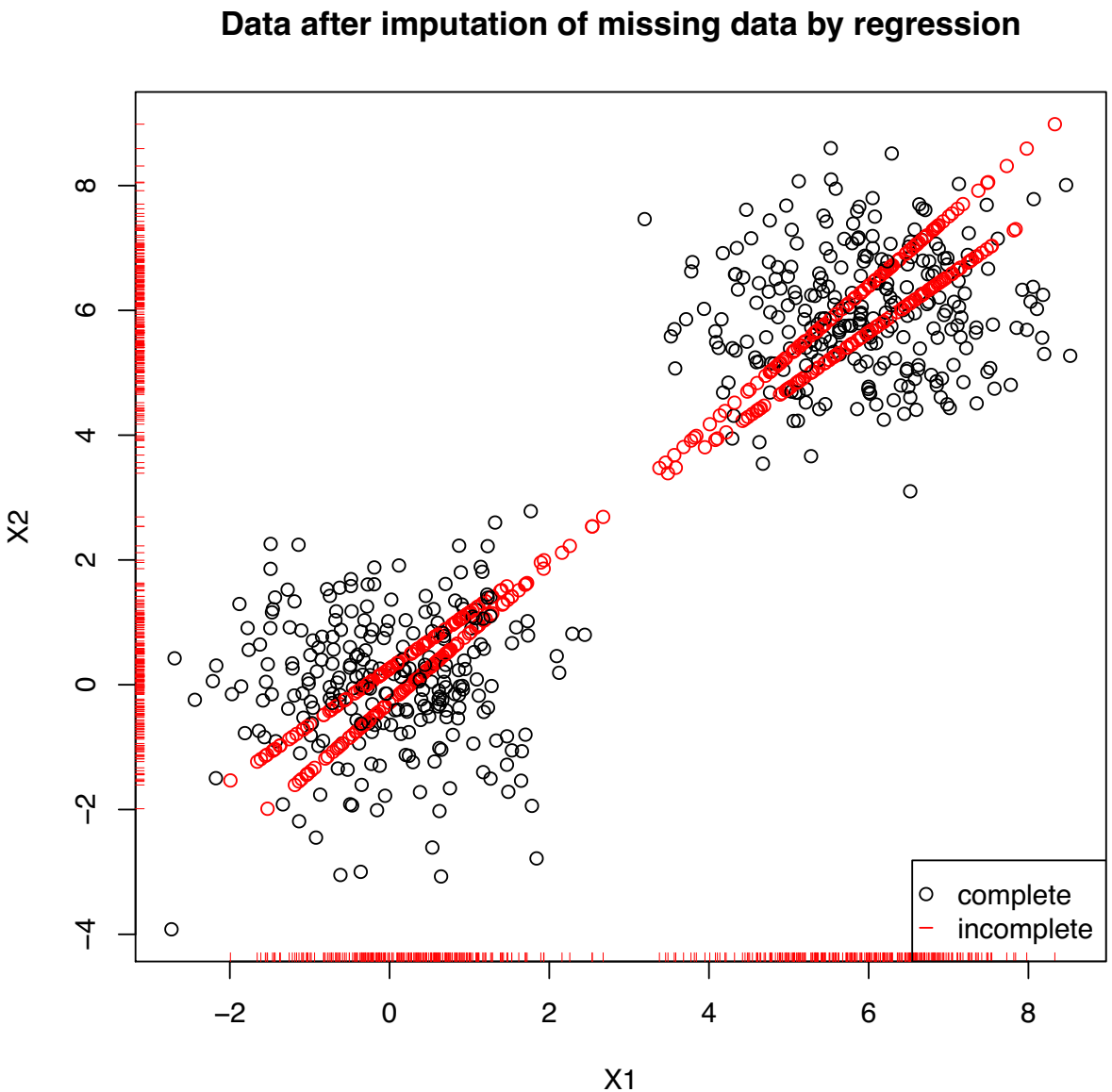


Illustration of the risk of imputation



Outline

Prostate cancer data[7]

- Individuals: 506 patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease
- Variables: $d = 12$ pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histolic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases)
- Some missing data: 62 missing values ($\approx 1\%$)

[7] Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488

Descriptors of the prostate cancer data set

Covariate	*Abbreviation*	*Number of Levels (if categorical)*
Age	Age	
Weight	Wt	
Performance rating	PF	4
Cardiovascular disease history	HX	2
Systolic Blood pressure	SBP	
Diastolic blood pressure	DBP	
Electrocardiogram code	EKG	7
Serum haemoglobin	HG	
Size of primary tumour	SZ	
Index of tumour stage and histolic grade	SG	
Serum prostatic acid phosphatase	AP	
Bone metastases	BM	2

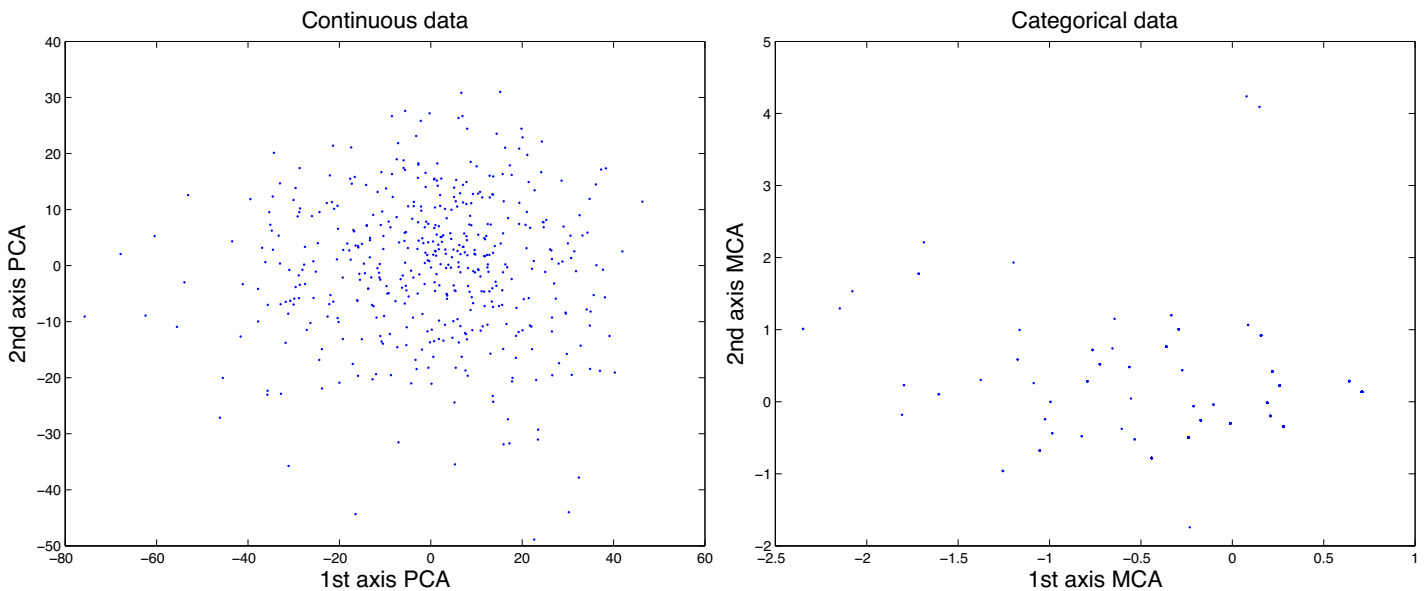
Aim

We forget the classes (Stages of the desease) for performing clustering

Questions

- How many clusters?
- Which partition?

Visually not so easy...



Two strategies in competition

- **Strategy "mice[8] + MixtComp[9]"**: MixtComp on the dataset completed by `mice`

```
> data.imp=mice(data)
> data.comp.mice=complete(data.imp)
```

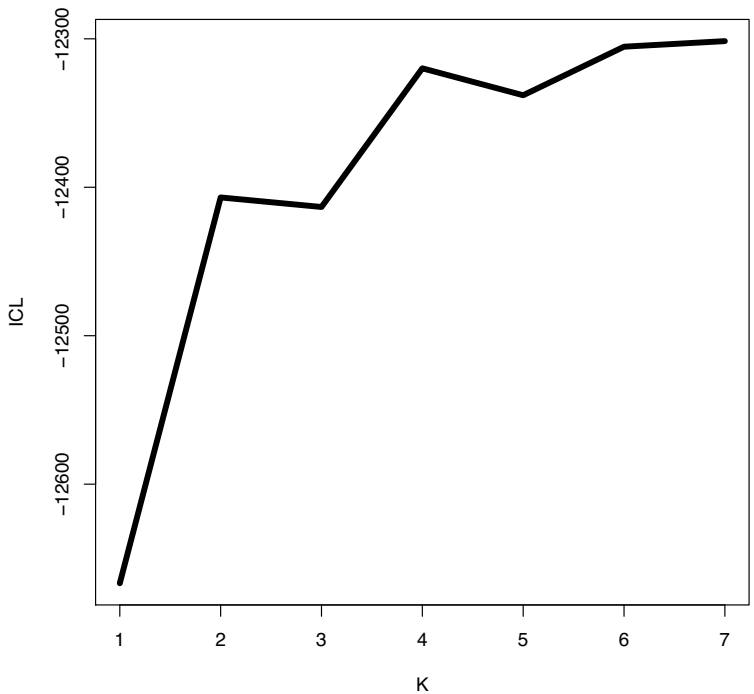
- **Strategy "full MixtComp"**: MixtComp on the observed (no completed) dataset

More information about Mixtcomp in Part II (be patient)...

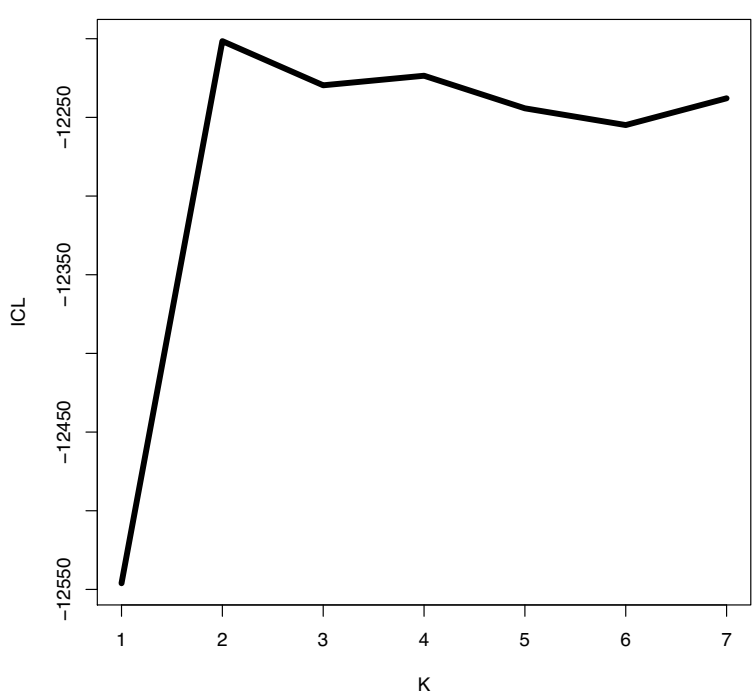
[8] See other lessons: http://cran.r-project.org/web/packages/mice/mice.pdf

[9] See Part II: https://cran.r-project.org/web/packages/RMixtComp/index.html

Choosing K with the ICL criterion[10]



mice + MixtComp
$\hat{K} = 7$



full MixtComp
$\hat{K} = 2$

. . . may lose some cluster information when imputation before clustering

[10] See Part II

Partition quality with $K = 2$

Strategy	mice + MixtComp	full MixtComp
% misclassified	12.8	8.1

To be compared also to missing data removal:

- 475 patients with non-missing data
- MixtComp for clustering
- possibility to consider continuous, categorical or mixed data

Strategy	continuous only	categorical only	mixed cont/cat
% misclassified	9.46	47.16	8.63

- risk of information lost when removing missing data lines/columns
- avoid to complete missing data (imputation depends on the purpose)

Outline

Full modeling relies on full observed data $(\boldsymbol{x}^o, \boldsymbol{c})$

From "partial" to "full" modeling

Partial modeling

- In statistics, most methods rely on
  1. modeling $\mathrm{p}(\boldsymbol{x};\boldsymbol{\theta})$
  2. estimating $\boldsymbol{\theta}$ from $\boldsymbol{x}$
- Examples : density estimation, regression (just add covariates), clustering...
- However, it involves both observed and missing data since $\boldsymbol{x} = (\boldsymbol{x}^o, \boldsymbol{x}^m)$
- It is the reason why imputation/deletion of missing entries is popular!

Full modeling

- However, strictly speaking
  - complete data are $(\boldsymbol{x}, \boldsymbol{c})$
  - observed data are $(\boldsymbol{x}^o, \boldsymbol{c})$
- Consequently
  1. modeling should concern $\mathrm{p}(\boldsymbol{x}, \boldsymbol{c};\boldsymbol{\theta},\psi) = \mathrm{p}(\boldsymbol{c}|\boldsymbol{x};\psi)\mathrm{p}(\boldsymbol{x};\boldsymbol{\theta})$
  2. estimating $\boldsymbol{\theta}$ should be performed only from $(\boldsymbol{x}^o, \boldsymbol{c})$

Missing data: typology of the missing mechanisms

- Missing completely at random (MCAR):

$$\mathrm{p}(\boldsymbol{c}|\boldsymbol{x};\psi) = \mathrm{p}(\boldsymbol{c};\psi) \quad \forall \boldsymbol{x}$$

- Missing at random (MAR):

$$\mathrm{p}(\boldsymbol{c}|\boldsymbol{x};\psi) = \mathrm{p}(\boldsymbol{c}|\boldsymbol{x}^o;\psi) \quad \forall \boldsymbol{x}^m$$

- Missing not at random (MNAR): the mechanism is not MCAR nor MAR

Example of MNAR data

The probability to have a missing value on income depends on the value of income (rich people less inclined to reveal their income).

Missing data: a seminal paper

Biometrika (1976), **63**, 3, *pp.* 581–92
Printed in Great Britain

Inference and missing data

BY DONALD B. RUBIN
Educational Testing Service, Princeton, New Jersey

SUMMARY

When making sampling distribution inferences about the parameter of the data, θ, it is appropriate to ignore the process that causes missing data if the missing data are 'missing at random' and the observed data are 'observed at random', but these inferences are generally conditional on the observed pattern of missing data. When making direct-likelihood or Bayesian inferences about θ, it is appropriate to ignore the process that causes missing data if the missing data are missing at random and the parameter of the missing data process is 'distinct' from θ. These conditions are the weakest general conditions under which ignoring the process that causes missing data always leads to correct inferences.

Some key words: Bayesian inference; Incomplete data; Likelihood inference; Missing at random; Missing data; Missing values; Observed at random; Sampling distribution inference.

Ignorable *vs.* non ignorable model

A missing mechanism is ignorable if likelihoods can be decomposed as

$$L(\boldsymbol{\theta}, \psi; \boldsymbol{x}^o, \boldsymbol{c}) = L(\psi; \boldsymbol{c}|\boldsymbol{x}^o) \times L(\boldsymbol{\theta}; \boldsymbol{x}^o)$$

Inference of $\boldsymbol{\theta}$

"If the missing mechanism is ignorable then likelihood-based inferences for $\boldsymbol{\theta}$ from $L(\boldsymbol{\theta}; \boldsymbol{x}^o)$ will be the same as likelihood based inference for $\boldsymbol{\theta}$ from $L(\boldsymbol{\theta}, \psi; \boldsymbol{x}^o, \boldsymbol{c})$." [a]

[a] Statistical Analysis With Missing Data. Roderick J. A. Little and Donald B. Rubin. New York: John Wiley & Sons, 1987, Section 6.2.

- M(C)AR is ignorable $\longrightarrow$ is the case in Part I
- MNAR is not ignorable $\longrightarrow$ will be the case in Part III

Likelihood-based inference concerns both frequentist and Bayesian paradigms

Outline

Observed-data log-likelihood estimation of $\boldsymbol{\theta}$

- Principle: MLE (Maximum Likelihood Estimate)

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}; \boldsymbol{x}^o)$$

where the observed-data log-likelihood is defined by

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}^o) = \ln \mathrm{p}(\boldsymbol{x}^o; \boldsymbol{\theta})$$

- Properties: we have, with $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}})$[11]

$$\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}^* \quad \text{and} \quad \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathrm{N}_\nu\left(\mathbf{0}, \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}\right)$$

where ν is the number of free continuous parameters in Θ, and

$$\mathbf{J} = -\mathrm{E}_{X_1} \nabla^2 \ln \mathrm{p}(X_1; \boldsymbol{\theta}^*)$$
$$\mathbf{K} = \mathrm{Var}_{X_1} \nabla \ln \mathrm{p}(X_1; \boldsymbol{\theta}^*)$$

- Algorithm: EM and variants (see later)

[11] p is the (unknown) true distribution, $\mathrm{p}_{\boldsymbol{\theta}} = \mathrm{p}(\cdot; \boldsymbol{\theta})$ and KL is the Kullback-Leibler divergence.

Expectation-Maximization (EM) algorithm

We need to defined first the complete-data log-likelihood as

$$\ell_c(\boldsymbol{\theta}; \boldsymbol{x}) = \mathrm{p}(\boldsymbol{x}; \boldsymbol{\theta})$$

- Initialisation: $\boldsymbol{\theta}^{(0)}$
- Iteration (q):
  - E-step: expectation of the complete-data log-likelihood

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \mathrm{E}\left[\ell_c(\boldsymbol{\theta}; \boldsymbol{x}) | \boldsymbol{x}^o; \boldsymbol{\theta}^{(q)}\right]$$

  - M-step: maximization of Q over $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$$

- Stopping rule: iteration number Q or criterion stationarity

Properties

- $\oplus$: simplicity, monotony of ℓ, low memory request
- $\ominus$: local maxima (depends on $\boldsymbol{\theta}^{(0)}$), linear convergence

Tutorial and practical work about EM (1/5)

- Theoretical part (1/2): express the MLE of μ in the case where
  - $\boldsymbol{x} = (x_1, \ldots, x_n) \sim \mathcal{N}(\mu, 1)$ with $\mu = 1$ to be estimated (variance is known)
  - $\boldsymbol{x}^o = (x_1, \ldots, x_{n^o})$ is the known sample of size n^o
  - $x^m = (x_{n^o+1}, \ldots, x_{n^o+n^m})$ is the known sample of size n^m
- Theoretical part (2/2): express EM steps in the previous case
- Practical part: write an R script
  - Implementing the previous EM algorithm
  - Check on the same figure both the evolution of $\mu^{(q)}$ values over the iteration number and the MLE value $\hat{\mu}$
  - Check on another figure the evolution of $\ell(\mu^{(q)}; \boldsymbol{x}^o)$ over q

Tutorial and practical work about EM (2/5)

- MLE:

$$\ell(\hat{\mu}; \boldsymbol{x}^o) = -\frac{n^o}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n^o}(x_i - \hat{\mu})^2$$

$$\frac{\partial \ell(\hat{\mu}^-; \boldsymbol{x}^o)}{\partial \hat{\mu}^-} = 0 \quad \Leftrightarrow \quad \hat{\mu} = \bar{\boldsymbol{x}}^o$$

Remark

Since MLE has a closed-form, the EM algorithm is dummy!

Tutorial and practical work about EM (3/5)

- E-step of EM:

$$\begin{aligned} Q(\mu^+, \mu^-) &= \mathrm{E}\left[-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu^+)^2 | \boldsymbol{x}^o; \mu^-\right] \\ &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n^o}(x_i - \mu^+)^2 - \frac{1}{2}\mathrm{E}\left[\sum_{i=n^o+1}^{n}(x_i - \mu^+)^2 | \boldsymbol{x}^o; \mu^-\right] \end{aligned}$$

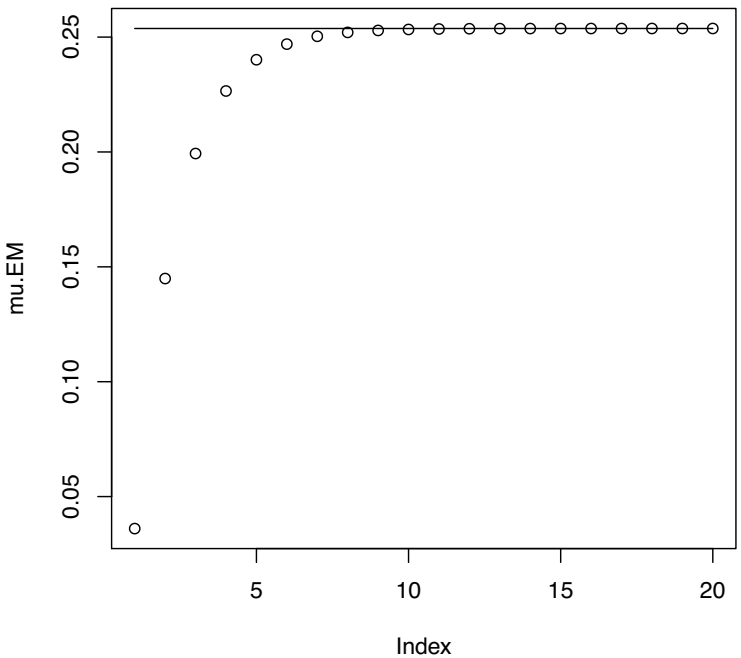
- M-step of EM:

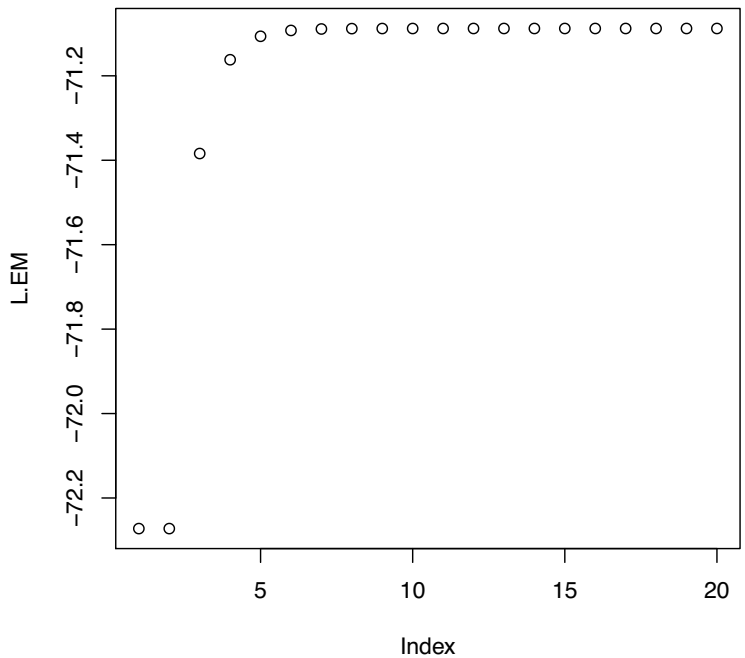
$$\begin{aligned} \frac{\partial Q(\mu^+, \mu^-)}{\partial \mu^+} = 0 \quad &\Leftrightarrow \quad -\frac{1}{2}\mathrm{E}\left[\partial\frac{\sum_{i=1}^{n}(x_i - \mu^+)^2}{\partial \mu^+} | \boldsymbol{x}^o; \mu^-\right] = 0 \\ &\Leftrightarrow \quad \sum_{i=1}^{n^o}(x_i - \mu^+) + n^m \mathrm{E}\left[(x - \mu^+); \mu^-\right] = 0 \\ &\Leftrightarrow \quad \mu^+ = \frac{n^o \bar{\boldsymbol{x}}^o + n^m \mu^-}{n} \end{aligned}$$

Tutorial and practical work about EM (4/5)

```
# data drawing
n.obs = 50
x.obs = rnorm(n.obs)
n.mis = 50

# closed-form ML
mu.ML = mean(x.obs)

# EM
n.it = 20
mu.EM = vector(mode = "numeric", length = n.it)
mu.EM[1] = rnorm(1) # starting value
L.EM = vector(mode = "numeric", length = n.it)
L.EM[1] = sum(dnorm(x.obs, mean = mu.EM[1], sd = 1, log = TRUE))
for (i.it in 2:n.it){
    mu.EM[i.it] = (sum(x.obs) + n.mis*mu.EM[i.it-1]) / (n.obs+n.mis)
    L.EM[i.it] = sum(dnorm(x.obs, mean = mu.EM[i.it-1], sd = 1, log = TRUE))
}
plot(mu.EM)
lines(1:n.it,rep(mu.ML,n.it))

# draw L
plot(L.EM)
```

Tutorial and practical work about EM (5/5)





Stochastic EM (SEM) algorithm

The purpose is still to maximize ℓ over $\boldsymbol{\theta}$

- Initialisation: $\boldsymbol{\theta}^{(0)}$
- Iteration (q):
  - SE-step: draw missing values
  $$\boldsymbol{x}^{m(q)} \sim \mathrm{p}(\cdot|\boldsymbol{x}^o; \boldsymbol{\theta}^{(q)})$$
  - M-step: maximize ℓ_c
  $$\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_c(\boldsymbol{x}^o, \boldsymbol{x}^{m(q)}; \boldsymbol{\theta})$$
- Stopping rule: iteration number Q

Properties

- The mean of the sequence $(\boldsymbol{\theta}^{(q)})$ approximates $\hat{\boldsymbol{\theta}}$:
$$\hat{\boldsymbol{\theta}} \simeq \frac{1}{Q - Q^{\text{burn-in}} + 1} \sum_{q=Q^{\text{burn-in}}}^{Q} \boldsymbol{\theta}^{(q)},$$
where $Q^{\text{burn-in}}$ is a so-called burn-in period to reach a stationary regime
- $\oplus$: often easier to express than EM, avoids local maxima, the standard deviation of the sequence $(\boldsymbol{\theta}^{(q)})$ produces a confidence interval
- $\ominus$: no punctual convergence (does not improve ℓ at each iteration), can be slower than EM

Tutorial and practical work about SEM (1/4)

Same exercise as EM but now with SEM.

Tutorial and practical work about SEM (2/4)

- SE-step:

$$\begin{aligned} \boldsymbol{x}^{m+} &\sim \mathrm{p}(\cdot|\boldsymbol{x}^o;\boldsymbol{\theta}^-) \\ &= \mathrm{p}(\dot{|};\boldsymbol{\theta}^-) \\ &= \mathcal{N}_{n^m}(\mu^-\mathbf{1},\mathbf{I}) \end{aligned}$$

- M-step: we have just to maximise the complete-data log-likeliood

$$\ell_c(\mu^+;\boldsymbol{x}^o,\boldsymbol{x}^{m+}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n^o}(x_i^o-\hat{\mu}^+)^2 - \frac{1}{2}\sum_{i=1}^{n^m}(x_i^{m+}-\mu^+)^2$$

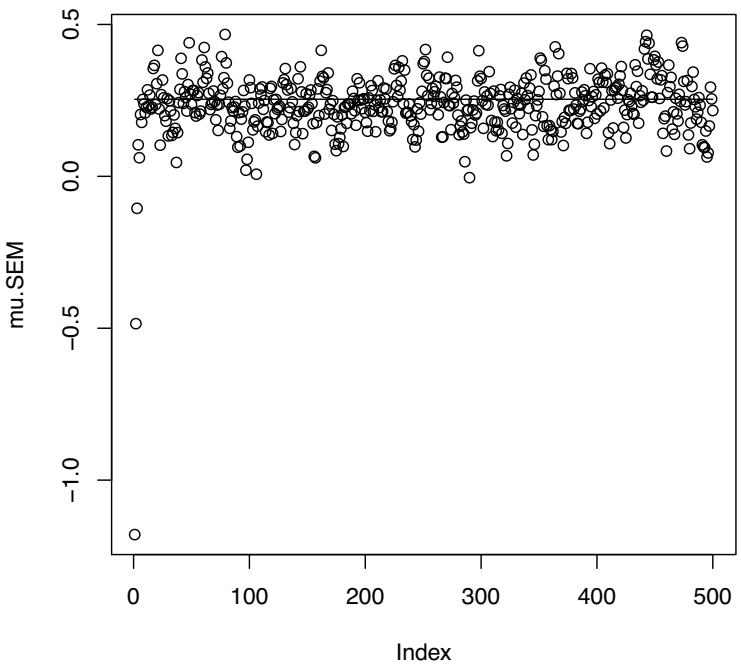
thus leading to the standard result

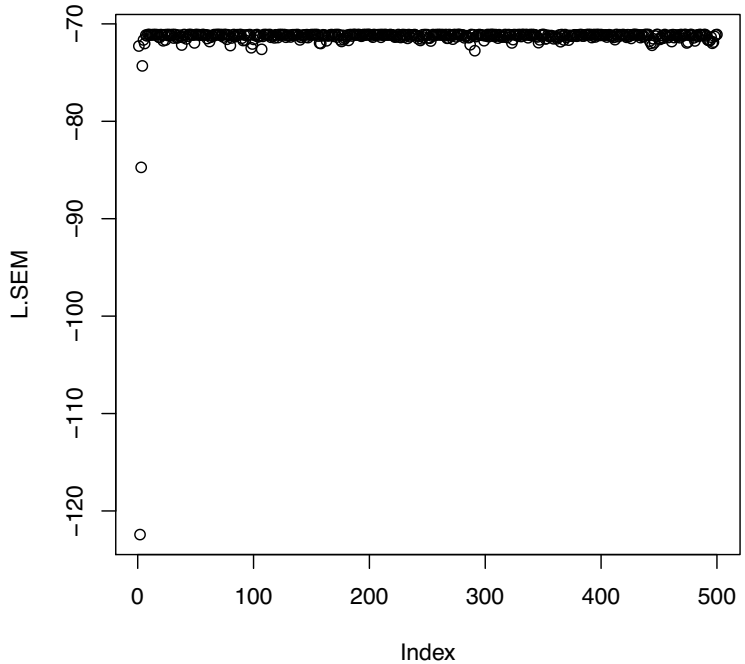
$$\mu^+ = \frac{n^o\bar{\boldsymbol{x}}^o + n^m\bar{\boldsymbol{x}}^{m+}}{n}.$$

Tutorial and practical work about SEM (3/4)

```
# data drawing
n.obs = 50
x.obs = rnorm(n.obs)
n.mis = 50

# closed-form ML
mu.ML = mean(x.obs)

# SEM
n.it = 500
mu.SEM = vector(mode = "numeric", length = n.it)
mu.SEM[1] = rnorm(1) # starting value
L.SEM = vector(mode = "numeric", length = n.it)
L.SEM[1] = sum(dnorm(x.obs, mean = mu.EM[1], sd = 1, log = TRUE))
for (i.it in 2:n.it){
    x.mis = rnorm(n.obs,mu.SEM[i.it-1],1)
    mu.SEM[i.it] = (sum(x.obs) + sum(x.mis)) / (n.obs+n.mis)
    L.SEM[i.it] = sum(dnorm(x.obs, mean = mu.SEM[i.it-1], sd = 1, log = TRUE))
}
plot(mu.SEM)
lines(1:n.it,rep(mu.ML,n.it))

# mu.SEM.final
n.burn = 10
mu.SEM.final = mean(mu.SEM[(n.burn+1):n.it])

# draw L
plot(L.SEM)
```

Tutorial and practical work about SEM (4/4)





Outline

Bayesian estimation of $\boldsymbol{\theta}$

- Preliminary remark: within the Bayesian paradigm, the parameter $\boldsymbol{\theta}$ is considered as a random vector thus it is classical to note $\mathrm{p}(\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})$
- Principle[12]:

$$\hat{\boldsymbol{\theta}}^{\text{Bayes}} = \mathrm{E}[\boldsymbol{\theta}|\boldsymbol{x}^o] = \int_\theta \boldsymbol{\theta}\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}^o)d\boldsymbol{\theta}$$

  where
  - $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}^o) = \ell(\boldsymbol{\theta};\boldsymbol{x}^o)\mathrm{p}(\boldsymbol{\theta})/\mathrm{p}(\boldsymbol{x}^o)$ is the posterior distribution of $\boldsymbol{\theta}$
  - $\mathrm{p}(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$
- Properties:
  - as for the MLE, asymptotic consistency and normality of $\hat{\boldsymbol{\theta}}^{\text{Bayes}}$
  - need to define a prior distribution...
  - posterior often not so easy to compute thus requires specific algorithms
- Algorithms: Gibbs, Metropolis-Hastings...

[12] Other definitions are possible.

Gibbs algorithm

The purpose is to generate a sequence $\boldsymbol{\theta}^{(Q^{\text{burn-in}})}, \ldots, \boldsymbol{\theta}^{(Q)}$ drawn from $\mathrm{p}(\boldsymbol{\theta}|\boldsymbol{x}^o)$

- Initialisation: $\boldsymbol{\theta}^{(0)}$
- Iteration (q):
  - Draw $\boldsymbol{x}^{m(q)}$:
  $$\boldsymbol{x}^{m(q)} \sim \mathrm{p}(\cdot|\boldsymbol{x}^o, \boldsymbol{\theta}^{(q)})$$
  - Draw $\theta^{m(q)}$
  $$\boldsymbol{\theta}^{(q+1)} \sim \mathrm{p}(\cdot|\boldsymbol{x}^o, \boldsymbol{x}^{m(q)})$$
- Stopping rule: iteration number Q

Properties

- The mean of the sequence $(\boldsymbol{\theta}^{(q)})$ approximates $\hat{\boldsymbol{\theta}}^{\text{Bayes}}$:

$$\hat{\boldsymbol{\theta}}^{\text{Bayes}} \simeq \frac{1}{Q - Q^{\text{burn-in}} + 1} \sum_{q=Q^{\text{burn-in}}}^{Q} \boldsymbol{\theta}^{(q)}$$

- Very similar to the SEM algorithm

Tutorial and practical work about Gibbs (1/4)

- Same exercise as EM and SEM but now with Gibbs.
- Take the prior $\mu \sim \mathcal{N}(0, 1)$.

Tutorial and practical work about Gibbs (2/4)

- Draw $\boldsymbol{x}^{m+}$: identical to the SE-step of SEM
- Draw μ^+:

$$\begin{aligned} \mu^+ &\sim \mathrm{p}(\cdot|x^o, \boldsymbol{x}^{m+}) \\ &= \mathcal{N}\left(\frac{n^o \bar{\boldsymbol{x}}^o + n^m \bar{\boldsymbol{x}}^{m+}}{n+1}, 1\right) \end{aligned}$$

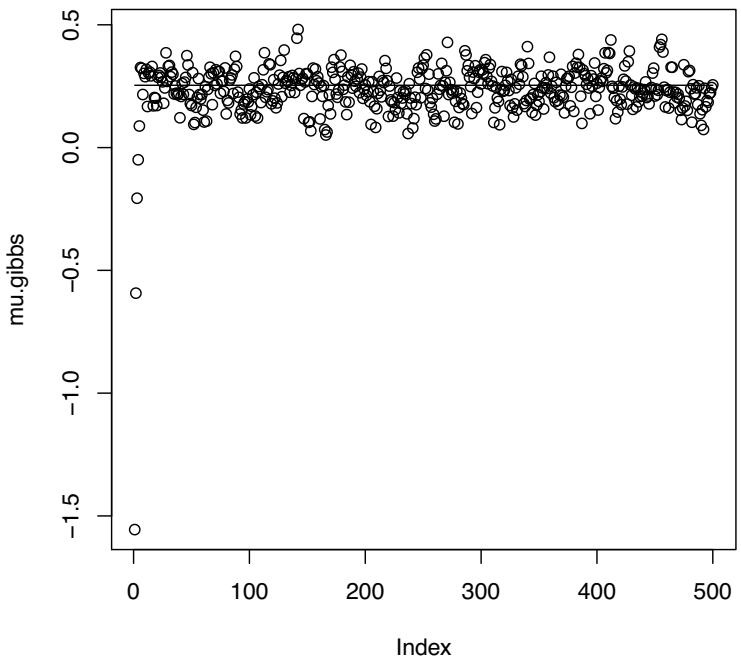
after using standard Gaussian results applied to the following expression

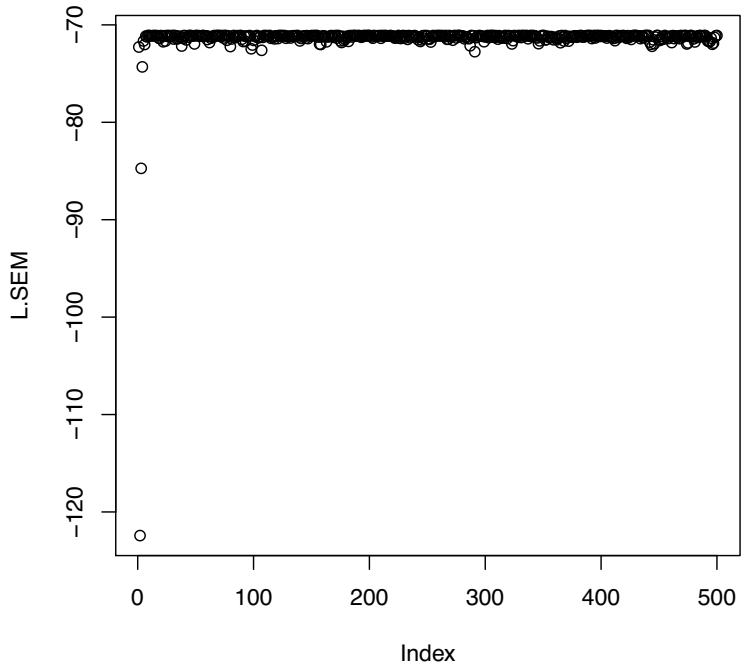
$$\mathrm{p}(\mu^+|x^o, \boldsymbol{x}^{m+}) = \frac{\mathrm{p}(x^o, \boldsymbol{x}^{m+}|\mu^+)\mathrm{p}(\mu^+)}{\mathrm{p}(x^o, \boldsymbol{x}^{m+})}$$

Tutorial and practical work about Gibbs (3/4)

```
# data drawing
n.obs = 50
x.obs = rnorm(n.obs)
n.mis = 50

# closed-form ML
mu.ML = mean(x.obs)

# Gibbs
n.it = 500
mu.gibbs = vector(mode = "numeric", length = n.it)
mu.gibbs[1] = rnorm(1) # starting value
for (i.it in 2:n.it){
    x.mis = rnorm(n.obs,mu.gibbs[i.it-1],1)
    mu.gibbs[i.it] = (sum(x.obs) + sum(x.mis)) / (n.obs+n.mis+1)
}
plot(mu.gibbs)
lines(1:n.it,rep(mu.ML,n.it))

# mu.gibbs.final
n.burn = 10
mu.gibbs.final = mean(mu.gibbs[(n.burn+1):n.it])

# draw L
plot(L.SEM)
```

Tutorial and practical work about Gibbs (4/4)





Outline

Model definition

- Model $\mathbf{m}$: it corresponds to a family of distributions $\mathrm{p}(\cdot;\boldsymbol{\theta})$

$$\mathbf{m} = \{\mathrm{p}(\boldsymbol{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

- Dimension ν: it corresponds to the number of free *continuous* parameters

$$\nu = \dim(\Theta)$$

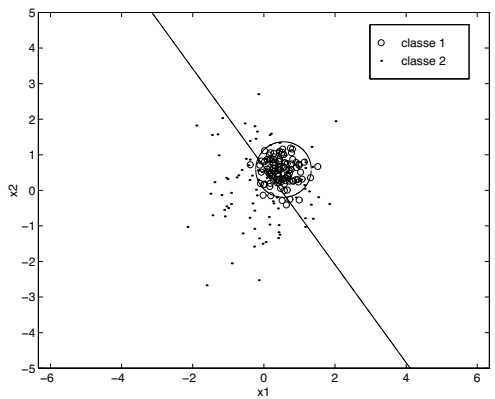
Interest of choosing a model

- A too simple model involves bias modeling
- A too complex model involves variance of estimation

Importance of model selection: example in supervised learning

Model **m** = parametric structure of the borderline shape between too groups

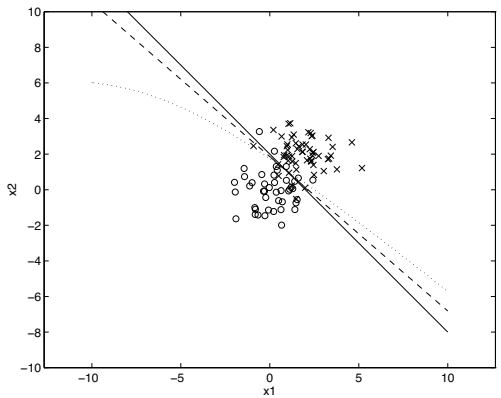
Too simple model: bias



true model: estimated spherical borderline

too simple model: estimated linear borderline

Too complex model: variance



——— true borderline

– – – estimated linear borderline

. . . estimated quadratic borderline

Integrated likelihood

- Posterior likelihood of $\mathbf{m}$:

$$\mathrm{p}(\mathbf{m}|\boldsymbol{x}^o) \propto \mathrm{p}(\boldsymbol{x}^o|\mathbf{m}) \quad \underbrace{\mathrm{p}(\mathbf{m})}_{\text{prior on } \mathbf{m}}$$

- Ideal model in a Bayesian context: with $\mathcal{M}$ a family of competing models

$$\hat{\mathbf{m}}^* \in \arg\max_{\mathbf{m}\in\mathcal{M}} \mathrm{p}(\mathbf{m}|\boldsymbol{x}^o)$$

- Integrated likelihood: if $\mathrm{p}(\mathbf{m}) = \text{cst}$, it is equivalent to maximize

$$\mathrm{p}(\boldsymbol{x}^o|\mathbf{m}) = \int_\Theta \mathrm{p}(\boldsymbol{x}^o;\boldsymbol{\theta},\mathbf{m}) \quad \underbrace{\mathrm{p}(\boldsymbol{\theta}|\mathbf{m})}_{\text{prior on } \boldsymbol{\theta}} \quad d\boldsymbol{\theta}$$

- Difficulties:
  - Choose the prior $\mathrm{p}(\boldsymbol{\theta}|\mathbf{m})$
  - Evaluate the integral

BIC criterion: genesis

- Laplace-Metropolis approximation: under standard regularity conditions, we have

$$\ln \mathrm{p}(\boldsymbol{x}^o|\mathbf{m}) = \ell(\hat{\boldsymbol{\theta}}_\mathbf{m}; \boldsymbol{x}^o) - \frac{\nu_\mathbf{m}}{2} \ln(n) + O_p(1)$$

  with $\hat{\boldsymbol{\theta}}_\mathbf{m}$ the MLE and $\nu_\mathbf{m}$ the number of parameters associated to $\mathbf{m}$
- BIC criterion (*Bayesian Information Criterion*): retain $\mathbf{m}$ maximizing

$$\mathrm{BIC}_\mathbf{m} = \ell(\hat{\boldsymbol{\theta}}_\mathbf{m}; \boldsymbol{x}^o) \; - \frac{\nu_\mathbf{m}}{2} \ln(n)$$

BIC criterion: consistency

- Consistency: BIC asymptotically selects

$$\mathbf{m}^* = \arg \inf_{\mathbf{m} \in \mathcal{M}} \mathrm{KL}(\mathrm{p}, \mathrm{p}_{\boldsymbol{\theta}^*_{\mathbf{m}}})$$

- Theoretical illustration of consistency: $\mathbf{m}_1 \subseteq \mathbf{m}_2$, $\mathbf{m}_1$ being the true model, $\Delta\nu = \nu_2 - \nu_1$, $\Delta\ell = \ell(\hat{\boldsymbol{\theta}}_2; \boldsymbol{x}^o) - \ell(\hat{\boldsymbol{\theta}}_1; \boldsymbol{x}^o)$, we have

$$2(\mathrm{BIC}_2 - \mathrm{BIC}_1) + \Delta\nu \ln(n) = 2\Delta\ell \xrightarrow{d} \chi^2_{\Delta\nu}$$

With $\mu = \Delta\nu$ and $\sigma^2 = 2\Delta\nu$ the mean and the variance of $\chi^2_{\Delta\nu}$

$$\mathrm{p}(\chi^2_{\Delta\nu} > \Delta\nu \ln(n)) \leq \mathrm{p}(|\chi^2_{\Delta\nu} - \mu| > \Delta\nu \ln(n) - \mu) \leq \frac{\sigma^2}{(\Delta\nu \ln(n) - \mu)^2} \overset{n\to\infty}{\longrightarrow} 0$$

by using the Chebyschev inequality. Thus, asymptotically, BIC will select $\mathbf{m}_1$

Tutorial and practical work about BIC (1/4)

- Let the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ where $x_1, \ldots, x_n \overset{iid}{\sim} \mathcal{N}(0, 1)$
- Let the uncertain sample $\boldsymbol{x}^\delta = ([x_1 - \delta, x_1 + \delta], \ldots, [x_n - \delta, x_n + \delta])$
- Let two competing models: $\mathbf{m}_1 = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$ and $\mathbf{m}_2 = \mathcal{N}(1, 1)$

- Compute BIC with $\mathbf{m}_1$ and with $\mathbf{m}_2$ associated to the sample $\boldsymbol{x}^\delta$
- Vary δ and see its effect on the model selection by BIC. Conclusion?

Tutorial and practical work about BIC (1/4)

Denoting by $\Phi(\cdot;\mu)$ the cumulative distribution of $\mathcal{N}(\mu,1)$, we have

$$\begin{aligned} \text{BIC}^{\delta} &= \ell(\hat{\mu}^{\text{MLE}};\boldsymbol{x}^{\delta}) - \frac{\nu}{2}\ln n \\ &= \sum_{i=1}^{n} \ln\left(\Phi(x_i+\delta;\hat{\mu}^{\text{MLE}}) - \Phi(x_i-\delta;\hat{\mu}^{\text{MLE}})\right) - \frac{\nu}{2}\ln n \end{aligned}$$

Then, use

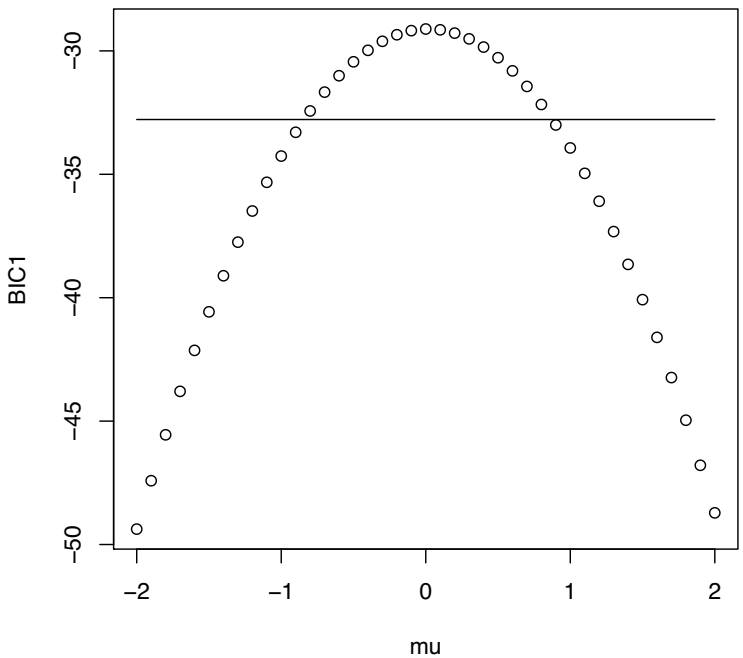
- $\nu = 1$ and $\hat{\mu} = \hat{\mu}_1^{\text{MLE}}$ for $\mathbf{m}_1$
- $\nu = 0$ and $\mu = 1$ for $\mathbf{m}_2$

Since the MLE $\hat{\mu}_1^{\text{MLE}}$ is not so easy to obtain, just graphically display the BIC^{δ} value for different values of μ on the same figure

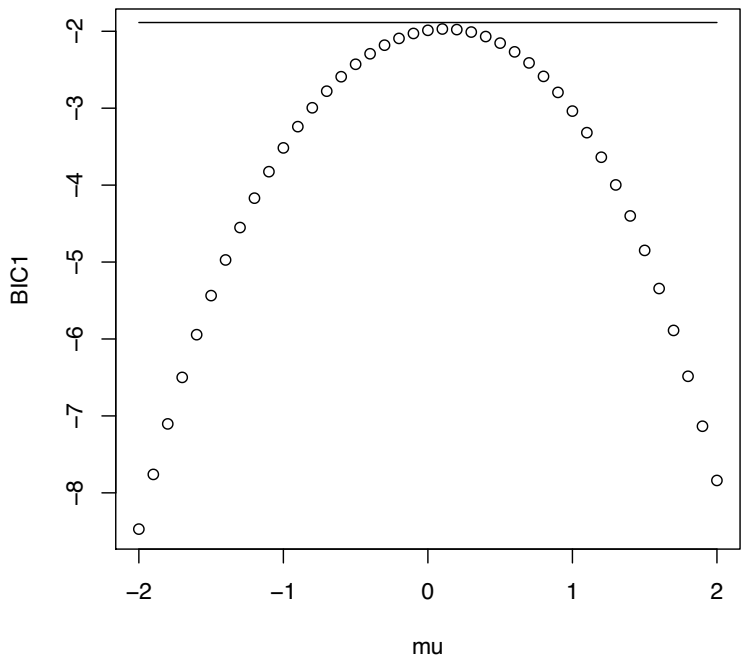
Tutorial and practical work about BIC (3/4)

```
# data drawing
n = 10
x = rnorm(n)

# paramerization
delta = 0.1
mu = seq(-2,2,0.1)
n.mu = length(mu)

# M1 : N(mu,1)
L1 = vector(mode = "numeric", length = n.mu)
for (i.mu in 1:n.mu){
    L1[i.mu] = sum(log(pnorm(x+delta,mu[i.mu],1)- pnorm(x-delta,mu[i.mu],1)))
}
BIC1 = L1 - 0.5*log(n)

# M2 : N(1,1)
L2 = vector(mode = "numeric", length = n.mu)
for (i.mu in 1:n.mu){
    L2[i.mu] = sum(log(pnorm(x+delta,1,1)- pnorm(x-delta,1,1)))
}
BIC2 = L2

# figure
plot(mu,BIC1)
lines(mu,BIC2)
```

Tutorial and practical work about BIC (4/4)



$\delta_1 = 0.1$



$\delta_2 = 2.2$

End of Part I