



**HAL**  
open science

# Traitement statistique des données manquantes-Part I Introduction to modeling

Christophe Biernacki

► **To cite this version:**

Christophe Biernacki. Traitement statistique des données manquantes-Part I Introduction to modeling. Doctorat. France. 2022. hal-03505648

**HAL Id: hal-03505648**

**<https://hal.science/hal-03505648>**

Submitted on 31 Dec 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Traitement statistique des données manquantes

Atelier Statistique de la SFdS – 10 et 11 mars 2021

—

## Part I Introduction to modeling

C. Biernacki

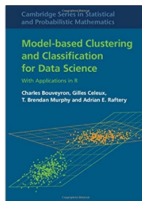
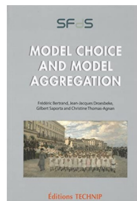
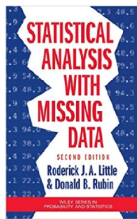
Laboratoire P. Painlevé, UMR CNRS 8524 & Université de Lille & Inria



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:

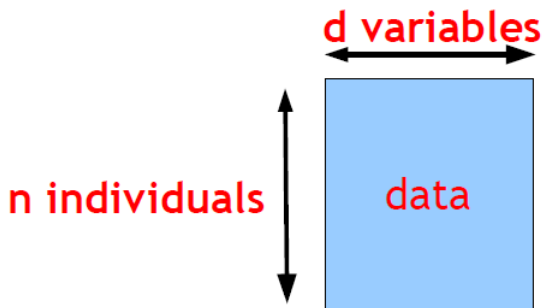
- 1 Part I: Introduction to modeling
- 2 Part II: Numerical and non-numerical data
- 3 Part III: Missing not at random data (MNAR) → ongoing research
- 4 Part IV: Binned data for big data analysis → ongoing research

# Outline

- 1 Data
- 2 Missing data
- 3 Embedding interest through an example
- 4 "Full" modeling
- 5 Frequentist estimation
- 6 Bayesian estimation
- 7 Model selection

Everything begins from data!

## Data sets structure



# Today's features: full mixed/missing



**categorical**  
Marital status  
**married**

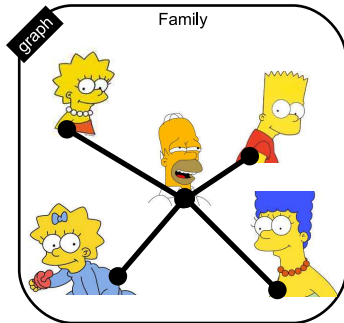
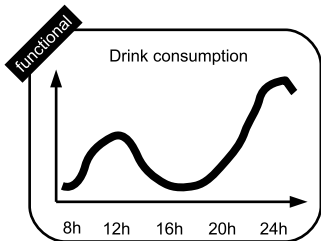
**integer**  
Children  
**3**

**missing**  
Size (m)  
**?**

**rank**  
Drink preference  
**beer > soda > water**

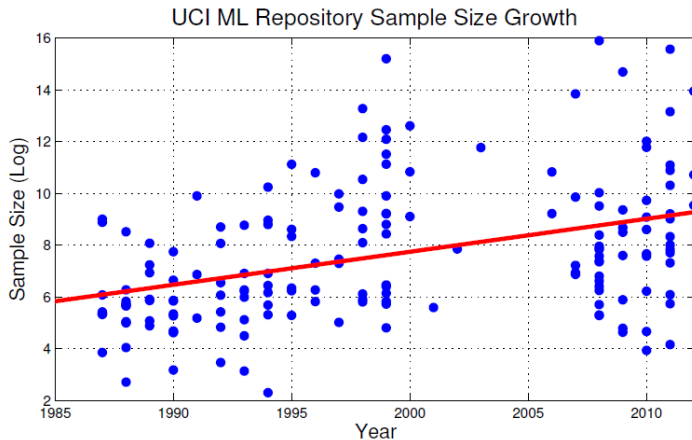
**ordinal**  
Intelligence  
**low**

**continuous**  
Weight (kg)  
**119.5**



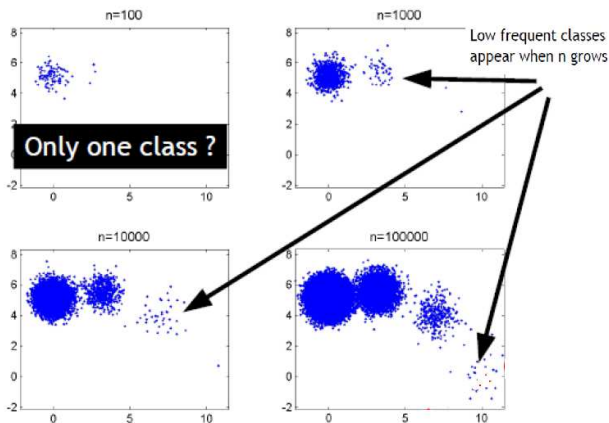
**And so on...**



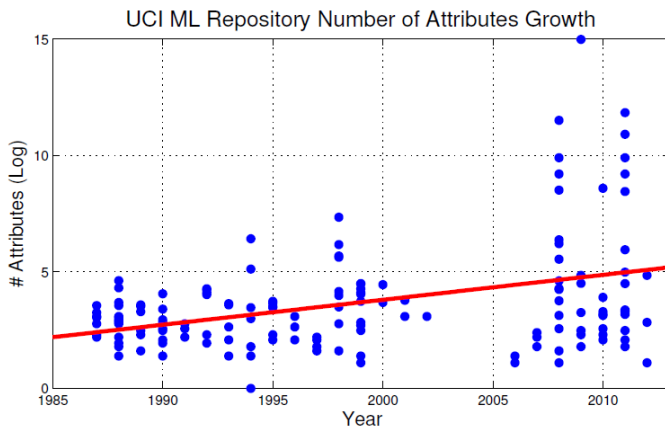
Large data sets ( $n$ )<sup>1</sup>

<sup>1</sup>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

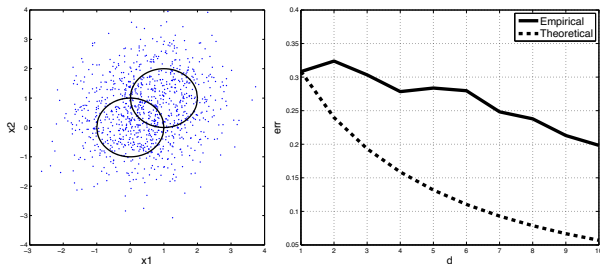


# High-dimensional data $(d)^2$



## 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

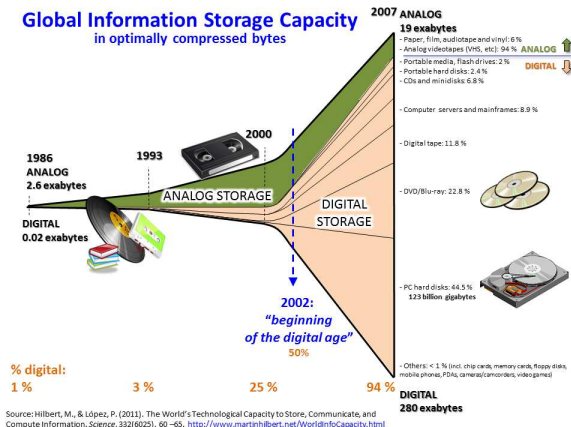
- **Storage cost per MB:** 700\$ in 1981, 1\$ in 1994, 0.01\$ in 2013  
→ price divided by 70,000 in thirty years
- **Storage capacity of HDDs:**  $\approx 1.02$  Go in 1982,  $\approx 8$  To today  
→ capacity multiplied by 8,000 over the same period
- **Computer processing speed:** 1 gigaFLOPS<sup>3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

---

<sup>3</sup>FLOP = Floating-point Operations Per Second

# Digital flow

- **Digital in 1986:** 1% of the stored information, 0.02 Eo<sup>4</sup>
- **Digital in 2007:** 94% of the stored information, 280 Eo (multiplied by 14,000)



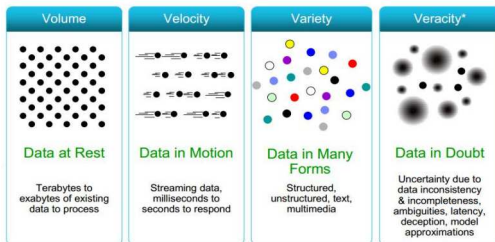
## Societal phenomenon

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 $\mathbf{x}$

- A **set** of  $n$  individuals

$$\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$$

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

$$\mathbf{x}_i = \{\mathbf{x}_{i1}, \dots, \mathbf{x}_{id}\}$$

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

- A  **$n$ -uplet** of individuals

$$\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$$

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

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

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

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

- 1 Data
- 2 Missing data**
- 3 Embedding interest through an example
- 4 "Full" modeling
- 5 Frequentist estimation
- 6 Bayesian estimation
- 7 Model selection

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 $x^o$			
?	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

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

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

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

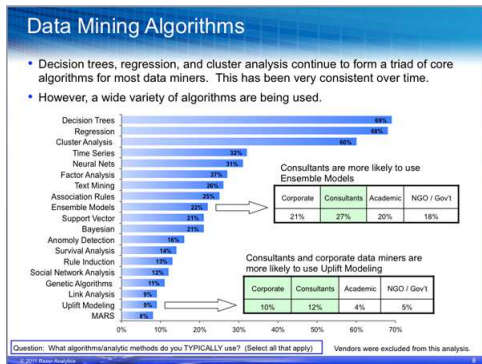
$\mathbf{x} = \{\mathbf{x}^o, \mathbf{x}^m\}$  is the full dataset with its observed and missing parts

### Notation illustration

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

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

Clustering everywhere<sup>5</sup>...



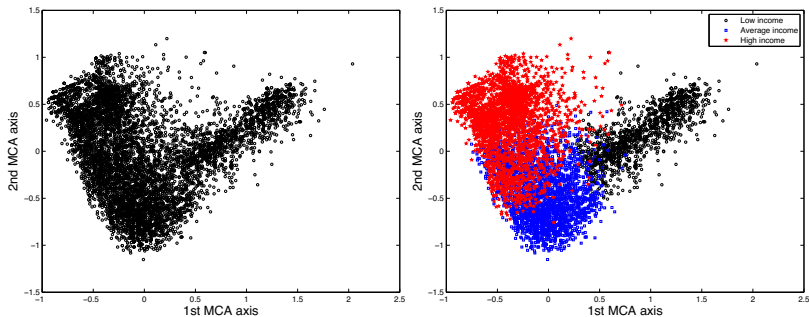
See Part II

<sup>5</sup>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  $G_1, \dots, G_K$ :  $z = (z_1, \dots, z_n)$ ,  
 $z_i = (z_{i1}, \dots, z_{iK})'$

$$x_i \in G_k \Leftrightarrow z_{ih} = \mathbb{I}_{\{h=k\}}$$

### Mixed, missing, uncertain

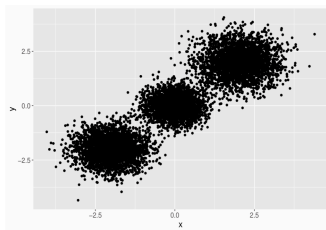
Individuals $x^o$				Partition $z$	$\Leftrightarrow$	Clusters
?	0.5	red	5	? ? ?	$\Leftrightarrow$	???
0.3	0.1	green	3	? ? ?	$\Leftrightarrow$	???
0.3	0.6	{red,green}	3	? ? ?	$\Leftrightarrow$	???
0.9	[0.25 0.45]	red	?	? ? ?	$\Leftrightarrow$	???
↓	↓	↓	↓			
continuous	continuous	categorical	integer			

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

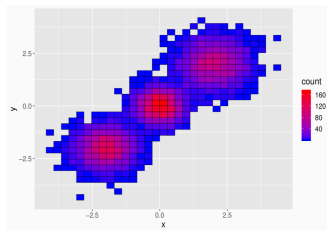


## Other more advanced missing data: binned data

- 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<sup>6</sup>: possible bias and underestimation of the variability

## General guidelines

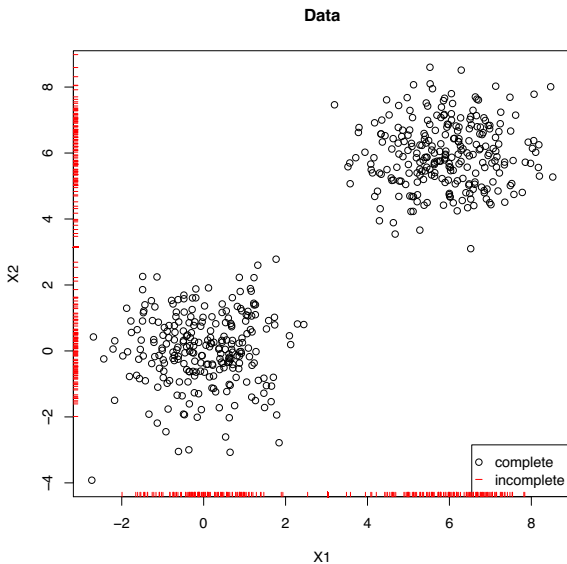
- Obtaining a complete dataset is **not** the final goal
- Missing data management should **take into account the initial analysis goal**

### Purpose of Part I

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

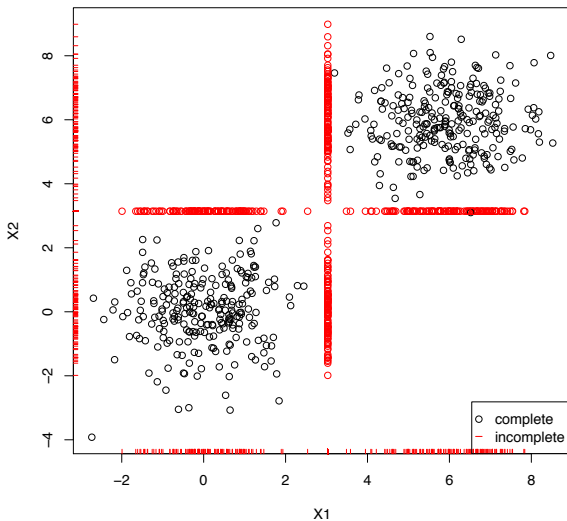
<sup>6</sup>See others lessons for classical methods of imputation

# Illustration of the risk of imputation



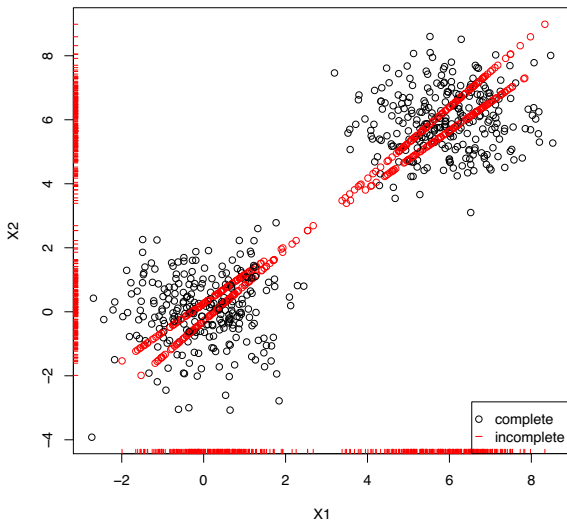
# Illustration of the risk of imputation

Data after imputation of missing data by the mean



# Illustration of the risk of imputation

Data after imputation of missing data by regression



# Outline

- 1 Data
- 2 Missing data
- 3 Embedding interest through an example**
- 4 "Full" modeling
- 5 Frequentist estimation
- 6 Bayesian estimation
- 7 Model selection

## Prostate cancer data<sup>7</sup>

- **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 histologic 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 ( $\approx 1\%$ )

---

<sup>7</sup>Byar DP, Green SB (1980): Bulletin Cancer, Paris 67:477-488

## Descriptors of the prostate cancer data set

<i>Covariate</i>	<i>Abbreviation</i>	<i>Number of Levels</i> (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 histologic grade	SG	
Serum prostatic acid phosphatase	AP	
Bone metastases	BM	2



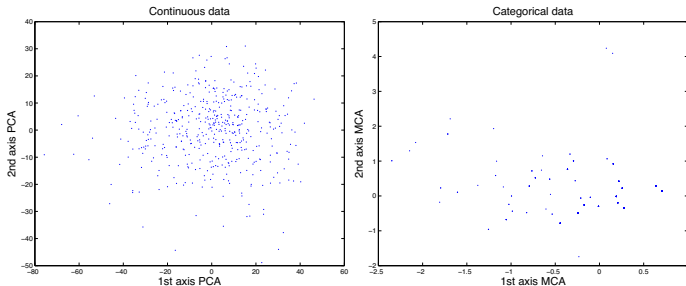
# Aim

We forget the classes (Stages of the disease) for performing [clustering](#)

## Questions

- How many clusters?
- Which partition?

Visually not so easy...



## Two strategies in competition

- **Strategy "mice<sup>8</sup> + MixtComp<sup>9</sup>"**: MixtComp on the dataset completed by mice

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

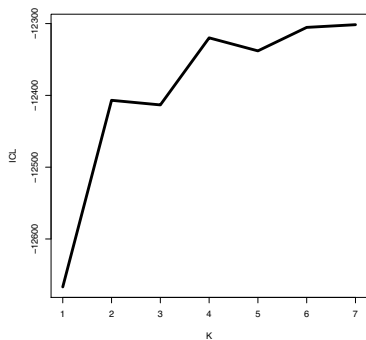
- **Strategy "full MixtComp"**: MixtComp on the observed (no completed) dataset

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

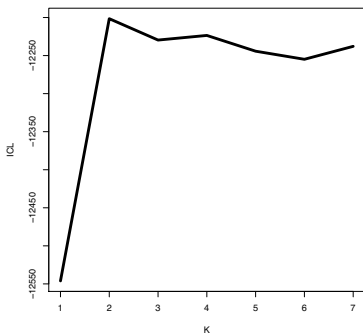
<sup>8</sup>See other lessons: <http://cran.r-project.org/web/packages/mice/mice.pdf>

<sup>9</sup>See Part II: <https://cran.r-project.org/web/packages/RMixtComp/index.html>

## Choosing $K$ with the ICL criterion<sup>10</sup>



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



full MixtComp  
 $\hat{K} = 2$

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

<sup>10</sup>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

- 1 Data
- 2 Missing data
- 3 Embedding interest through an example
- 4 "Full" modeling**
- 5 Frequentist estimation
- 6 Bayesian estimation
- 7 Model selection

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

# From "partial" to "full" modeling

## Partial modeling

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

## Full modeling

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

## Missing data: typology of the missing mechanisms

- Missing completely at random (**MCAR**):

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

- Missing at random (**MAR**):

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

- 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*

581

## 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,  $\theta$ , 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  $\theta$ , 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  $\theta$ . 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

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

### Inference of $\theta$

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

---

<sup>a</sup>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 → is the case in Part I
- MNAR is not ignorable → will be the case in Part III

Likelihood-based inference concerns both frequentist and Bayesian paradigms

# Outline

- 1 Data
- 2 Missing data
- 3 Embedding interest through an example
- 4 "Full" modeling
- 5 Frequentist estimation**
- 6 Bayesian estimation
- 7 Model selection

## Observed-data log-likelihood estimation of $\theta$

- **Principle:** MLE (Maximum Likelihood Estimate)

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

where the **observed**-data log-likelihood is defined by

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

- **Properties:** we have, with  $\theta^* = \arg \min_{\theta \in \Theta} \text{KL}(p, p_\theta)$ <sup>11</sup>

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


where  $\nu$  is the number of free continuous parameters in  $\Theta$ , and

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

$$\mathbf{K} = \text{Var}_{X_1} \nabla \ln p(X_1; \theta^*)$$

- **Algorithm:** EM and variants (see later)

---

<sup>11</sup> $p$  is the (unknown) true distribution,  $p_\theta = p(\cdot; \theta)$  and KL is the Kullback-Leibler divergence. 

# Expectation-Maximization (EM) algorithm

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

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

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

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

- **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  $\ell$ , 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  $\mu$  in the case where
  - $\mathbf{x} = (x_1, \dots, x_n) \sim \mathcal{N}(\mu, 1)$  with  $\mu = 1$  to be estimated (variance is known)
  - $\mathbf{x}^o = (x_1, \dots, x_{n^o})$  is the known sample of size  $n^o$
  - $\mathbf{x}^m = (x_{n^o+1}, \dots, x_{n^o+n^m})$  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  $\ell(\mu^{(q)}; \mathbf{x}^o)$  over  $q$

## Tutorial and practical work about EM (2/5)

### ■ MLE:

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

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

### Remark

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

## Tutorial and practical work about EM (3/5)

### ■ E-step of EM:

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

### ■ M-step of EM:

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



## Tutorial and practical work about EM (4/5)

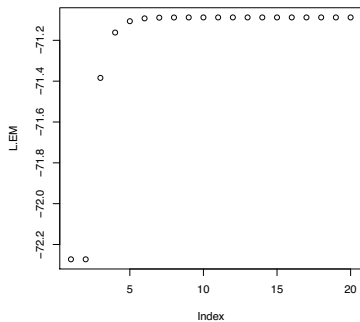
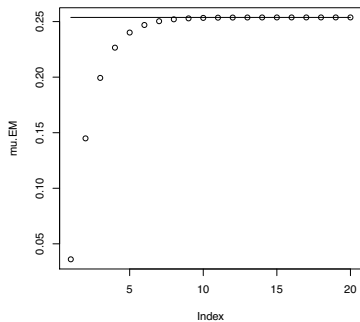
```
# data drawing
n.obs = 50
x.obs = rnorm(n.obs)
n.mis = 50

# 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  $\ell$  over  $\theta$

- Initialisation:  $\theta^{(0)}$
- Iteration ( $q$ ):
  - SE-step: draw missing values

$$\mathbf{x}^{m(q)} \sim p(\cdot | \mathbf{x}^o; \theta^{(q)})$$

- M-step: maximize  $\ell_c$

$$\theta^{(q+1)} = \arg \max_{\theta \in \Theta} \ell_c(\mathbf{x}^o, \mathbf{x}^{m(q)}; \theta)$$

- Stopping rule: iteration number  $Q$

### Properties

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

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

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  $(\theta^{(q)})$  produces a confidence interval
- $\ominus$ : no punctual convergence (does not improve  $\ell$  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:

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

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

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

thus leading to the standard result

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

## Tutorial and practical work about SEM (3/4)

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

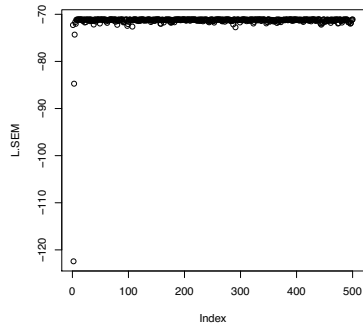
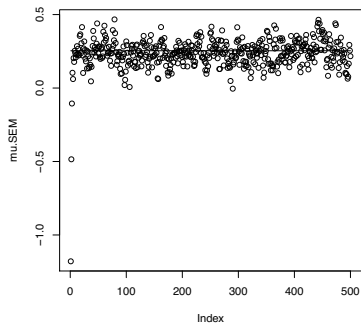
# closed-form ML
mu.ML = mean(x.obs)

# 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.SEM[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

- 1 Data
- 2 Missing data
- 3 Embedding interest through an example
- 4 "Full" modeling
- 5 Frequentist estimation
- 6 Bayesian estimation**
- 7 Model selection



## Bayesian estimation of $\theta$

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

- **Principle**<sup>12</sup>:

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

where

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

---

<sup>12</sup>Other definitions are possible.

## Gibbs algorithm

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

- Initialisation:  $\theta^{(0)}$

- Iteration ( $q$ ):

- Draw  $\mathbf{x}^{m(q)}$ :

$$\mathbf{x}^{m(q)} \sim p(\cdot|\mathbf{x}^o, \theta^{(q)})$$

- Draw  $\theta^{m(q)}$

$$\theta^{(q+1)} \sim p(\cdot|\mathbf{x}^o, \mathbf{x}^{m(q)})$$

- Stopping rule: iteration number  $Q$

### Properties

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

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

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

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

after using standard Gaussian results applied to the following expression

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

## Tutorial and practical work about Gibbs (3/4)

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

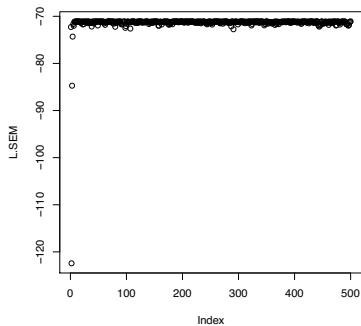
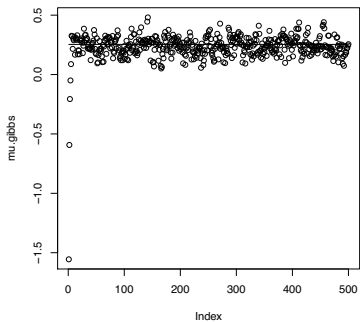
# closed-form ML
mu.ML = mean(x.obs)

# 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

- 1 Data
- 2 Missing data
- 3 Embedding interest through an example
- 4 "Full" modeling
- 5 Frequentist estimation
- 6 Bayesian estimation
- 7 Model selection**

## Model definition

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

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

- **Dimension  $\nu$** : it corresponds to the number of free *continuous* parameters

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

### Interest of choosing a model

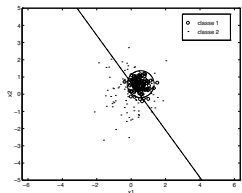
- A too simple model involves bias modeling
- A too complex model involves variance of estimation



# Importance of model selection: example in supervised learning

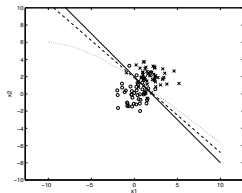
Model  $m$  = parametric structure of the borderline shape between two groups

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  $\mathbf{m}$ :

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

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

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

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

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

- Difficulties:

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

## BIC criterion: genesis

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

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

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

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

## BIC criterion: consistency

- **Consistency:** BIC asymptotically selects

$$\mathbf{m}^* = \arg \inf_{\mathbf{m} \in \mathcal{M}} \text{KL}(p, p_{\theta_{\mathbf{m}}^*})$$

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

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

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

$$p(\chi_{\Delta\nu}^2 > \Delta\nu \ln(n)) \leq p(|\chi_{\Delta\nu}^2 - \mu| > \Delta\nu \ln(n) - \mu) \leq \frac{\sigma^2}{(\Delta\nu \ln(n) - \mu)^2} \xrightarrow{n \rightarrow \infty} 0$$

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

## Tutorial and practical work about BIC (1/4)

- Let the sample  $\mathbf{x} = (x_1, \dots, x_n)$  where  $x_1, \dots, x_n \stackrel{iid}{\sim} \mathcal{N}(0, 1)$
- Let the uncertain sample  $\mathbf{x}^\delta = ([x_1 - \delta, x_1 + \delta], \dots, [x_n - \delta, x_n + \delta])$
- 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  $\mathbf{x}^\delta$
- Vary  $\delta$  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

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

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  $\text{BIC}^\delta$  value for different values of  $\mu$  on the same figure

## Tutorial and practical work about BIC (3/4)

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

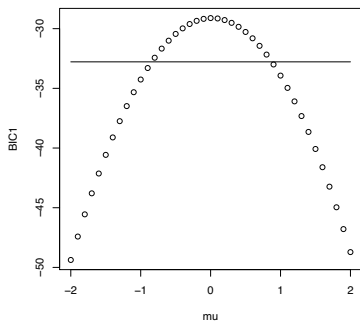
# parameterization
delta = 0.1
mu = seq(-2,2,0.1)
n.mu = length(mu)

# 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)

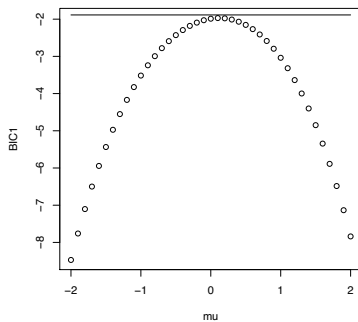
# M2 : N(1,1)
L2 = vector(mode = "numeric", length = n.mu)
for (i.mu in 1:n.mu){
  L2[i.mu] = sum(log(pnorm(x+delta,1,1)- pnorm(x-delta,1,1)))
}
BIC2 = L2

# 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